Semi-supervised Acoustic Scene Classification under Spatial-Temporal Variability with a CRNN-based Model

Haowen Li*, Ziyi Yang*
* Nanyang Technological University, Singapore
E-mail: haowen.li@ntu.edu.sg, ziyi016@e.ntu.edu.sg

Abstract—**This report presents a CNN-RNN based model for acoustic scene classification (ASC) under spatial-temporal variability, which is the APSIPA ASC 2025 Grand Challenge. The proposed architecture, referred to as MobileASCNet, combines depthwise separable convolutions and ResNet-inspired modules to extract efficient spatial and residual features, and employs a GRU-based recurrent branch to model temporal dependencies. After city and time feature fusion, a multi-layer perceptron (MLP) along with additional residual blocks is used to further enhance classification performance. Unlike many recent approaches, our model is trained from scratch without using pretraining or knowledge distillation. Experimental results on the development set demonstrate the effectiveness of our approach, achieving a classification accuracy of 99.0%, outperforming the official baseline model (96.0%) with lower model complexity.**

I. Introduction

Acoustic Scene Classification (ASC) aims to recognize the type of environment in which an audio signal was recorded, such as a street, shopping mall, or metro station, based on its acoustic characteristics [1], [2]. ASC has been one of the core tasks in the Detection and Classification of Acoustic Scenes and Events (DCASE) challenges for nearly a decade [3], [4], and has attracted extensive research interest due to its applications in context-aware services, intelligent devices, and urban monitoring [5]. Recent advances in deep learning have significantly improved ASC accuracy on benchmark datasets, driven by powerful Convolutional Neural Networks (CNNs) [6]. However, the high computational complexity and memory requirements of such models hinder their deployment on resource-constrained platforms like smartphones, wearables, or embedded audio sensors. Moreover, ASC systems often suffer from performance degradation when evaluated on audio captured by previously unseen devices or recording conditions [7], [8], highlighting a need for models that are both efficient and robust to domain shift.

To address these issues, recent effort have investigated the development of compact neural network architectures [9] and efficient training paradigms, including knowledge distillation and model ensembling [10], [11]. These strategies aim to reduce model complexity while maintaining or improving generalization capability. In the ASC domain, low-complexity models like CP-Mobile [9] have demonstrated competitive accuracy with a fraction of the parameters of large networks. Our previous work contributes to this line of research by proposing a dual-level knowledge distillation framework that incorporates both output-level supervision and intermediate feature alignment to guide the training of low-complexity student models [12]. Similar approaches have been explored in DCASE challenges, where ensemble teacher models such as PaSST [6] and CP-ResNet [13], together with lightweight architectures implemented without distillation, exemplified by our Convolutional Neural Networks-Gated Recurrent Unit (CNN–GRU) based system for DCASE2025 Task 1 [14], have been employed to meet strict complexity constraints while maintaining competitive accuracy.

Extending beyond these approaches, the APSIPA ASC 2025 Grand Challenge building upon the IEEE ICME 2024 Grand Challenge [15], which focused on domain shift across cities and time, introduces a more contextually rich and realistic evaluation setting [16]. In this benchmark, each 10-second audio segment is annotated with both city-level location and timestamp metadata, encompassing 22 cities across China and spanning various time periods. The challenge adopts a semi-supervised learning protocol in which only a small portion of the development data is labeled, while the remainder remains unlabeled. This formulation imposes additional challenges by requiring systems to address label scarcity, spatial-temporal variability, and domain shifts simultaneously. Consequently, it encourages the development of models that are not only computationally efficient but also capable of leveraging contextual information and unlabeled data to improve robustness in real-world scenarios.

In this paper, we build upon these findings and shift our focus from knowledge transfer to architectural design. Specifically, to align with the challenge's objective of leveraging contextual metadata, we incorporate both city-level location and timestamp information during training. Within the semi-supervised learning framework, our method effectively utilizes both labeled and pseudo-labeled data to enhance model robustness across different cities and temporal contexts. We refer to this architecture as **MobileASCNet**, which emphasizes both its suitability for low-complexity deployment and its ability to incorporate spatial-temporal metadata for acoustic scene classification.

