ASCMamba: Multimodal Time-Frequency Mamba for Acoustic Scene Classification

Bochao Sun* and Dong Wang†§ and Zhanlong Yang* and Jun Yang† and Han Yin‡
* School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, China
† School of Automation, Northwestern Polytechnical University, Xi'an, China
‡ School of Electrical Engineering, KAIST, Daejeon, Republic of Korea
§Corresponding author: auto_wd@mail.nwpu.edu.cn

***Abstract*—Acoustic Scene Classification (ASC) is a fundamental problem in computational audition, which seeks to classify environments based on the distinctive acoustic features. In the ASC task of the APSIPA ASC 2025 Grand Challenge, the organizers introduce a multimodal ASC task. Unlike traditional ASC systems that rely solely on audio inputs, this challenge provides additional textual information as inputs, including the location where the audio is recorded and the time of recording. In this paper, we present our proposed system for the ASC task in the APSIPA ASC 2025 Grand Challenge. Specifically, we propose a multimodal network, ASCMamba, which integrates audio and textual information for fine-grained acoustic scene understanding and effective multimodal ASC. The proposed ASCMamba employs a DenseEncoder to extract hierarchical spectral features from spectrograms, followed by a dual-path Mamba blocks that capture long-range temporal and frequency dependencies using Mamba-based state space models. In addition, we present a two-step pseudo-labeling mechanism to generate more reliable pseudo-labels. Results show that the proposed system outperforms all the participating teams and achieves a 6.2% improvement over the baseline. Code, model and pre-trained checkpoints are available at https://github.com/S-Orion/ASCMamba.git**

I. INTRODUCTION

Acoustic scene classification (ASC) is a crucial research problem in computational audition that aims to recognize the unique acoustic characteristics of an environment [1]–[3]. Potential applications of ASC techniques include environmental monitoring and smart devices. Yet prevailing methods often assume static scenes, neglecting spatiotemporal variability across locations and record times. Ignoring such context undermines model generalization in real-world deployments.

Unlike the ASC task in the ICME 2024 Challenge [4], the APSIPA ASC 2025 Grand Challenge focuses on two critical factors influencing the performance of the ASC task: additional contextual information and scarcity of labeled data. The problem of leveraging additional contextual information, such as city-level location data and precise timestamps, is explored in this challenge. Another key issue is utilizing abundant unlabelled data to train robust ASC systems.

In this paper, we present our approach for the ASC task in the APSIPA ASC 2025 Grand Challenge. Specifically, we adopt Mamba [5] as the backbone of our model, as it exhibits more prominent efficiency and performance advantages over Transformer [6] in processing long-duration audio sequences. To fully exploit the temporal and spectral dependencies in audio signals, ASCMamba applies multiple Mamba blocks for dynamic modeling in both the record time and frequency domains. Furthermore, to facilitate multi-modal information interaction, we adopt a Conditional Layer Normalization (CLN) mechanism to incorporate text embeddings into ASCMamba.

The Challenge offers an extensive collection of unlabeled data, which can be leveraged for semi-supervised learning approaches. In this work, we first pre-train the proposed ASCMamba on TAU Urban Acoustic Scenes (UAS) 2020 Mobile development dataset [7] and CochlScene dataset [8]. The labeled data from Chinese Acoustic Scene (CAS) development dataset to fine-tune the pre-trained ASCMamba. For unlabeled CAS samples, we use the first fine-tuned ASCMamba to generate pseudo labels. For certain unlabeled samples, the pseudo-labels generated by the ASCMamba model have low confidence. To improve the quality of these pseudo-labels, we develop a secondary system dedicated to generating pseudo-labels for above mentioned low-confidence cases, and then use the intersection of these pseudo-labels with those predicted by the ASCMamba model to produce the reliable pseudo-labels. The second system is based on the challenge baseline, i.e., SE-Trans [9]. We improve the SE-Trans architecture by incorporating multi-scale pooling to enhance the ability of feature representation. In addition, we introduce an extra fully connected layer for indoor/outdoor binary classification as a prior. We then adjust the ASC class confidence scores based on the binary classification results to further improve accuracy. Finally, the ASCMamba is fine-tuned once more on the union of labeled and pseudo-labeled data, which serves as the final ASC model for evaluation.

