ADAPTF-SEPNET: AUDIOSET-DRIVEN ADAPTIVE PRE-TRAINING OF TF-SEPNET FOR MULTI-DEVICE ACOUSTIC SCENE CLASSIFICATION

Technical Report

Ziyang Zhou[1]*, Zeyu Yin*[1]*, Yiqiang Cai*[1]*, Shengchen Li*[1]*, Xi Shao*[2]

[1] Xi'an Jiaotong-Liverpool University, School of Advanced Technology, Suzhou, China,
Ziyang.Zhou22,Zeyu.Yin22,Yiqiang.Cai21@student.xjtlu.edu.cn, Shengchen.Li@xjtlu.edu.cn
[2] Nanjing University of Posts and Telecommunications,
College of Telecommunications and Information Engineering, Nanjing, China,
shaoxi@njupt.edu.cn

ABSTRACT

This technical report presents our submission to DCASE 2025 Challenge Task 1: Low-Complexity Acoustic Scene Classification with Device Information. We propose a multi-device framework that leverages device-specific models trained with knowledge distillation techniques and enhanced through AudioSet pre-training. Our approach utilizes TF-SepNet as the backbone architecture, pre-trained on the large-scale AudioSet dataset to learn robust acoustic representations. For each of the known devices, a dedicated model is trained. At inference time, the system identifies the device source of the audio clip and selects the corresponding pre-trained model for classification. Evaluated on the test set, our device-specific system achieves an overall accuracy of 59.5%.

Index Terms— Acoustic scene classification, low-complexity models, device-specific adaptation, knowledge distillation, TF-SepNet, AudioSet pre-training, transfer learning

1. INTRODUCTION

Acoustic Scene Classification (ASC) [1] is a fundamental task in computational audio analysis that aims to classify audio recordings into predefined environmental scene categories such as "metro station," "urban park," or "public square." This task has gained significant attention due to its wide range of applications including context-aware mobile devices, intelligent monitoring systems, and audio content analysis for multimedia applications. The annual Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge [2] has been instrumental in advancing ASC research, progressively introducing more challenging and realistic scenarios that reflect real-world deployment conditions. The DCASE 2025 Challenge Task 1 represents a significant evolution from previous editions, building upon the foundations established in DCASE 2022-2024 while introducing novel challenges that better reflect practical deployment scenarios.

The most significant change in this year's challenge is the availability of recording device information at inference time. Unlike previous editions where device information was withheld during evaluation, encouraging the development of device-agnostic systems, DCASE 2025 provides device IDs for recordings in the evaluation set. This paradigm shift enables participants to develop device-specific models that can leverage device characteristics to improve classification performance, reflecting real-world scenarios where the target deployment device is known. The dataset consists of recordings from multiple devices including real devices (A: Soundman OKM II Klassik, B: Samsung Galaxy S7, C: iPhone SE, D: GoPro Hero5 Session) and simulated devices (S1-S10) created through impulse response convolution and dynamic range compression. The development set contains data from devices A, B, C, and S1-S6, while the evaluation set introduces unknown devices (D, S7-S10) marked with "unknown" IDs, maintaining the challenge of generalizing to unseen recording conditions.

Our approach addresses these challenges through ADAPTF-SepNet, a comprehensive framework that combines AudioSet pre-training with device-specific adaptation strategies. We leverage the TF-SepNet architecture as our backbone and enhance it with transfer learning from the large-scale AudioSet dataset, enabling the model to learn rich acoustic representations before fine-tuning on the target ASC task. The device information is exploited through specialized model adaptation, where separate models are trained for each known device while maintaining a general model for unknown devices. The remainder of this report details our methodology, experimental setup, and results, demonstrating how transfer learning and device-specific adaptation can be effectively combined to achieve competitive performance under strict complexity constraints.

2. METHODOLOGY

2.1. System Architecture

Our system is built around the **TF-SepNet** [3] architecture, a lightweight CNN designed for efficient acoustic scene classification. The key components include:

- **Backbone Network:** We adopt TF-SepNet (Time-Frequency Separate Network) as our core model architecture. TF-SepNet employs separate convolutions for time and frequency dimensions, enabling efficient capture of both temporal dynamics and spectral characteristics in acoustic scenes. The architecture consists of 17 layers with Time-Frequency Separate Convolution blocks, using 64 base channels and processing single-channel mel-spectrograms with 512 frequency bins. This design provides an optimal balance between model complexity and computational efficiency for mobile deployment scenarios.
- **Pre-trained Feature Extractor:** AudioSet pre-trained

`TF-SepNet` to provide stronger time-frequency feature priors.

