A Semi-Supervised Acoustic Scene Classification Network Based on Multi-Modal Information Fusion

Junkang Yang*, Hongqing Liu†, Liming Shi‡, Lu Gan§, Hiromitsu Nishizaki*¶ and Chee Siang Leow*
* Integrated Graduate School of Medicine, Engineering, and Agricultural Sciences, University of Yamanashi, Japan
† Chongqing Key Lab of Mobile Communications Technology, Chongqing University of Posts and Telecommunications, China
‡ School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, China
§ College of Engineering, Design and Physical Science, Brunel University, U.K.
E-mail: hnishi@yamanashi.ac.jp

***Abstract*—This paper presents our semi-supervised acoustic scene classification (ASC) framework submitted to the APSIPA ASC 2025 Grand Challenge, which focuses on city- and time-aware ASC under limited labeled data. Our approach leverages a multi-modal network architecture that fuses audio mel-spectrograms with spatiotemporal metadata (city identity and timestamps) to capture dynamic acoustic scene variations across urban environments. The model employs a residual-based CNN with attention mechanisms for robust feature extraction, enhanced by multi-modal fusion. To address label scarcity, we adopt a staged semi-supervised pipeline: pre-training on TAU Urban Acoustic Scenes 2020 and CochlScene datasets with specaugment and mixup augmentations, and then iterative fine-tuning on challenge data with pseudo-labeling to expand the training set was conducted, resulting in performance improvement. Experimental results demonstrate the efficacy of our city/time-aware design and semi-supervised strategies on our validation data.**

I. Introduction

Acoustic scene classification (ASC) has become a critical research area in computational audition, with applications ranging from smart city monitoring to intelligent audio devices. Traditional ASC systems typically treat acoustic scenes as static categories [1], failing to account for the significant variations that occur across different geographical locations and temporal contexts. This limitation becomes particularly evident in real-world scenarios where the acoustic characteristics of the same scene category such as a public square or a shopping district can vary dramatically between cities due to cultural differences and urban design [2], as well as across different times of day or days of the week [3]. The City and Time-Aware Semi-supervised Acoustic Scene Classification Challenge in APSIPA ASC 2025 seeks to address these gaps by incorporating city-level location data and precise timestamps alongside audio samples, pushing the boundaries of current ASC technology toward more context-aware and adaptable solutions.

The challenge builds upon previous work in semi-supervised learning for ASC while introducing novel dimensions of complexity. While the ICME 2024 [4] challenge focused on addressing domain shift across geographic regions, it did not explicitly leverage city identity and temporal metadata as discriminative features. This year's competition provides a unique opportunity to explore how these contextual cues can enhance classification performance, particularly when labeled data is scarce, which is a common constraint in real-world applications [5]. By encouraging participants to develop methods that effectively utilize both labeled and unlabeled data in conjunction with spatiotemporal information, the challenge aims to foster innovations in semi-supervised and domain adaptation techniques that can better handle the dynamic nature of acoustic environments.

From a practical standpoint, the outcomes of this challenge hold significant potential for industrial applications. Urban sound monitoring systems, smart devices, and acoustic analytics platforms stand to benefit from models that can adapt to city-specific soundscapes and temporal patterns [6]. For instance, a subway station in Beijing may exhibit different acoustic characteristics compared to one in Shanghai, and these may further vary between morning rush hours and late-night operations. By capturing these nuances, the developed solutions can lead to more robust and context-sensitive ASC systems. Moreover, the focus on semi-supervised learning aligns with the industry's need for scalable solutions that can leverage abundant unlabeled data, making the research impactful.

In this paper, we present our approach submitted to the challenge, detailing our methodology for integrating spatiotemporal metadata with audio features in a semi-supervised framework. Our work explores novel techniques for feature representation, domain adaptation, and contextual fusion, with the goal of improving classification accuracy across diverse urban environments and time periods. Through extensive experimentation and analysis, we demonstrate how city and time awareness can significantly enhance ASC performance while maintaining generalizability.

II. Proposed Method

A. Netork Architechture

The overall network structure is shown in Fig. 1. It processes input audio spectrograms (shape: [batchsize, 1, frames, bins]) by first passing them through a 7×7 convolutional kernel for

¶ Corresponding author.