II. DATASETS

A. Overview

The TAU UAS 2020 Mobile development dataset [7] and the CochlScene dataset [8] serve as the sources for pre-training ASCMamba model. TAU UAS 2020 Mobile comprises 23,046 samples, each delivered in binaural format at a 48 kHz sampling rate. CochlScene provides 76,115 single-channel audio files sampled at 44.1 kHz. Because these datasets cover different acoustic-scene taxonomies, we remove selected scene labels and merge others to construct a unified pre-training set. Table I lists the resulting counts of audio recordings per scene, where the data are used to pre-train the proposed ASCMamba

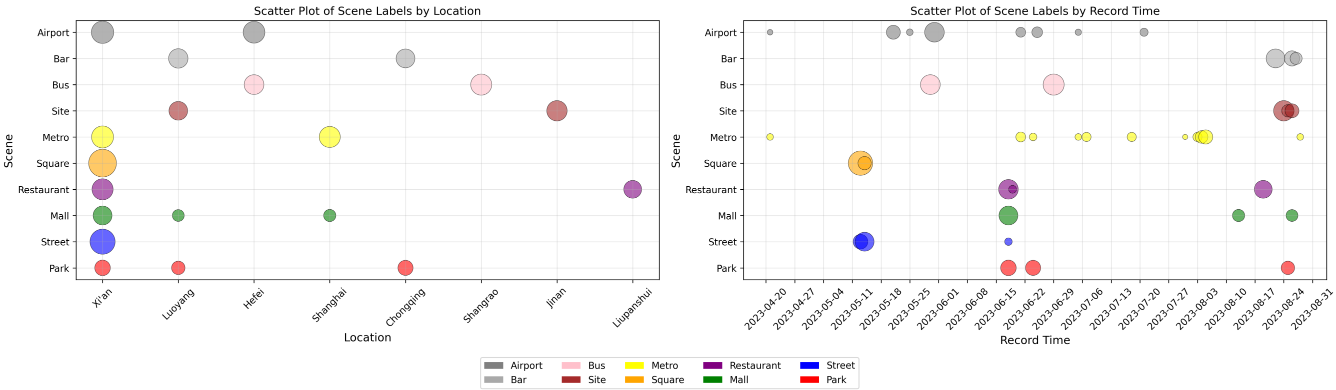

Fig. 1. Distribution of acoustic scenes in development set by location and record time.

TABLE I
THE NUMBER OF AUDIO RECORDINGS FOR EACH SCENE IN THE NEW GENERATED

Pre-training dataset

Scene	Number of audio recordings
Airport	2,302
Bus	8,125
Car	5,845
Metro	8,201
Metro Station	8,201
Public square	2,303
Restaurant	5,933
Shopping Mall	2,303
Traffic Street	8,049
Urban Park	8,048
Total	59,310

model. The CAS 2023 development dataset comprises 8,700 audio clips, 20% of which are labeled. We extract the labeled data as the initial labeled dataset to fine-tune the ASCMamba and improved SE-Trans models.

B. Location and Record Time Distribution Analysis

In the development dataset, the number of labeled samples is 1,740, and the spatial and temporal distribution of scene labels exhibit significant characteristics. From a spatial perspective, the development dataset comprises data from 8 locations, including *Xi'an*, *Luoyang*, *Hefei*, *Shanghai*, *Chongqing*, *Shangrao*, *Jinan*, and *Liupanshui* in descending order of representation. The distribution of scene labels over location and record time is shown in Fig. 1. The size of the circle represents the density of scene labels, and a larger circle indicates a higher number of scene labels for a specific location or record time period. As can be seen, the spatial and temporal information exhibit a strong correlation with scene labels. This provides an important basis for utilizing spatiotemporal information as context for the classification task.

