ASCMamba: Multimodal Time-Frequency Mamba for Acoustic Scene Classification

Bochao Sun* and Dong Wang* and Han Yin†
* Northwestern Polytechnical University, Xi'an, China
† School of Electrical Engineering, KAIST, Daejeon, Republic of Korea

***Abstract*—Acoustic Scene Classification (ASC) is a fundamental problem in computational audition, which seeks to classify environments based on their distinctive acoustic signatures. In the ASC task of the APSIPA ASC 2025 Grand Challenge, the organizers introduced a multimodal ASC task. Unlike traditional ASC systems that rely solely on audio input, this challenge provides additional textual information as inputs, including the city where the audio was recorded and the time of recording. In this paper, we present our submission system for the ASC task in the APSIPA ASC 2025 Grand Challenge. Specifically, we propose a multimodal network, ASCMamba, which integrates spectral, temporal, and contextual 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 systems achieve superior performance, with an improvement ranging from 4% to 5% over the challenge baseline.**

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. Potential applications of ASC techniques include environmental monitoring and smart devices. Yet prevailing methods often assume static scenes, neglecting spatio-temporal variability across cities and times. Ignoring such context undermines model generalization in real-world deployments.

Unlike the ASC task in the ICME 2024 Challenge [1], the APSIPA ASC 2025 Grand Challenge focuses two critical factors influencing the performance of 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 propose **ASCMamba**, a multimodal network for ASC tasks. To fully exploit the temporal and spectral dependencies in audio signals, ASCMamba applies multiple Mamba [2] blocks for dynamic modeling in both the time and frequency domains. Furthermore, to facilitate multimodal 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 [3] and CochlScene dataset [4]. These two datasets are then combined with the labeled data from Chinese Acoustic Scene (CAS) development dataset to fine-tune the pre-trained ASCMamba. For unlabeled CAS samples, we use the pre-trained 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 the 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 [5]. We improve the SE-Trans architecture by incorporating multi-scale pooling to enhance 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

The TAU UAS 2020 Mobile development dataset [3] and the CochlScene dataset [4] serve as the sources for pre-training ASCMamba model. TAU UAS 2020 Mobile comprises 23,040 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 removed selected scene categories and merged 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 and improved SE-Trans model.

The CAS 2023 development dataset comprises 8,700 audio clips, 20% of which are annotated. These labeled data are merged with the new pre-training dataset to create the initial labeled dataset. These labeled data are merged with the new pre-training dataset to create the initial labeled dataset.

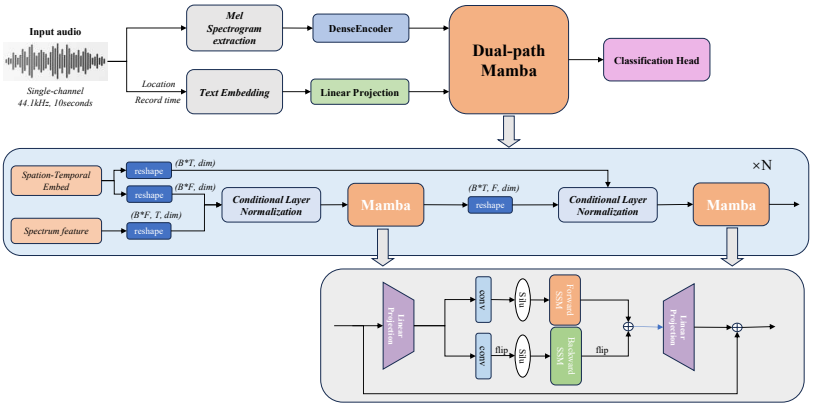

Fig. 1. Overview of the proposed ASCMamba, which is composed of a DenseEncoder and a Dual-path Mamba Block.

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

Pre-training dataset	
Scene	Number of audio recordings
Airport	2302
Bus	8125
Car	5845
Metro	8201
Metro Station	8201
Public square	2303
Restaurant	5933
Shopping Mall	2303
Traffic Street	8049
Urban Park	8048
Total	59310

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 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 Figure 1, the proposed ASCMamba model is composed of two 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 depth 4 is applied to encourage feature reuse and gradient propagation across layers. Inspired by DenseNet [6], 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 audio spectrum.

