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

Haowen Li haowen.li@ntu.edu.sg
Smart Nation TRANS Lab
Nanyang Technological University, Singapore

Mou Wang wangmou21@mail.nwpu.edu.cn
Institute of Acoustics
Chinese Academy of Sciences, China

Zhengding Luo [*] luoz0021@e.ntu.edu.sg
Smart Nation TRANS Lab
Nanyang Technological University, Singapore

Ee-Leng Tan etanel@ntu.edu.sg
Smart Nation TRANS Lab
Nanyang Technological University, Singapore

Ziyi Yang ziyi016@e.ntu.edu.sg
Smart Nation TRANS Lab
Nanyang Technological University, Singapore

Woon-Seng Gan ewsgan@ntu.edu.sg
Smart Nation TRANS Lab
Nanyang Technological University, Singapore

Editors: Tatsuya Komatsu, Keisuke Imoto, Xiaoxue Gao, Nobutaka Ono, Nancy F. Chen

Abstract

In this work, we present MobileASCNet, a lightweight CRNN-based model designed for acoustic scene classification (ASC) under spatial-temporal variability, as defined in the APSIPA ASC 2025 Grand Challenge. The model combines depthwise separable convolutions and ResNet-inspired residual blocks for efficient spatial feature extraction, and employs a gated recurrent unit (GRU) branch to capture temporal dependencies. City and time embeddings are fused to enhance context-awareness. We conduct extensive comparisons under different training strategies, including training from scratch, pretraining with fine-tuning, and feature freezing. Without relying on knowledge distillation, MobileASCNet achieves a competitive classification accuracy on the development set, with low model complexity.

Code: https://github.com/HaowenLi/AudioAAAI2026

Keywords: Acoustic Scene Classification, Lightweight CRNN, Spatial-Temporal Variability, Context-Aware Learning

[*] Corresponding author.

1. Introduction

Acoustic scene classification (ASC) involves identifying the recording environment of audio signals, such as parks, shopping malls, or airports, based on through analysis of their distinctive acoustic features Barchiesi et al. (2015). As one of the core tasks in the Detection and Classification of Acoustic Scenes and Events (DCASE) challenges for nearly a decade Mesaros et al. (2025), ASC has garnered significant attention for its utility in context-aware services, intelligent devices, and urban monitoring Ding et al. (2024). Moreover, it serves as useful auxiliary task for improving the performance of sound event detection, speech enhancement, and other downstream audio processing applications Igarashi et al. (2022); Qian et al. (2025). Recent advances in deep learning have substantially improved the accuracy of ASC on benchmark datasets, driven by the success of powerful convolutional neural networks (CNNs) Koutini et al. (2021). However, the high computational and memory demands of these models limit their applicability in resource-constrained settings, such as smartphones, wearables, and embedded audio sensors. In addition, ASC systems often experience significant performance degradation when applied to audio recorded under previously unseen conditions or with unfamiliar devices Stowell et al. (2015); Tan et al. (2024), underscoring the need for models that are not only efficient but also robust to domain shifts.

To address these challenges, recent research has focused on designing compact neural network architectures Kim et al. (2021) and developing efficient training paradigms such as knowledge distillation and model ensembling Schmid and et al. (2023); Hinton et al. (2015). These strategies aim to reduce model complexity while preserving or enhancing generalization capability. In the context of ASC, lightweight models such as CP-Mobile Murauer and Schuller (2023) and variants Li et al. (2025a) have achieved competitive performance with significantly fewer parameters compared to larger networks. Our previous work contributes to this direction 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 Li et al. (2025b). Similar approaches have been explored in the DCASE challenges, where ensemble teacher models such as PaSST Koutini et al. (2021) and CP-ResNet Horváth et al. (2021) have been adopted. In addition, lightweight models trained without distillation, exemplified by our convolutional neural networks-gated recurrent unit (CRNN) based system for DCASE2025 Task 1 Tan et al. (2025), have also been employed to meet strict complexity constraints while maintaining high classification accuracy.