It should be noted that there are 12 locations in the evaluation set, with 6 overlapping in the development dataset and the remaining 6 being exclusive to the evaluation dataset. Regarding temporal coverage, the development dataset spans from April 21, 2023, to August 28, 2023, while the evaluation dataset spans from April 22, 2023, to August 27, 2023.

TABLE II
THE RESHAPED SEQUENCES

Sequences	Means
$\boldsymbol{X}_t \in R^{(B*F)\times T\times C}$	each frequency bin as a sequence over record time
$\boldsymbol{X}_f \in R^{(B*T)\times F\times C}$	each frame as a sequence over frequency

III. PROPOSED APPROACH

A. ASCMamba

As shown in Fig 2, the proposed ASCMamba model is composed of 2 blocks: DenseEncoder and Dual-path Mamba Block. Details are described as follows.

1) DenseEncoder: The DenseEncoder is a two-dimensional convolutional feature extraction module designed for time-frequency representation learning in audio processing tasks. It consists of three primary components: an initial channel projection block, a dense connectivity-based feature refinement block, and a frequency-axis down sampling block. Specially, a DenseBlock with a depth of 4 is applied to encourage feature reuse and gradient propagation across layers. Inspired by DenseNet [10], each layer within the block receives feature maps from all preceding layers as input, promoting the learning of compact and discriminative spectral patterns. It mainly serves as an efficient front-end feature extractor to effectively capture local and hierarchical features in the spectrogram.

2) Dual-path Mamba: The core of the block is a dual-path Mamba architecture, which separately models dynamics along the time and frequency dimensions. Given an input spectrogram feature, which is the encoded representation extracted by the preceding DenseEncoder from log-mel inputs. We denote this feature as $\boldsymbol{X} \in R^{B\times C\times T\times F}$, which is reshaped into 2 sequences, i.e., $\boldsymbol{X}_t$ and $\boldsymbol{X}_f$, as shown in Table II. Each sequence is processed independently by a MambaBlock, capturing long-range dependencies along their respective axes.

To enable **multimodal integration**, the model accepts text embeddings based on the input location and record time information, which are projected into a shared conditional space $R^{D_{cond}}$. This conditional vector is used to modulate the

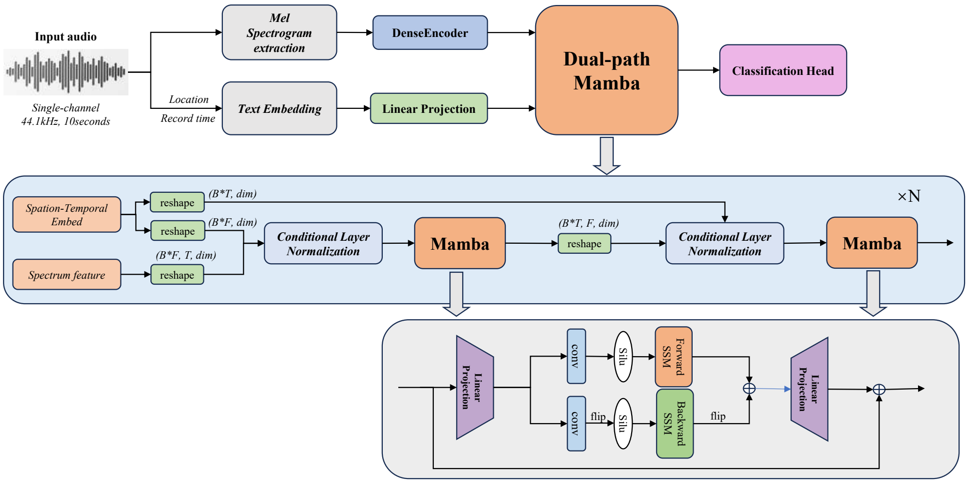

Fig. 2. Overview of the proposed ASCMamba, which is composed of a DenseEncoder and a Dual-path Mamba Block.

