JOINT FEATURE AND OUTPUT DISTILLATION FOR LOW-COMPLEXITY ACOUSTIC SCENE CLASSIFICATION

Technical Report

Haowen Li[1], *Ziyi Yang*[1], *Mou Wang*[2], *Ee-Leng Tan*[1], *Junwei Yeow*[1],
Santi Peksi[1], *Woon-Seng Gan*[1],

[1] Smart Nation TRANS Lab, Nanyang Technological University, Singapore
[2] Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
haowen.li@ntu.edu.sg, ziyi016@e.ntu.edu.sg, wangmou21@mail.nwpu.edu.cn, etanel@ntu.edu.sg,
junwei004@e.ntu.edu.sg, speksi@ntu.edu.sg, ewsgan@ntu.edu.sg

ABSTRACT

This report presents a dual-level knowledge distillation framework with multi-teacher guidance for low-complexity acoustic scene classification (ASC) in DCASE2025 Task 1. We propose a distillation strategy that jointly transfers both soft logits and intermediate feature representations. Specifically, we pre-trained PaSST and CP-ResNet models as teacher models. Logits from teachers are averaged to generate soft targets, while one CP-ResNet is selected for feature-level distillation. This enables the compact student model (CP-Mobile) to capture both semantic distribution and structural information from teacher guidance. Experiments on the TAU Urban Acoustic Scenes 2022 Mobile dataset (development set) demonstrate that our submitted systems achieve up to 59.30% accuracy.[1]

Index Terms— Acoustic Scene Classification, Knowledge Distillation, Data Augmentation, Feature Distillation

1. INTRODUCTION

Acoustic Scene Classification (ASC) aims to identify the environment in which an audio recording was captured, such as a street, shopping mall, or park, based on its acoustic characteristics [1, 2]. The DCASE 2025 Challenge Task 1 focuses on developing low-complexity ASC models that are robust to domain shifts across mobile recording devices and diverse urban environments. The challenge emphasizes generalization across devices under strict constraints on model size and computational cost. The task uses the TAU Urban Acoustic Scenes 2022 Mobile dataset [3], which contains approximately 64 hours of audio recordings under 10 acoustic scenes. Compared with previous editions, the 2025 challenge emphasizes on robustness to unseen-device conditions and data efficiency. Specifically, models must be trained using only a 25% subset of the official training set. In addition, submitted systems must not exceed 128kB parameter memory and 30M multiply-accumulate operations (MACs) per inference pass.

To address these challenges, knowledge distillation (KD) has been widely adopted to train compact student networks under the supervision of larger teacher networks [4, 5]. KD was first introduced by Hinton et al.[6] as a technique to compress large models into smaller ones by transferring soft target distributions. Since then, KD has evolved into a general learning paradigm with various forms of knowledge, including output logits, intermediate features, attention maps to relational structures [7]. In DCASE challenges, prior work has largely focused on output-level distillation using logits, CPJKU's submission in DCASE2023 achieved strong performance under low-complexity constraints [8]. However, few studies have explored feature-level supervision, which has been shown to offer additional benefits in general deep learning settings [9, 10].

Models such as the Patchout Spectrogram Transformer (PaSST) [11] and Convolutional Patch-ResNet (CP-ResNet) [12] have demonstrated strong performance in ASC and served as effective teachers for compact CNN-based architectures [8], and early investigations into model frameworks for ASC [13, 14] have also offered guidance for this work.

In this work, we employ a dual-level knowledge distillation framework that combines output-level and feature-level supervision to improve the training of a compact CP-Mobile [15] student model. Specifically, we ensemble multiple high-performing teacher models (CP-ResNet and PaSST) to provide complementary guidance through soft target distributions. Inspired by advances in feature-based distillation such as FitNets [9] and SimKD [10], we further align intermediate representations between a designated teacher and the student network to enhance structural transfer.

This report is organized as follows. Section 2 describes the input feature extraction process and data augmentation techniques. Section 3 details the knowledge distillation framework, including student-teacher architecture, feature matching strategies, and teacher model training procedures. Section 4 reports the configuration and evaluation results of submitted systems for DCASE 2025 Task 1. Finally, Section 5 summarizes the key findings and concludes the report.

2. DATA PREPROCESSING AND AUGMENTATION

2.1. Preprocessing

All models operate on 32 kHz audio. Log-Mel spectrograms are extracted using configurations customized for each model to balance time-frequency resolution and computational efficiency.