Building upon the IEEE ICME 2024 Grand Challenge Bai et al. (2024), which emphasized domain shift across cities and time, the APSIPA ASC 2025 Grand Challenge introduces a more contextually rich and realistic evaluation setting by leveraging the Chinese Acoustic Scene (CAS) 2023 dataset. In this benchmark, each 10-second audio segment is annotated with both city-level location and timestamp metadata, covering 22 cities across China and spanning diverse time periods. The challenge adopts a semi-supervised learning protocol, where only a small portion of the development data is labeled and the remainder remains unlabeled. This setup presents additional challenges, requiring models to handle label scarcity, spatial-temporal variability, and domain shifts concurrently. As a result, it promotes the development of models that are not only computationally efficient but also capable of leveraging contextual information and unlabeled data to enhance robustness in real-world acoustic environments. Our work is developed based on the CAS 2023 dataset, with the goal of evaluating ASC performance under these challenging conditions.

In this paper, we shift the focus from knowledge transfer techniques to the architectural and training strategy design for ASC under spatial-temporal variability. To align with the challenge objective of leveraging contextual metadata, we incorporate both city-level location and timestamp information into model training. We propose a lightweight yet effective architecture, termed **MobileASCNet**, which integrates multi-modal metadata and is optimized for real-world deployment scenarios. Under a semi-supervised learning framework, the model utilizes both labeled and pseudo-labeled data to enhance generalization across unseen cities and time periods. To better understand the design space, we conduct extensive ablation studies, including comparisons of training with and without pretraining, freezing strategies, the application of data augmentation, and various fusion methods for temporal and spatial metadata. These studies validate the effectiveness of MobileASCNet in achieving high accuracy under strict complexity constraints while maintaining robustness to domain shifts.

2. Pre-training on External Datasets

To improve model generalization and provide a robust initialization for downstream fine-tuning on the APSIPA ASC 2025 Challenge dataset, we adopt a pre-training strategy based on two large-scale public acoustic scene datasets: **TAU Urban Acoustic Scenes 2020** Toni et al. (2020) and **CochlScene** Jeong and Park (2022). Both datasets cover diverse real-world sound environments and have been widely used in ASC research.

Table 1: Category alignment and sample statistics across CAS 2023 dataset, TAU 2020 dataset, and CochlScene dataset.

CAS 2023 dataset	**TAU 2020 dataset**	**Samples**	**CochlScene dataset**	**Samples**	**Selected for Pre-training**
Airport	✓ (airport)	2302	–	–	✓ 2303
Bus	✓ (bus)	2304	✓ (bus)	5821	✓ 8152
Metro	✓ (metro)	2304	✓ (subway)	5897	✓ 8201
Restaurant	–	–	✓ (restaurant)	5933	✓ 5933
Shopping Mall	✓ (shopping_mall)	2303	–	–	✓ 2303
Public Square	✓ (public_square)	2303	–	–	✓ 2303
Urban Park	✓ (park)	2304	✓ (park)	5744	✓ 8048
Traffic Street	✓ (street_traffic)	2304	✓ (street)	5745	✓ 8049
Construction Site	–	–	–	–	✗
Bar	–	–	✗ (restroom / cafe)	–	✗

The official CAS 2023 dataset which used in this challenge defines 10 acoustic scene categories: *Bus, Airport, Metro, Restaurant, Shopping Mall, Public Square, Urban Park, Traffic Street, Construction Site, and Bar.* The TAU2020 dataset contains 10 categories, including *Airport, Bus, Metro (Subway), Metro Station, Shopping Mall, Traffic Street, Street Pedestrian, Park, Tram, and Public Square.* The CochlScene dataset provides 13 categories: *Bus, Cafe, Car, Crowded Indoor, Elevator, Kitchen, Park, Residential Area, Restaurant, Restroom, Street, Subway, and Subway Station.* However, the label sets of these datasets are not fully aligned. To maximize the overlap between the external datasets and CAS 2023 dataset, we perform a *label harmonization and selection* process. Classes that share the same or semantically similar meanings across datasets are retained for pre-training, while others are discarded. The final mapping and sample statistics are summarized in Table 1. 8 classes (*Airport, Bus, Metro, Restaurant, Shopping Mall, Public Square, Urban Park, Traffic Street*) are retained for pre-training, ensuring a domain-consistent initialization for the challenge model. To address the mismatch in label space, we remove the classification head after pre-training and retain only the feature extraction backbone during the fine-tuning stage.

3. Data Preprocessing and Feature Extraction

The raw audio recordings are first converted into log-mel spectrogram representations, which are widely used in ASC 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 2.