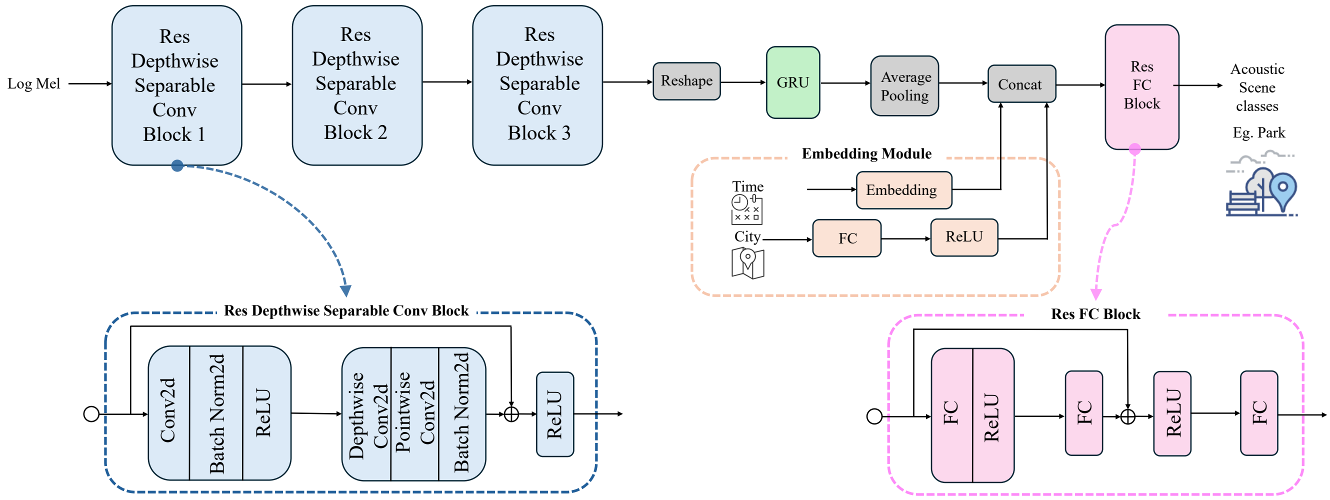

Fig. 1. Overview of the proposed MobileASCNet architecture.

II. Data Preprocessing and Feature Extraction

In this work, we adopt the same data preprocessing and feature extraction pipeline as the official APSIPA ASC 2025 Challenge baseline. The raw audio recordings are first converted into log-mel spectrogram representations, which are widely used in acoustic scene classification tasks due to their compactness and perceptual relevance.

Specifically, we apply short-time Fourier transform (STFT) to each audio waveform using a Hann window. The magnitude spectrogram is then projected onto the mel scale using a mel filter bank with 64 mel bands. Finally, logarithmic compression is applied to obtain the log-mel spectrogram. The detailed parameters used in the extraction are as shown in Table I.

TABLE I
Parameters for log-mel spectrogram extraction.

Parameter	Value
Sampling rate	44,100 Hz
FFT size (n_{fft})	2,048
Window length (win_{length})	1,764
Hop length (hop_{length})	882
Number of mel bands	64
Frequency range	50 Hz – 22,050 Hz
Window type	Hann

The configuration follows the official ASC Challenge baseline settings to ensure comparability and reproducibility. The resulting log-mel spectrograms are used as input to our model MobileASCNet described in the Section III.

III. Proposed Model Architecture

A. Model Architecture

As illustrated in Figure 1, the network consists of three main components:

- **Residual Depthwise Separable Convolutional Block:** The input log-mel spectrogram is first passed through a series of three residual depthwise separable convolutional blocks (ResDepthwise Separable Conv Block)as illustrated by the blue components in Figure 1. This design enables efficient extraction of local spatial features while maintaining strong representational capacity.
- **Temporal Modeling:** Following the ResDepthwise Separable Conv Blocks, the resulting feature map is reshaped along the temporal axis and processed by a GRU module, which captures temporal dependencies across time frames. The GRU outputs are then aggregated using temporal average pooling, producing a fixed-length representation of the entire acoustic scene.
- **Feature Fusion and Classification:** The outputs from GRU and are concatenated and passed through a embedding module, which is illustrated by the light orange components in Figure 1, followed by a multi-layer perceptron (MLP) and additional residual blocks (Res FC Block), which is illustrated by the light orange components in Figure 1.

B. Training Procedure

The training procedure largely follows the official baseline setup of the APSIPA ASC 2025 Grand Challenge. The CAS 2023 development dataset [15]contains approximately 24.1 hours of audio data, including both labeled (approximately 4.8 hours) and unlabeled (approximately 19.3 hours) segments, each accompanied by city and time metadata.

The model is trained in a three-stage semi-supervised learning pipeline:

1) **Initial Training:** The model is first trained using only the labeled data to learn discriminative acoustic features.
2) **Pseudo Labeling:** The trained model is then used to generate pseudo labels for the unlabeled portion of the dataset.
3) **Secondary Fine-tuning:** Both the original labeled and pseudo-labeled data are used to further fine-tune the model and improve its generalization across different acoustic conditions.

While the overall training structure is consistent with the challenge baseline, our method does not rely on any pretrained model from external datasets (e.g., TAU2020 [17]). Instead, we train the model from scratch, ensuring that no external data is involved throughout the entire training pipeline. This also demonstrates the model's strong learning capacity under limited supervision.

In all stages, the model incorporates city and time metadata through learnable embeddings, enabling it to adapt to spatial-temporal variability in real-world acoustic scenes.

IV. Experimental Results

A. Experimental Setup

All experiments were conducted using the development dataset provided in the APSIPA ASC 2025 Grand Challenge, derived from the CAS 2023 dataset [15]. Our model MobileASCNet was implemented in PyTorch and trained from scratch. The training followed the official semi-supervised three-stage protocol described in Section III-B, with no external pretraining or distillation applied.