Teacher models: We use 2 architectures as teacher models: PaSST and CP-ResNet. For each architecture, we employ different spectrogram preprocessing settings to emphasize either frequency or temporal resolution:

[1] `https://github.com/HaoWLee/dcase2025_task1_inference`

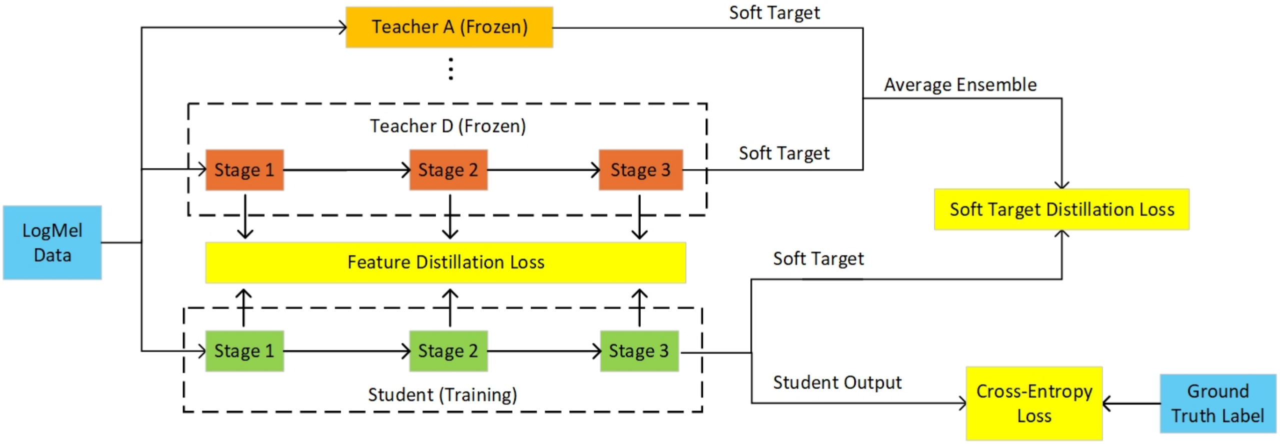

Figure 1: Overview of the joint feature and logit distillation framework. Teacher A-D provide ensembled soft targets for output-level distillation, and teacher D supervises feature-level distillation.

- **PaSST-1:** 1024-point FFT, 800-point window, 320-point hop, 128 Mel bins
- **PaSST-2:** 4096-point FFT, 800-point window, 320-point hop, 128 Mel bins
- **CP-ResNet-1:** 4096-point FFT, 3072-point window, 750-point hop, 256 Mel bins
- **CP-ResNet-2:** 4096-point FFT, 3072-point window, 500-point hop, 256 Mel bins

Student model: **CP-Mobile** [15] uses the same 32 kHz audio input and adopts a 4096-point FFT, 3072-point window, and a 500-point hop size, producing 256 Mel bins. Compared to CP-ResNet, this configuration provides improved temporal resolution, which aligns better with the receptive fields of lightweight models under complexity constraints.

2.2. Data Augmentation

To improve model generalization under limited supervision and domain shift, three augmentation strategies are employed: time-domain rolling, frequency-domain MixStyle (Freq-MixStyle), and device impulse response (DIR) convolution.

Time Roll: Input waveforms are circularly shifted along the time axis to introduce temporal variability while preserving semantic content. For example, CP-Mobile and PaSST-1 apply a shift of up to 312 ms (10,000 samples at 32 kHz), while other models adopt a shorter shift of 125 ms (4,000 samples).

Freq-MixStyle: Following the domain generalization framework MixStyle [16], we adopt a frequency-wise variant tailored to log-Mel spectrograms. With probability p, sample-wise channel statistics are interpolated using a Beta distribution with parameter α_{mix} to perturb style-related information and improve cross-device robustness. This augmentation is only applied to CP-ResNet teachers.

DIR Augmentation: To simulate the acoustic coloration introduced by different devices, waveforms are convolved with DIRs sourced from the MicIRP dataset [17]. Each training example undergoes DIR-based augmentation with a specified probability to encourage invariance to microphone and channel characteristics.

Table 1: Data augmentation configurations for teacher and student models.

Model	Time Roll	DIR Prob.	Freq-MixStyle
PaSST-1	312 ms	0.6	($\alpha_{\text{mix}} = 0.4, p = 0.4$)
PaSST-2	125 ms	0.4	($\alpha_{\text{mix}} = 0.4, p = 0.8$)
CP-ResNet-1	125 ms	0.4	($\alpha_{\text{mix}} = 0.4, p = 0.8$)
CP-ResNet-2	125 ms	0.6	($\alpha_{\text{mix}} = 0.3, p = 0.4$)
CP-Mobile	312 ms	0.6	None

The augmentation configurations for each model are summarized in Table 1.

3. TRAINING AND KNOWLEDGE DISTILLATION

3.1. Teacher Model Training

All teacher models are trained independently on the 25% subset of the TAU Urban Acoustic Scenes 2022 Mobile dataset, using their respective log-Mel spectrogram configurations detailed in Section 2. All teachers adopt Freq-MixStyle augmentation, while DIR convolution is selectively applied to improve robustness against microphone variability.