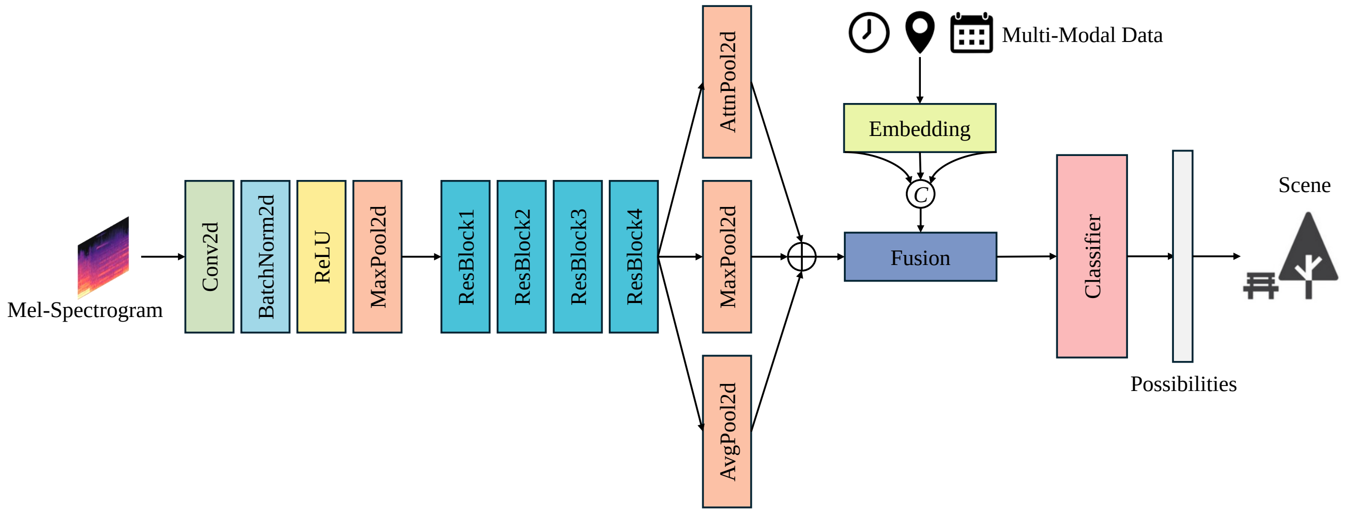

Fig. 1. Overal network architechture.

initial spatial feature extraction, followed by batch normalization and ReLU activation before entering a 3×3 max pooling layer for spatial compression. The data subsequently flows through four residual blocks, where each block employs 3×3 convolutions coupled with attention mechanisms (sequentially applying channel then spatial attention [7]). As the channel dimensions progressively expand to [64,128,256,512] respectively, four downsampling operations (stride=2) are performed to deeply extract time-frequency features. The feature map then enters an innovative pooling fusion stage: spatial attention weights are element-wise multiplied with the feature map and summed across time-frequency dimensions, while separate global average pooling and max pooling operations are computed, and these three vectors are additively combined to form robust audio feature representations. When multimodal processing is enabled, this 512-dimensional vector is concatenated along the feature dimension with location embedding tensors (mapping discrete location IDs to embeddings) and temporal feature tensors (processed through a two-layer fully connected network), then compressed to 256 dimensions with a fusion layer incorporating batch normalization, ReLU activation, and dropout [8]. Finally, regardless of modality mode, the features flow through a three-layer fully connected classifier (each containing batch normalization, ReLU activation, and 0.4 dropout regularization), progressively compressed through 128 to 64 dimensions before outputting class probability distributions at the target classification dimension. The detailed designs of resblock and classifier in the network are depicted in Fig. 2.