Table 2: Parameters for log-mel spectrogram extraction.

Parameter	**Value**
Sampling rate	44,100 Hz
FFT size	2,048
Window length	1,764
Hop length	882
Number of mel bands	64
Frequency range	50 Hz – 22,050 Hz
Window type	Hann

To further improve the robustness and generalization of the model, we incorporate several data augmentation techniques during training, as follows.

- **Rolling augmentation**: Each audio clip is circularly shifted along the time axis by a random offset uniformly sampled from 0.1 to 0.3 seconds. This preserves the spectral structure while introducing temporal diversity without altering the scene label.
- **Mixup** Zhou et al. (2021): Two spectrograms and their corresponding one-hot labels are linearly interpolated using a mixing coefficient $\lambda \sim \mathrm{Beta}(\alpha, \alpha)$ with α randomly sampled in the range [0.6, 0.8]. Mixup is applied with a probability randomly sampled from [0.3, 0.8] during training, which encourages linear behavior between training examples and improves robustness to label ambiguity.
- **SpecAugment** Park et al. (2019): Time masking and frequency masking are applied to the spectrograms by randomly zeroing out continuous time steps or frequency bins, which encourages the model to rely on more distributed and robust features.

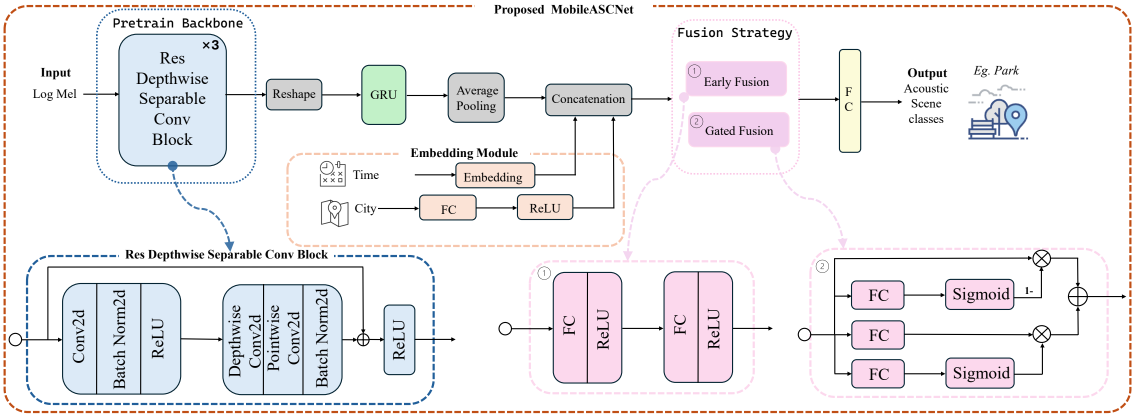

Figure 1: Overview of the proposed MobileASCNet architecture.

Note that all data augmentation techniques are applied exclusively during the fine-tuning stage on CAS 2023 dataset, and only to the training subset. The configuration follows the official ASC Challenge baseline settings to ensure comparability and reproducibility.

4. Proposed Model Architecture

4.1. Model Architecture

As illustrated in Figure 1, the proposed MobileASCNet 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. Serving as the pretrained backbone from external datasets, 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 the GRU layer are concatenated with the city and time embeddings and then passed through two alternative fusion strategies, as illustrated in Figure 1. The first strategy, *Early Fusion*, integrates acoustic and metadata features via fully connected (FC) layers. The second strategy, *Gated Fusion*, employs a learnable gating mechanism to adaptively balance the contributions of acoustic and contextual information. The fused representations are subsequently fed into the final FC classification layers to produce the acoustic scene predictions.

4.2. Training Procedure

The training procedure largely follows the official baseline setup of the APSIPA ASC 2025 Grand Challenge. The CAS 2023 dataset Bai et al. (2024) 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 using a three-stage semi-supervised learning pipeline:

1. **Initial Training or Fine-tuning:** Depending on the availability of pretrained weights, the model follows two strategies.
    - Initializes from scratch and is trained solely on the labeled portion of CAS 2023 dataset to learn discriminative acoustic features;
    - Loads a pretrained backbone from external datasets and is fine-tuned on the labeled CAS subset. In this case, we experiment with two strategies: (a) fully fine-tuning all layers, and (b) freezing the low-level convolutional blocks while only updating the high-level semantic layers.
2. **Pseudo Labeling:** The model trained in Stage 1 is used to infer pseudo labels for the unlabeled subset, generating soft targets for subsequent learning.
3. **Secondary Fine-tuning:** Both the original labeled and pseudo-labeled data are jointly used to fine-tune the model, improving its generalization across diverse spatial and temporal contexts.