internal feature representations through a CLN mechanism. Specifically, CLN dynamically adjusts the affine parameters (scale γ and β bias) of layer normalization based on the context, which can be formulated as:

$$CLN(x, c) = \gamma(c) \cdot LN(x) + \beta(c) \tag{1}$$

where $\gamma(x)$ and $\beta(x)$ are generated by linear projections from the conditional vector c. This modulation is applied before both the temporal and frequency Mamba paths, allowing the model to adapt its feature space according to spatiotemporal contexts, such as emphasizing different frequency patterns depending on the time of day or geographic region.

B. Improved SE-Trans

In order to make full use of the official development dataset, we develop a second system dedicated to generating pseudo-labels for the data with low confidence predicted by the ASCMamba model, and then use the intersection of these pseudo-labels with the labels predicted by the ASCMamba model as the reliable pseudo-labels. The second system adopts an **improved SE-Trans** architecture to enhance the ability of feature expression and the accuracy of classification through multi-scale pooling and two-step classification strategy.

Specifically, the improved SE-Trans uses 2 Squeeze-and-Excitation (SE) modules (with 2 convolutional layers, a multi-scale SE layer, and pooling), where the SE layers apply 1×1, 2×2, and 3×3 multi-scale pooling to strengthen feature representation. Features are then processed by a Transformer encoder to model spatiotemporal dependencies, followed by 2 fully connected layers outputting 10 specific scene labels and 2 rough labels. The two-class fully connected layer aims to classify an input audio clip into 1 of 2 main classes, including in-door and out-door. The final prediction of scene class is obtained by score fusion of these 2 classifiers [11], which is expressed as follows:

$$c = argmax_{c,i\supset c} y_c^1 * y_i^2 \tag{2}$$

where y_c^1 denotes the probability of class c predicted by the ten-class classifier, while y_i^2 represents the probability of class i predicted by the binary classifier, with $c \in \{1, 2, ...10\}$, and $i\,\{1, 2\}$. Since $i \supset c$, means that class i is a superset of class c. For instance, the indoor scene category is a superset that includes bus, metro, restaurant, shopping mall and bar. This design significantly improves recognition accuracy and robustness in complex scenarios.

C. Two-step Pseudo-labeling

To exploit the unlabeled data, we introduce a two-step pseudo-labeling scheme. In the first step, the pre-trained ASCMamba model is fine-tuned on the initial labeled dataset and subsequently used to assign pseudo labels to the unlabeled clips. The predicted posterior probabilities of unlabeled data are sorted from high to low, and we select the top 90% of the pseudo-labeled data. In the second step, the ASCMamba and the improved SE-Trans are employed to generate pseudo labels for the left 10% of the unlabeled data in the development set. Audio samples predicted to belong to the same scene category by both of these 2 models are selected as reliable pseudo-labeled data. These samples are finally combined with the initial labeled dataset to form the definitive labeled dataset used to train our submission system.

IV. Experimental Setups

A. Validation Sets Creation

To explore the performance of the baseline and proposed system, we first analyze the distribution discrepancy between

TABLE III
THE ACC (%) OF BASELINE AND PROPOSED SYSTEMS ON VALID-EASY AND VALID-HARD. "L&RT" MEANS "LOCATION AND RECORD TIME".