- **Feature Extraction:** CpMel spectrograms with `512 mel bins` and `32kHz` sampling rate, providing rich spectral-temporal resolution.
- **Device-Specific Adaptation:** Separate models are trained for each known device (e.g., `a`, `s1`, `s2`) to mitigate device mismatch issues.
- **Knowledge Distillation:** A teacher-student framework is employed, where the teacher model (BEATs) guides the student (TF-SepNet) to achieve enhanced performance through soft label supervision.

2.2. Data Augmentation

Data augmentation plays a pivotal role in acoustic scene classification, particularly when confronted with limited labeled training data as imposed by the 25% subset constraint. Our framework incorporates a comprehensive augmentation pipeline comprising three complementary techniques: Device Impulse Response (DIR) augmentation, Frequency-wise MixStyle, and Soft Mixup. These augmentation methods are designed to be modular and can be seamlessly integrated during the training process.

Device Impulse Response (DIR) Augmentation This technique addresses the domain gap between different recording devices by simulating cross-device recordings. We convolve the input waveform with randomly selected device impulse responses from the MicIRP dataset [4]. Given an input waveform $x(t)$ and an impulse response $h(t)$, the augmented signal is computed as:

$$x_{\text{aug}}(t) = x(t) * h(t) \tag{1}$$

where $*$ denotes convolution. The application probability is controlled by hyperparameter p_{dir}, allowing for fine-tuned control over augmentation intensity.

Frequency-wise MixStyle Adapted from the original MixStyle approach [5], this method performs domain randomization in the frequency domain rather than across feature channels. Freq-MixStyle [6] normalizes spectral components across frequency bands, effectively mitigating device-induced domain shifts. For a spectrogram $S(f, t)$ with frequency bins f and time frames t, the normalization is applied frequency-wise to enhance cross-device generalization.

Soft Mixup Building upon the conventional Mixup strategy [7], our Soft Mixup variant extends the linear interpolation to incorporate both hard labels and soft teacher predictions. For two training samples $(x_i, y_i, \tilde{y}_i)$ and $(x_j, y_j, \tilde{y}_j)$, where x represents input features, y denotes ground truth labels, and $\tilde{y}$ contains teacher logits, the augmented sample is generated as:

$$x_{\text{mix}} = \lambda x_i + (1 - \lambda) x_j \tag{2}$$

$$y_{\text{mix}} = \lambda y_i + (1 - \lambda) y_j \tag{3}$$

$$\tilde{y}_{\text{mix}} = \lambda \tilde{y}_i + (1 - \lambda) \tilde{y}_j \tag{4}$$

where $\lambda \sim \text{Beta}(\alpha, \alpha)$ is sampled from a Beta distribution with $\alpha = 0.2$.

3. TRAINING SETUP

Our training methodology employs a four-stage progressive learning approach: AudioSet pre-training, base model training, knowledge distillation, and device-specific fine-tuning.

3.1. Stage 1: AudioSet Pre-training

We first pre-train TF-SepNet on AudioSet's 527 classes using balanced training segments. The model uses single-channel input, 64 base channels, depth 17, and 512 mel-frequency bins. Training employs Adam optimizer (lr=0.004), CosineAnnealingWarmRestarts ($T_0 = 10$, $T_{mult} = 2$), batch size 128, and 300 epochs. Only DIR augmentation is applied (p=0.4) to simulate diverse acoustic environments.

3.2. Stage 2: Base Model Training

The pre-trained model is fine-tuned on DCASE Task 1 (10 acoustic scenes) using the TAU Urban Acoustic Scenes 2022 Mobile Development dataset. We initialize from the best AudioSet checkpoint with `load_classifier: false`. Training configuration: Adam optimizer (lr=0.004), batch size 256, 200 epochs, monitoring validation accuracy.

Data Augmentation: We apply three augmentation techniques: (1) MixUp ($\alpha = 0.3$), (2) Freq-MixStyle ($\alpha = 0.4$, p=0.8), and (3) DIR augmentation (p=0.4).

3.3. Stage 3: Knowledge Distillation

We employ teacher-student learning using an ensemble of three BEATS models (SSL, SSL+SL, SSL*) as teachers. The student TF-SepNet is initialized from the best base model (epoch 59, acc=0.5619). Knowledge distillation parameters follow DCASE2023 methodology: temperature $\tau = 2.0$, distillation weight $\lambda = 0.02$. SoftMixUp replaces standard MixUp to preserve teacher knowledge. Training uses 300 epochs with identical hyperparameters as Stage 2.