While the overall training structure is consistent with the official challenge baseline, our core design focuses on architectural efficiency and contextual adaptation rather than relying on knowledge transfer from large-scale pretrained models and incorporates city and time metadata through learnable embeddings, enabling it to adapt to spatial-temporal variability in real-world acoustic scenes.

5. Experimental Results

5.1. Experimental Setup

All experiments were conducted using the CAS 2023 development dataset provided in the APSIPA ASC 2025 Grand Challenge Bai et al. (2024). The training followed the official semi-supervised three-stage protocol described in the *Training Procedure* section, with no distillation applied.

We train our model MobileASCNet for a maximum of 100 epochs with a batch size of 64, using the adaptive moment estimation (Adam) optimizer Kingma and Ba (2014). The initial learning rate is set to 10^{-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 Prechelt (2002) with a patience of 10 epochs based on validation accuracy. Unless otherwise specified, all experiments, including training from scratch or fine-tuning from external pretrained backbones (e.g., TAU 2020 dataset or CochlScene dataset), are conducted under the same optimization and learning configurations.

To better understand the trade-offs in model design, we conduct a comprehensive series of ablation studies, including: (1) comparisons between training from scratch and using pre-trained backbones, (2) analyses of whether freezing certain model layers affects generalization, (3) evaluation of the impact of data augmentation techniques such as Mixup, and (4) comparisons of different strategies for incorporating multi-modal metadata, including early fusion and late fusion of city and time embeddings.

Due to the unavailability of ground truth labels for the official evaluation set, all experiments were conducted on the development set, and classification accuracy was used as the primary evaluation metric.

5.2. Performance Comparison

We compare our proposed MobileASCNet with the official challenge baseline Bai et al. (2023) 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 3 shows the classification accuracy and parameters of different models on the development set.

Table 3: Classification accuracy (%) and model size of MobileASCNet under different training strategies on the development set.

Training Strategy	**Accuracy (%)**	**Params (M)**
Baseline	96.0	0.44
From scratch	97.3	**0.37**
Backbone with external set	86.8	
Pretrained backbone	96.8	
Pretrained + Freeze	96.9	
Pretrained + Mixup	96.3	
Pretrained + Early fusion	97.2	
Pretrained + Gated fusion	**97.5**	

[1] M denotes million parameters.

As shown in Table 3, our proposed model MobileASCNet achieves the highest classification accuracy of **97.5%** on the development set when using the Gated Fusion strategy, significantly outperforming the official baseline of **96.0%**, while using fewer parameters (0.37M vs. 0.44M). This demonstrates a superior balance between model complexity and recognition performance.

Across different training strategies, training the model entirely from scratch already yields a strong accuracy of 97.3%, demonstrating the effectiveness of the proposed architecture even without external supervision. Backbone with external set means the backbone trained solely on the external dataset, before any fine-tuning, achieves 86.8% accuracy. When using this pretrained backbone, the model reaches 96.8% after fine-tuning on CAS 2023 dataset, which is slightly lower than the scratch-trained variant. Further freezing the low-level layers after pretraining provides a marginal improvement, with accuracy increasing to 96.9%.

Regarding data augmentation, applying Mixup results in a small performance decline to 96.3%. On the other hand, integrating contextual metadata yields consistent improvements: early fusion leads to an accuracy of 97.2%, while the proposed Gated Fusion achieves the best result at **97.5%**, highlighting its effectiveness in leveraging both acoustic and contextual information.

Overall, the ablation results validate the effectiveness of MobileASCNet under various training paradigms, with Gated Fusion emerging as the most effective configuration under the semi-supervised setting.

6. Conclusions

In this work, we proposed **MobileASCNet**, a lightweight yet effective CRNN-based architecture for ASC under spatial-temporal variability, developed as part of the APSIPA ASC 2025 Grand Challenge. We conducted comparisons across multiple training strategies, including different pretraining schemes, data augmentation methods, and fusion approaches. Experimental results on the official development set demonstrate that the proposed MobileASCNet achieves a classification accuracy of 97.5%, outperforming the official baseline with fewer parameters without relying on knowledge distillation.