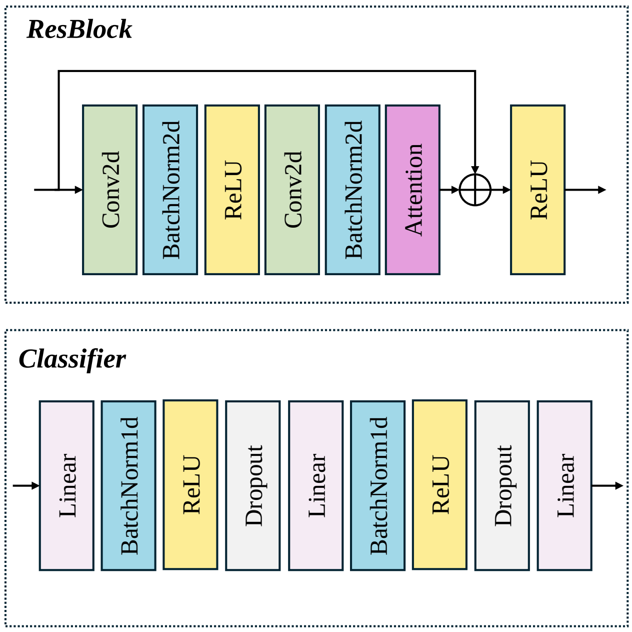

Fig. 2. Details of resblock and classifier.

B. Data & Augmentation

For pre-training, we use the development dataset of TAU urban acoustic scenes 2020 mobile [9] and CochlScene [10]. In order to keep the same format with the data proposed by challenge, we normalized these data to 44.1 kHz and 10 seconds each. The types of scene in these 3 dataset are not the same, so we manually perform the classification for pre-training data to adapt the challenge's requirements. The details are illustrated in Table I. During pre-training we use 20% of the data for validation and 80% for training, and the following augmentation methods are applied.

SpecAugment. SpecAugment [11] is a data augmentation method that operates directly on audio mel-spectrograms, enhancing model robustness through three key operations: time shifting, frequency masking, and time masking. This approach

TABLE I
RELATIONSHIP BETWEEN THE LABEL TYPE OF CHALLENGE DATA AND OUR PRE-TRAINING DATA

	Challenge Data	TAU Urban Acoustic Scenes 2020	CochlScene
Labels	bus	bus	bus
	airport	airport	-
	metro	metro station, metro	subway, subway station
	restaurant	-	restaurant
	shopping mall	shopping mall	-
	public square	public square	-
	urban park	park	park
	traffic street	street traffic	street
	construction site	-	-
	bar	-	cafe

simulates variations in real-world audio signals by disrupting local continuity in the mel-spectrogram, forcing the model to learn more global features rather than local details. It is particularly effective for addressing common environmental noise interference in acoustic scene classification. It does not require additional computational resources to generate synthetic samples and can improve model's performance by simply applying masking operations to the original spectrogram .

Mixup. Mixup [12] employs linear interpolation to blend samples and their labels from different categories, constructing new samples that lie between the original ones to enhance model generalization. In ASC, this method combines spectrograms of two different scenes at a random ratio while mixing their corresponding labels proportionally, mimicking the gradual transitions and overlaps of real-world soundscapes. This augmentation technique effectively mitigates model overfitting to specific samples, leading to smoother decision boundaries—especially useful for handling cases where different classes share similar acoustic characteristics. Unlike SpecAugment, Mixup expands the training data distribution by implicitly modeling relationships between samples. In our system, we follow the setting of mixup in [13] and the parameter α is set to 1.0.

C. Training

During training, the development set's metadata files are shuffled after setting the random seed to 1234 and split into training and validation sets at an 8:2 ratio. The training process employs a batch size of 64. Training process sets an upper limit of 1000 epochs with an equally stringent early stopping mechanism. Training is terminated if validation performance fails to improve for 20 consecutive epochs. This stopping condition provides ample time for model convergence. The Adam [14] optimizer is used with an initial learning rate of 5×10^{-4}, a relatively small value conducive to stable training. The learning rate scheduling adopts a step decay strategy, multiplying the learning rate by a decay factor of 0.9 every 2 epochs, forming a smooth decay curve. The loss function in all stages is cross entropy [15].

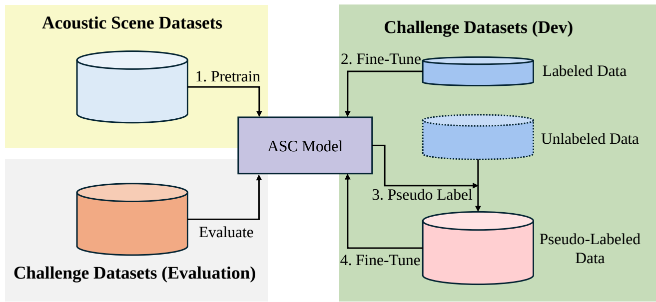

Fig. 3. Training and evaluation process.