System	Airport	Bar	Bus	Construction Site	Metro	Republic Square	Restaurant	Shopping Mall	Traffic Street	Urban Park	Average
Valid-Easy											
SE-Trans (Baseline)	76.92	94.59	96.88	97.06	90.91	97.44	93.33	88.89	100	91.30	92.73
ASCMamba w/ L&RT	97.07	100	100	100	90.01	100	93.01	100	97.00	100	**97.71**
ASCMamba w/o L&RT	84.62	91.30	95.15	94.12	96.08	93.33	93.33	92.73	96.88	100	93.75
Valid-Hard (5%)											
ASCMamba w/ L&RT	82.61	93.17	96.15	94.87	97.10	95.25	93.26	93.00	96.00	100	94.14
ASCMamba w/o L&RT	84.62	100	97.73	90.00	97.44	100	94.33	100	100	100	**96.41**

the validation set and the training set, and find that their distributions are generally consistent. However, since this competition requires participants to pay attention to the impact of domain shift on audio classification, we divide the validation set into **Valid-Easy** and **Valid-Hard** with the help of the official evaluation set to simulate the potential distribution difference between the training data and the final test data. Specifically, we first extract the Log-Mel filter bank (LMFB) features from audio files in both the evaluation set and the validation set, then calculate the statistical distribution of each audio feature, such as the mean and variance. In the second step, we use cosine similarity to quantify the differences between the feature distribution of the validation set and the evaluation set. A threshold of 0.9 is used: all samples with cosine similarity higher than 0.9 are considered to have a consistent distribution with the evaluation set and thus assigned to Valid-Hard, while the remainder are allocated to Valid-Easy.

Furthermore, considering that in real-world applications the spatiotemporal distribution of audio data may differ from the training domain, we randomly shuffle the recording time and location metadata of a portion of the Valid-Hard set. The proportion of shuffled data is denoted as x%, referred to as Valid-Hard (x%). Specifically, we set x to 5, 10, 20, 50, 70, and 100, in order to more comprehensively evaluate the impact of spatiotemporal distribution shifts on model performance.

B. Evaluation Metric

Following the challenge baseline, we evaluate the performance of the ASC system using accuracy (ACC) as the primary metric. Accuracy measures the proportion of correctly classified samples over the total number of samples, providing a straightforward and interpretable assessment of the model's classification effectiveness across different scene categories.

C. Training Details

We first resample the audio recordings in the TAU UAS 2020 and CAS 2023 datasets to 44.1 kHz. All audio clips have a fixed length of 10 seconds. LMFB were extracted as audio features by using the Librosa [12] library with 2048 short-time Fourier transform (STFT) points, a 40ms Hann window, and a frame shift of 20ms. We apply 64 Mel-filter bands on the spectrograms and generate a feature tensor shape of $500 \times 64 \times 1$. Dropout rate is set to 0.1. We train our model using the Adam [13] optimizer. The Batch size is set to 4 and the learning rate is set to 0.0001. All of our models are trained using the PyTorch toolkit [14] due to its ease of use and efficient support for rapid prototyping and debugging.

We adopt a four-stage semi-supervised training framework that progressively improves model performance by effectively leveraging labeled and unlabeled data. Firstly, we train the ASCMamba model using our self-generated pretraining dataset to obtain a base pretrained model. In the second stage, we fine-tun the model using the labeled portion of the development dataset. During the third phase, we generate pseudo-labels for unlabeled development data based on the fine-tuned model from the previous step, identifying reliable data within this subset. Specifically, to identify reliable data for the second round of fine-tuning in the third stage, we implement a two-stage selection strategy as follows: first, guided by the confidence scores of pseudo-labels, the top 90% data instances with the highest reliability are directly designated as reliable data. Second, for the remaining 10% of data instances, ASCMamba and the improved SE-Trans are employed to generate pseudo-labels. Audio samples predicted to belong to the same scene category by both of 2 models are selected as reliable pseudo-labeled data. This strategy is designed to maximize the utilization of officially provided training data, thereby enhancing the model's generalization capability.

Finally, we perform another fine-tuning using both the real labeled and reliable pseudo-labeled data from the development set on the fine-tuned model, ultimately obtaining a final model that balances accuracy and generalization capability.