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 two 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 location and temporal embeddings, which are projected into a shared conditional space $\mathbb{R}^{D_{cond}}$. This conditional vector is used to modulate the 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 spatio-temporal 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 data set, we developed a second system dedicated to generating pseudo-labels for the data with low confidence predicted by ASCMamba model, and then used 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

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)	0.76	0.94	0.96	0.97	0.90	0.97	0.93	0.88	1.0	0.91	0.92
ASCMamba w/ L&RT	0.97	1.0	1.0	1.0	0.90	1.0	0.93	1.0	0.97	1.0	0.97
ASCMamba w/o L&RT	0.84	0.91	0.95	0.94	0.96	0.93	0.93	0.92	0.96	1.0	0.93
Valid-Hard											
ASCMamba w/ L&RT	0.82	0.93	0.96	0.94	0.97	0.95	0.93	0.93	0.96	1.0	0.94
ASCMamba w/o L&RT	0.84	1.0	0.97	0.90	0.97	1.0	0.94	1.0	1.0	1.0	0.96

and the accuracy of classification through multi-scale pooling and two-step classification strategy.

Specifically, the improved SE-Trans uses two Squeeze-and-Excitation (SE) modules (with two 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 [7] encoder to model spatio-temporal dependencies, followed by two fully connected layers outputting 10 specific scene categories and 2 rough labels. The two-class fully connected layer aims to classify an input audio clip into one of two main classes, including in-door and out-door. The final prediction of scene class is obtained by score fusion of these two classifiers[8], 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 predicted by the ten-class classifier, while y_i^2 represents the probability of class predicted by the binary classifier, $c \in \{1, 2, ...10\}$, $i \{1, 2\}$. Since $i \supset c$, means that class i is a super set of class. For example, the indoor scene category is the super set for bus, metro, restaurant, shopping mall and bar. This design significantly enhances 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 dataset. Audio samples predicted to belong to the same scene category by both of these two 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. Evaluation Metric

Following 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.

B. 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. Log-Mel filter bank (LMFB) were extracted as audio features by using Librosa [9] 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 generates a feature tensor shape of 500 × 64 × 1. Dropout rate is set to 0.1. We train our model using Adam [10] optimizer. The Batch size is set to 4 and learning rate is set to 0.0001. All of our models are trained using the PyTorch toolkit [11].

Because of the size of the data from the challenge dataset is much smaller than that from the pre-training dataset. To ensure that the model sees the challenge data more frequently during fine-tuning, We employ a hybrid approach, which proceeds as follows: first, pseudo-labels are predicted, and the confidence score for each data instance is generated. Guided by the confidence scores, the top 90% of data instances are selected as high-quality data for the second round of fine-tuning. For the remaining 10% of data instances, the ASCMamba and the improved SE-Trans are employed to generate reliable pseudo labels. Audio samples predicted to belong to the same scene category by both of these two models are selected as reliable pseudo-labeled data. Eventually, these two subsets, combined with the data bearing genuine labels, are integrated to fine-tune the target model. This strategy is designed to maximize the utilization of officially provided training data, thereby enhancing the model's generalization capability.

V. RESULTS

To explores the performance of the baseline and proposed systems, we split the validation data into two subsets: **Valid-Easy** and **Valid-Hard**. Specifically, the samples in Valid-Easy

exhibit relatively small spatio-temporal distribution differences from the training data. In contrast, considering that in real-world applications the spatio-temporal distribution of data may differ substantially from that of the training set, we assign samples with larger distributional discrepancies to Valid-Hard.

The experimental results are presented in Table III. On Valid-Easy, the accuracy rate of model classification has reached 97%.

For Valid-Hard, ASCMamba without city and time performs better, which can prove its sensitivity to spatio-temporal information. If the distribution of spatio-temporal information between the training data and the final test data is inconsistent, it will lead to a deterioration of the model's performance. Therefore, we use ASCMamba w/o L&RT as the final submitted system.

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 for ASC tasks, 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 show that when incorporate location and record time prior, ASCMamba's classification accuracy rate is 97%, outperforming the challenge baseline.

REFERENCES

[1] 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.

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

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

[4] 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)*, IEEE, 2022, pp. 17–21.

[5] 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.

[6] 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.

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

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

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

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

[11] 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.