As shown in Fig. 3, our semi-supervised training framework follows a four-stage pipeline [4] that progressively enhances model performance by effectively utilizing both labeled and unannotated data. The process begins with pre-training on pre-training data using the proposed network with specaugment [11] and mixup [12] augmentation, establishing a robust pre-trained model. In the second supervised fine-tuning phase, we adapt this pre-trained model to our specific classification task by carefully monitoring validation metrics to save the best fine-tuned model. The third pseudo-labeling stage then leverages this optimized model to generate high-confidence predictions for unlabeled examples of developing data, creating expanded training sets that combine original annotations with these labeled samples. Finally, the pseudo-label training phase retrains the model on this augmented dataset, maintaining identical hyperparameters but benefiting from significantly more training examples, ultimately producing our final model. This structured approach achieves an optimal balance between supervised learning precision and semi-supervised learning's ability to extract knowledge from unlabeled data, resulting in models with both high accuracy and excellent generalization capabilities.

TABLE II
TRAINING ACCURACY ON VALIDATION DATA OF DIFFERENT STAGES.

Stage	Accuracy (Average)
Pre-Training	93.70%
First Round Fine-Tuning	87.00%
Second Round Fine-Tuning	87.60%

III. RESULTS

During the experiment, the model's training accuracy at different stages was verified. The results is illustrated in Table II The average accuracy on the validation data during the pre-training phase reached 93.70%, demonstrating that the model was able to effectively learn the data characteristics during pre-training. During the first round of fine-tuning, the average accuracy on the validation data decreased slightly to 87.00%. This may be due to the need for the model to adjust to the new data distribution during fine-tuning, resulting in a temporary drop in accuracy. After the second round of fine-tuning, the average accuracy on the validation data increased slightly to 87.60%, indicating that after multiple rounds of fine-tuning, the model's performance on the validation data has gradually stabilized and improved. Overall, as the training phase progressed, the model's finally became stable on validation data.

TABLE III
FINAL RESULTS ON EVALUATION DATA.

Item	Accuracy
Bus	0.440
Airport	0.693
Metro	0.920
Restaurant	0.750
Shoppingmall	0.580
Public square	0.040
Urban park	0.700
Traffic street	0.650
Construction site	0.510
Bar	0.850
Macro-accuracy	0.613

Table III presents the classification accuracy and macro accuracy of the model across different scenarios in this challenge's evaluation data. From specific items, there are significant variations in classification accuracy among different scenarios. The metro and bar scenarios achieve classification accuracies of 0.920 and 0.850, respectively, demonstrating particularly outstanding performance. This may be because these two scenarios possess distinctive and easily recognizable acoustic features or patterns such as the sound of trains moving and station announcements in metro, and the presence of music and crowd conversations in bar, which exhibit relatively stable audio characteristics. The model can effectively capture these key features during training, thereby achieving high accuracy.

In contrast, the classification accuracy for the public square is only 0.040, which is the lowest among all scenarios, indicating significant difficulties the model faces in processing sounds from public squares. This may be related to the complexity, diversity, and unpredictability of sounds in public squares, where various activity noises, conversations among different groups of people, and ambient background noises interweave, making it challenging for the model to extract discriminative features for accurate classification. Additionally, the accuracies for construction site (0.510) and shopping mall (0.580) suggest that the model's classification performance for these scenarios still requires improvement.

The model's macro accuracy reaches 0.613, ranking third among all teams. It performs well in some scenarios, future efforts could focus on deeper research into the characteristics of sound data from scenarios with lower classification accuracy, as well as collecting more relevant training data to further enhance the overall macro accuracy.

TABLE IV
COMPAXITY ANALYSIS OF PROPOSED MODEL.

Item	Value
#Params	21.65M
MACs	2.34G
CPU Inference Time	40ms

Furthermore, we test the model's complexity and the results are depicted in table IV. The CPU platform is Intel(R) Core(TM) i9-10940X CPU @ 3.30GHz and the input include 3 parts: a 44.1 kHz audio lasts 10 seconds and two character strings representing city and time information.