V. RESULTS AND DISCUSSIONS

To demonstrate the effectiveness of our model in audio classification tasks, we conduct the following ablation studies. We evaluate the performance of the baseline and the 2 model variants separately on the Valid-Easy and Valid-Hard datasets. The results are listed in Table III. When evaluated on Valid-Easy, ASCMamba w/ L&RT achieves the best performance with an average accuracy of 97%. This strongly validates the model's ability to extract and learn audio signal features. In contrast, when tested on Valid-Hard, ASCMamba w/o L&RT outperforms ASCMamba w/ L&RT. Additionally, following the division method described in IV-A, we separately evaluate the performance of the 2 model variants on Valid-Hard when spatiotemporal information is randomly shuffled at different

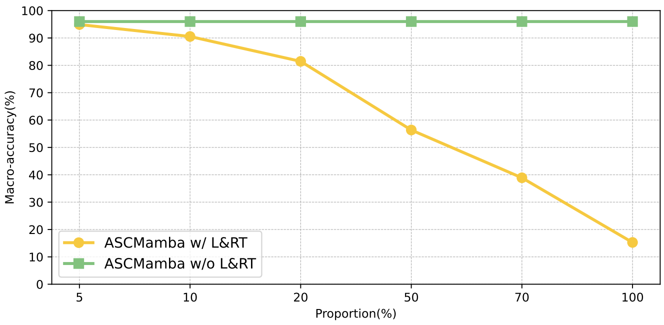

Fig. 3. The performance of the two systems after shuffling the Valid-Hard data at different proportions

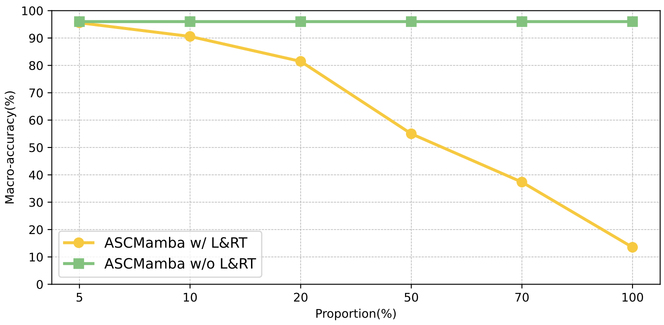

Fig. 4. The performance of the two systems with the gradual increase in the proportion of unseen locations within the validation set

proportions, and the results can be found in Fig. 3.

In addition to randomly shuffling spatiotemporal information, we also notice that the evaluation set contains locations unseen in the training set, and thus we conduct further experiments: we select different proportions of data from the validation set. The proportions of these selected data are the same as those of Valid-Hard, and randomly change the location to those seen in the evaluation set but not in the validation set, such as *Nanchang*, *Shenyang*, *Guangzhou*, *Changchun*, *Tianjin*, and *Taiyuan*. The experimental results are visualized in Fig. 4.

Results on Fig. 3 and Fig. 4 show that, when the spatiotemporal distribution of the test data differs from the training domain, or when unseen locations appear in the test set, ASCMamba w/ L&RT consistently underperforms ASCMamba w/o L&RT. Furthermore, greater distributional discrepancies lead to a more substantial degradation in the performance of ASCMamba w/ L&RT relative to its counterpart.

We attribute this to the reliance of ASCMamba w/ L&RT on temporal and positional information, which arises from its extensive utilization of multi-modal information. This also implies that when external multi-modal information is ambiguous, it is preferable to solely utilize audio-specific features for analysis; whereas when multi-modal information is clear, employing multi-modal fusion can significantly enhance the model's discriminative performance.