We train our model MobileASCNet for a maximum of 200 epochs with a batch size of 64, using the Adaptive Moment Estimation (Adam) optimizer [18]. The initial learning rate is set to 1e-4 and is updated using a Step Learning Rate (StepLR) scheduler, which decays the learning rate by a factor of 0.9 every 2 epochs. To prevent overfitting and reduce training time, we apply an early stopping strategy [19] with a patience of 10 epochs based on validation accuracy.

All models were evaluated on the development set, and performance was measured using classification accuracy.

B. Performance Comparison

We compare our proposed **MobileASCNet** with the official challenge baseline [20] which adopts a cross-task SE-Trans architecture that combines Squeeze-and-Excitation and Transformer encoders to capture channel-wise and temporal dependencies in acoustic features. Table II shows the classification accuracy and parameters of different models on the development set.

TABLE II
Classification accuracy and model size comparison on the development set.

Model	Accuracy (%)	Params
Baseline	96.0	0.44M
MobileASCNet	**99.0**	**0.37M**

As shown in Table II, our proposed MobileASCNet achieves a classification accuracy of 99.0%, surpassing the official baseline by 3.0%. Notably, this performance is achieved without any external pretraining, highlighting the effectiveness of our architecture when trained from scratch. In addition, MobileASCNet uses fewer parameters, demonstrating a better trade-off between accuracy and model complexity. The observed performance gain can be attributed to the effective integration of depthwise separable convolutions, residual connections, GRU-based temporal modeling, and contextual embeddings.

V. Conclusions

In this report, we proposed **MobileASCNet**, a lightweight yet effective CRNN based architecture for ASC under spatial-temporal variability, as part of the APSIPA ASC 2025 Grand Challenge. The model integrates depthwise separable convolutions and ResNet-inspired modules to efficiently capture spatial and residual features, while a GRU models temporal dynamics. Additionally, contextual metadata such as city and time information are incorporated through an embedding module, further enhancing the scene recognition capability of model. Experimental results on the official development set demonstrate the effectiveness of our approach, achieving a classification accuracy of 99.0%, significantly outperforming the official baseline (96.0%) with fewer parameters without external data pretrained and knowledge distillation.

References

[1] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: A review of features, classifiers and datasets," *IEEE Trans. Audio Speech Lang. Process.*, vol. 23, no. 3, pp. 512–529, 2015.

[2] T. Mesaros, T. Heittola, and T. Virtanen, "Multi-device dataset for acoustic scene classification and sound event detection," in *Proc. DCASE Workshop*, 2018.

[3] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *2016 24th European signal processing conference (EUSIPCO)*, IEEE, 2016, pp. 1128–1132.

[4] A. Mesaros, R. Serizel, T. Heittola, T. Virtanen, and M. D. Plumbley, "A decade of dcase: Achievements, practices, evaluations and future challenges," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2025, pp. 1–5.

[5] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE press, 2006.

[6] K. Koutini, H. Eghbal-Zadeh, D. Widmann, C. Mertes, G. Schuller, and B. Schuller, "Efficient training of audio transformers with patchout," *arXiv preprint arXiv:2110.05069*, 2021.

[7] T. Mesaros, T. Heittola, K. Drossos, and T. Virtanen, "Tau urban acoustic scenes 2022 mobile: Three-device dataset for acoustic scene classification," DCASE2021 Challenge, Tech. Rep. 2021.

[8] Y. Tan, H. Ai, S. Li, and M. D. Plumbley, "Acoustic scene classification across cities and devices via feature disentanglement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1286–1297, 2024.

[9] B. Murauer and B. Schuller, “Efficient acoustic scene classification with cp-mobile,” DCASE2023 Challenge, Tech. Rep. 2023.

[10] C. Schmid and et al., “Efficient teacher-student training for acoustic scene classification using passt,” DCASE2023 Challenge, Tech. Rep. 2023.

[11] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.

[12] H. Li, Z. Yang, M. Wang, *et al.*, “Joint feature and output distillation for low-complexity acoustic scene classification,” *arXiv preprint arXiv:2507.19557*, 2025.

[13] K. Koutini, H. Eghbal-Zadeh, C. Mertes, G. Schuller, D. Widmann, and B. Schuller, “Receptive-field-regularized cnn variants for acoustic scene classification,” in *Proc. DCASE Workshop*, 2021.

[14] E.-L. Tan, J. W. Yeow, S. Peksi, H. Li, Z. Yang, and W.-S. Gan, “Sntl-ntu dcase25 submission: Acoustic scene classification using CNN-GRU model without knowledge distillation,” DCASE2025 Challenge, Tech. Rep., May 2025.

[15] 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].

[16] APSIPA ASC 2025 Grand Challenge Organizers, *Apsipa asc 2025 grand challenge*, https://ascchallenge.xshengyun.com/2025/index.html, Accessed: 2025-08-07, 2025.

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

[18] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[19] L. Prechelt, “Early stopping-but when?” In *Neural Networks: Tricks of the trade*, Springer, 2002, pp. 55–69.

[20] 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.