References

Jisheng Bai, Jianfeng Chen, Mou Wang, Muhammad Saad Ayub, and Qingli Yan. A squeeze-and-excitation and transformer-based cross-task model for environmental sound recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 15(3):1501–1513, 2023.

Jisheng Bai, Mou Wang, Haohe Liu, Han Yin, Yafei Jia, Siwei Huang, Yutong Du, Dongzhe Zhang, Dongyuan Shi, Woon-Seng Gan, Mark D. Plumbley, Susanto Rahardja, Bin Xiang, and Jianfeng Chen. Description on ieee icme 2024 grand challenge: Semi-supervised acoustic scene classification under domain shift, 2024.

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.*, 23(3):512–529, 2015.

Biyun Ding, Tao Zhang, Chao Wang, Ganjun Liu, Jinhua Liang, Ruimin Hu, Yulin Wu, and Difei Guo. Acoustic scene classification: A comprehensive survey. *Expert Systems with Applications*, 238:121902, 2024.

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

Kristóf Horváth, Harsh Purohit, Yohei Kawaguchi, Ryo Tanabe, Kota Dohi, Takashi Endo, Masaaki Yamamoto, and Tomoya Nishida. Using arcface metric learning for low-complexity acoustic scene classification. Technical report, DCASE2021 Challenge, June 2021.

Ami Igarashi, Keisuke Imoto, Yuka Komatsu, Shunsuke Tsubaki, Shuto Hario, and Tatsuya Komatsu. How information on acoustic scenes and sound events mutually benefits event detection and scene classification tasks. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 7–11. IEEE, 2022.

Il-Young Jeong and Jeongsoo Park. Cochlscene: Acquisition of acoustic scene data using crowdsourcing. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 17–21. IEEE, 2022.

Byeonggeun Kim, Seunghan Yang, Jangho Kim, and Simyung Chang. QTI submission to DCASE 2021: Residual normalization for device-imbalanced acoustic scene classification with efficient design. Technical report, DCASE2021 Challenge, June 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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.

Haowen Li, Zhengding Luo, Dongyuan Shi, Boxiang Wang, Junwei Ji, Ziyi Yang, and Woon-Seng Gan. Doa estimation with lightweight network on llm-aided simulated acoustic scenes, 2025a. URL https://arxiv.org/abs/2511.08012.

Haowen Li, Ziyi Yang, Mou Wang, Ee-Leng Tan, Junwei Yeow, Santi Peksi, and Woon-Seng Gan. Joint feature and output distillation for low-complexity acoustic scene classification. *arXiv preprint arXiv:2507.19557*, 2025b.

Annamaria Mesaros, Romain Serizel, Toni Heittola, Tuomas Virtanen, and Mark 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)*, pages 1–5. IEEE, 2025.

B. Murauer and B. Schuller. Efficient acoustic scene classification with cp-mobile. Tech. rep., DCASE2023 Challenge, 2023.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*, interspeech_2019. ISCA, September 2019. URL http://dx.doi.org/10.21437/Interspeech.2019-2680.

Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 2002.

Xinyuan Qian, Jiaran Gao, Yaodan Zhang, Qiquan Zhang, Hexin Liu, Leibny Paola Garcia, and Haizhou Li. Sav-se: Scene-aware audio-visual speech enhancement with selective state space model. *IEEE Journal of Selected Topics in Signal Processing*, 2025.

C. Schmid and et al. Efficient teacher-student training for acoustic scene classification using passt. Tech. rep., DCASE2023 Challenge, 2023.

Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D Plumbley. Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, 17(10):1733–1746, 2015.

Ee-Leng Tan, Jun Wei Yeow, Santi Peksi, Haowen Li, Ziyi Yang, and Woon-Seng Gan. Sntl-ntu dcase25 submission: Acoustic scene classification using CNN-GRU model without knowledge distillation. Technical report, DCASE2025 Challenge, May 2025.

Yizhou Tan, Haojun Ai, Shengchen Li, and Mark D Plumbley. Acoustic scene classification across cities and devices via feature disentanglement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1286–1297, 2024.

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

Kaichun Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations (ICLR)*, 2021.