The evaluation results of each team on the official blind test set in Table IV. It is clearly evident that our system achieves a 6.2% improvement over the official baseline system and outperforms all systems submitted by other participating teams, securing first place in this competition. Rank-2 System [15] proposes a city-separated cross-validation scheme to evaluate the model's generalization capability on unseen locations. It employs low-frequency feature constraints to mitigate overfitting risks and replaces max-pooling with average pooling for aggregating temporal frame information. However, it has limitations such as restricted applicability due to the low-frequency assumption, under-utilization of multi-modal information, and unknown robustness to extreme data cases. Rank-3 System [16] employs an attention-based ResNet as the audio backbone network, which effectively captures key features in the time-frequency domain. However, this system fails to propose an intrinsic mechanism to address the model's heavy reliance on positional information, which may hinder its generalization to unseen locations.

TABLE IV
THE EVALUATION RESULTS OF EACH TEAM ON THE OFFICIAL BLIND TEST SET

System	Score(Macro-accuracy)
Rank-1 (ASCMamba w/o L&RT)	**64.4%**
Rank-2 System	62.8%
Rank-3 System	61.3%
Rank-4 System	58.6%
SE-Trans (Baseline)	58.2%

VI. CONCLUSION

In this paper, we present our approach to tackle the ASC task of the APSIPA ASC 2025 Grand Challenge. In detail, we propose a novel architecture named ASCMamba, which uses a DenseEncoder to extract local and hierarchical features, and applies a Dual-path Mamba block for sequence modeling. In addition, we employ a two-step mechanism to generate reliable pseudo-labels for unlabeled data with low confidence. Experimental results in demonstrate that ASCMamba exhibits distinct advantages in fusing multi-modal information. On the official blind test dataset, the proposed ASCMamba outperforms the existing baseline system by 6.2% in macro ACC, and achieves the first rank among all participating teams, which sufficiently demonstrate the effectiveness of the proposed approach in acoustic scene classification tasks.

REFERENCES

[1] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.

[2] H. Yin, J. Bai, Y. Xiao, *et al.*, "Exploring text-queried sound event detection with audio source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2025, pp. 1–5.

[3] H. Yin, J. Chen, J. Bai, *et al.*, "Multi-granularity acoustic information fusion for sound event detection," *Signal Processing*, vol. 227, p. 109 691, 2025.

[4] J. Bai, M. Wang, H. Liu, *et al.*, "Description on ieee icme 2024 grand challenge: Semi-supervised acoustic scene classification under domain shift," *arXiv preprint arXiv:2402.02694*, 2024.

[5] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[6] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[7] H. Toni, M. Annamaria, and V. Tuomas, "Tau urban acoustic scenes 2020 mobile development dataset [data set]," *Zenodo*, 2020.

[8] I.-Y. Jeong and J. Park, "Cochlscene: Acquisition of acoustic scene data using crowdsourcing," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, IEEE, 2022, pp. 17–21.

[9] J. Bai, J. Chen, M. Wang, M. S. Ayub, and Q. Yan, "A squeeze-and-excitation and transformer-based cross-task model for environmental sound recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 15, no. 3, pp. 1501–1513, 2023.

[10] D. Kim, B. Heo, and D. Han, "Densenets reloaded: Paradigm shift beyond resnets and vits," in *European Conference on Computer Vision*, Springer, 2024, pp. 395–415.

[11] H. Hu, C.-H. H. Yang, X. Xia, *et al.*, "A two-stage approach to device-robust acoustic scene classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 845–849.

[12] B. McFee, C. Raffel, D. Liang, *et al.*, "Librosa: Audio and music signal analysis in python.," *SciPy*, vol. 2015, pp. 18–24, 2015.

[13] K. D. B. J. Adam *et al.*, "A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, vol. 1412, no. 6, 2014.

[14] A. Paszke, S. Gross, F. Massa, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[15] T. Kawamura, M. Sera, and N. Ono, "Evaluation of low-frequency feature restriction and average pooling for acoustic scene classification under unseen-city conditions," *Technical Report*, 2025.

[16] J. Yang, H. Liu, L. Shi, L. Gan, H. Nishizaki, and C. S. Leow, "A semi-supervised acoustic scene classification network based on multi-modal information fusion," *Technical Report*, 2025.