Our model has a total of 21.65M parameters, indicating a certain level of complexity, though not excessively large, which helps balance performance and computational resource consumption. The model requires 2.34 billion FLOPs, which means that it performs a substantial amount of numerical computations during sound scene classification tasks. This could be a factor affecting the model's inference speed. However, when deployed on CPU, the model achieves an inference time of just 40ms, demonstrating its ability to meet real-time or near-real-time ASC requirements in practical applications.

IV. CONCLUSION

In this paper, we detailed our system submitted to APSIPA ASC 2025 Grand Challenge: City and Time-Aware Semi-supervised Acoustic Scene Classification. We used publicly released datasets including TAU and CochlScene during pre-training and fine-tune the model with the data provided by oganizers to address the ASC task. We employed a self-designed model to infuse the multi-modal data and generate reliable pseudo-labeled data. Additionally, we used specaugment and mixup augmentations to obtain the final results. We provide the code and checkpoint at https://github.com/JunkangYang/ALPS-ASC.

REFERENCES

[1] B. Ding, T. Zhang, C. Wang, *et al.*, “Acoustic scene classification: A comprehensive survey,” *Expert Systems with Applications*, vol. 238, p. 121 902, 2024, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2023.121902.

[2] Y. Cai, Y. Tan, S. Li, X. Shao, and M. D. Plumbley, *Improving acoustic scene classification with city features*, 2025. arXiv: 2503.16862 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2503.16862.

[3] R. Parikh, H. Sundar, M. Sun, C. Wang, and S. Matsoukas, “Impact of acoustic event tagging on scene classification in a multi-task learning framework,” in *Interspeech 2022*, 2022, pp. 4192–4196. DOI: 10.21437/Interspeech.2022-10905.

[4] J. Bai, M. Wang, H. Liu, *et al.*, *Description on ieee icme 2024 grand challenge: Semi-supervised acoustic scene classification under domain shift*, 2024. arXiv: 2402.02694 [eess.AS]. [Online]. Available: https://arxiv.org/abs/2402.02694.

[5] L. Alzubaidi, J. Bai, A. Al-Sabaawi, *et al.*, “A survey on deep learning tools dealing with data scarcity: Definitions, challenges, solutions, tips, and applications,” *Journal of Big Data*, vol. 10, no. 1, p. 46, 2023. DOI: 10.1186/s40537-023-00727-2.

[6] B. İşler, “Urban sound recognition in smart cities using an iot–fog computing framework and deep learning models: A performance comparison,” *Applied Sciences*, vol. 15, no. 3, 2025, ISSN: 2076-3417. DOI: 10.3390/app15031201.

[7] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[8] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014, ISSN: 1532-4435. DOI: 10.5555/2627435.2670313.

[9] T. Heittola, A. Mesaros, and T. Virtanen, *Tau urban acoustic scenes 2020 mobile, development dataset*, Zenodo, Feb. 2020. DOI: 10.5281/zenodo.3670167. [Online]. Available: https://doi.org/10.5281/zenodo.3670167.

[10] I.-Y. Jeong and J. Park, “Cochlscene: Acquisition of acoustic scene data using crowdsourcing,” in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 17–21. DOI: 10.23919/APSIPAASC55919.2022.9979822.

[11] D. S. Park, W. Chan, Y. Zhang, *et al.*, “Specaugment: A simple data augmentation method for automatic speech recognition,” in *Interspeech 2019*, 2019, pp. 2613–2617. DOI: 10.21437/Interspeech.2019-2680.

[12] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, *Mixup: Beyond empirical risk minimization*, 2018. arXiv: 1710.09412 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1710.09412.

[13] Q. Wang, G. Zhong, H. Hong, *et al.*, “The nercslip-ustc system for semi-supervised acoustic scene classification of icme 2024 grand challenge,” in *2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2024, pp. 1–4. DOI: 10.1109/ICMEW63481.2024.10645399.

[14] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2017. arXiv: 1412.6980 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1412.6980.

[15] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, “A tutorial on the cross-entropy method,” *Annals of operations research*, vol. 134, no. 1, pp. 19–67, 2005. DOI: 10.1007/s10479-005-5724-z.