3.4. Stage 4: Device-Specific Fine-tuning

Individual models are trained for each device (a, b, c, s1, s2, s3, unknown) using `DCASEDataModuleByDevice`. Each model initializes from the knowledge distillation checkpoint with `load_classifier: true`. Training configuration: reduced learning rate (0.0008), extended epochs (400), disabled DIR augmentation for device specificity.

3.5. Post-Training Optimization

Post-training static quantization to INT8 precision is applied using Intel Neural Compressor with maximum 1% accuracy loss tolerance. A multi-device inference system (`LitMultiDeviceInference`) automatically selects device-specific models during deployment.

3.6. Implementation Details

Training uses PyTorch Lightning framework with TensorBoard logging, 2 data loading workers, and pin memory optimization. All models use 32kHz sampling rate and maintain consistent 512 mel-frequency bin extraction throughout the pipeline.

Table 1: Class-wise accuracy (%) comparison **with** and **without** device-specific finetuning on the DCASE2025 Task 1 evaluation set

Scene Class	Airport	Bus	Metro	Metro Station	Park	Public Square	Shopping Mall	Street Pedestrian	Street Traffic	Tram
With device-specific finetuning	48.30	78.10	53.70	50.40	75.20	49.20	66.00	34.00	76.90	63.10
Without device-specific finetuning	48.00	76.90	54.00	46.90	73.90	48.79	66.20	31.00	76.90	60.90

4. SUBMISSION AND RESULT

Table 1 presents a detailed comparison of class-wise scene classification accuracy with and without device-specific finetuning. Most acoustic scenes benefit from device-level adaptation, with notable gains observed in complex environments such as metro station, public square, and tram. The results suggest that device-aware training improves robustness in noisy or overlapping acoustic environments, while classes like street traffic maintain high performance regardless of the setting due to their strong acoustic signatures.

Table 2: Performance evaluation on different devices with and without finetuning the specific device

Setting	a	b	c	s1	s2	s3	s4	s5	s6	Overall
With	0.676	0.603	0.619	0.583	0.568	0.599	0.570	0.586	0.532	0.595
Without	0.655	0.583	0.612	0.579	0.553	0.598	0.582	0.568	0.521	0.583

Table 2 compares classification accuracy across different devices with and without device-specific finetuning. The results show consistent improvements across nearly all devices when models are individually finetuned for each device, demonstrating the effectiveness of adapting models to device-specific signal characteristics. During the device-specific finetuning stage, we applied DIR augmentation (p=0.8) to enhance performance on unknown devices. The results reveal mixed outcomes: while simulated devices s4 and s6 show improved performance, s6 exhibits reduced accuracy, indicating that device variability continues to present challenges when generalizing to certain unseen domains.

5. CONCLUSION

In this work, we present a comprehensive four-stage progressive training methodology for DCASE 2025 Task 1 that effectively addresses the challenges of device domain mismatch and limited labeled data in acoustic scene classification. Our approach combines AudioSet pre-training, comprehensive data augmentation (MixUp, Freq-MixStyle, DIR augmentation), ensemble knowledge distillation using multiple BEATS teacher models, and device-specific fine-tuning to achieve robust performance across diverse recording conditions. The TF-SepNet backbone with time-frequency separate convolutions efficiently captures both temporal and spectral patterns, while the device-aware inference system ensures optimal performance through automatic model routing. Experimental results demonstrate the effectiveness of each training stage, with notable improvements from knowledge distillation and device adaptation, validating our progressive learning strategy for real-world acoustic scene classification applications.

6. 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] F. Schmid, P. Primus, T. Heittola, A. Mesaros, I. Martín-Morató, and G. Widmer, "Low-complexity acoustic scene classification with device information in the dcase 2025 challenge," *arXiv preprint arXiv:2505.01747*, 2025.

[3] Y. Cai, M. Lin, S. Li, and X. Shao, "Dcase2024 task1 submission: Data-efficient acoustic scene classification with self-supervised teachers," DCASE Challenge, Tech. Rep, Tech. Rep., 2024.

[4] T. Morocutti, F. Schmid, and G. Widmer, "Device simulation with impulse responses for the device shift problem in acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023)*, 2023, pp. 166–170.

[5] K. Zhou, Y. Yang, T. Hospedales, and T. Xiang, "Domain generalization with mixstyle," in *International Conference on Learning Representations*, 2021.

[6] F. Schmid, K. Koutini, and G. Widmer, "Efficient large-scale audio tagging via transformer-to-cnn knowledge distillation," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, 2022, pp. 62–66.

[7] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.