We train 4 teacher models: 2 based on PaSST and 2 on CP-ResNet. For each architecture, one variant uses a spectrogram configuration emphasizing high frequency resolution (longer FFT and window), and the other favors higher temporal resolution (shorter window and hop size). This dual-resolution setup provides complementary time-frequency perspectives for KD.

To improve teacher model generalization and stability of soft targets, we apply model soup [18] within each teacher model. Specifically, we select the top 5 checkpoints after training convergence and compute the average of their weights. This simple weight averaging strategy helps mitigate overfitting and produces more robust teacher ensembles for distillation.

To generate soft targets, we compute the mean of softmax outputs from all 4 teacher models on the training set. These ensembled logits are used to supervise the student model via output-level distillation.

3.2. Knowledge Distillation Framework

We adopt a joint knowledge distillation framework that integrates both output-level and feature-level supervision, as illustrated in Fig. 1. The objective is to transfer knowledge from multiple high-capacity teacher models to a compact student network.

Soft target distillation transfers knowledge from teacher outputs. Given teacher logits z_t and student logits z_s, we compute softened probability distributions using a temperature scaling factor T. The soft target loss is defined via the Kullback-Leibler divergence:

$$\mathcal{L}_{\text{soft}} = T^2 \cdot \mathrm{KL}\left(\mathrm{softmax}\left(\frac{z_s}{T}\right), \|, \mathrm{softmax}\left(\frac{z_t}{T}\right)\right). \quad (1)$$

Feature-level distillation enforces alignment between intermediate representations, using either direct activation matching or self-similarity alignment (detailed in Section 3.3). The corresponding feature loss is denoted as $\mathcal{L}_{\text{feat}}$.

Cross-entropy loss is applied between the student prediction z_s and the ground-truth label y:

$$\mathcal{L}_{\text{ce}} = \mathrm{CE}\left(\mathrm{softmax}(z_s), , y\right). \quad (2)$$

The final objective is a weighted sum of the three components:

$$\mathcal{L}\text{student} = \alpha \cdot \mathcal{L}\text{soft} + \beta \cdot \mathcal{L}\text{feat} + \gamma \cdot \mathcal{L}\text{ce}, \quad (3)$$

where α, β, and γ are the respective weights for soft-target, feature, and cross-entropy losses. Unless otherwise stated, we set $T = 2$, $\alpha = 1.0$, $\beta = 0.1$, and $\gamma = 0.05$ in our experiments.

3.3. Feature Projection

To bridge architectural differences between teacher and student models, we explore two distinct feature matching strategies for intermediate layer distillation:
Direct Feature Matching (DFM) aligns feature maps from teacher and student directly in the activation space [9]. Specifically, we select intermediate feature maps with similar spatial dimensions and use $1{\times}1$ convolutional adapters to match channel dimensions. For example, features after the second residual block of CP-ResNet are mapped to early-stage outputs of CP-Mobile. The matching is supervised using the mean squared error (MSE) loss:

$$\mathcal{L}_{\text{feat}}^{\text{DFM}} = \left\| f_s - \text{Adapter}(f_t) \right\|_2^2. \quad (4)$$

This approach is inspired by conventional intermediate feature matching frameworks such as FitNets [9].

Self-Similarity Feature Matching (SSFM), adopts a self-similarity based distillation method originally proposed for speech enhancement tasks in [19], computes the time-frequency self-similarity Gram matrices of intermediate features for each input and minimizes the discrepancy between teacher and student similarity structures:

$$\mathcal{L}_{\text{feat}}^{\text{SSFM}} = \left\| G(f_s) - G(f_t) \right\|_2^2, \quad (5)$$

where $G(\cdot)$ denotes the Gram matrix capturing internal correlation across time-frequency bins.

The feature layers are manually selected based on spatial alignment and semantic consistency between teacher and student networks. We do not employ any dynamic attention or automated feature matching mechanisms in this study.

Table 2: Detailed configuration of submissions. DFM: Direct Feature Matching, SSFM: Self-Similarity Feature Matching.

Submission	**S1**	**S2**
Feature KD Method	SSFM	DFM
Feature KD Stages	Stage 1–3	Stage 3
Feature KD Teacher	CP-ResNet	CP-ResNet
Output KD Teacher	2×PaSST + 2×CP-ResNet	
Student	CP-Mobile	
Total Params	61.16 K	
MACs	17.05 M	
Accuracy (%)	58.80	59.30

4. SUBMISSIONS AND RESULTS

We submitted systems S1 (Li_NTU_task1_1) and S2 (Li_NTU_task1_2) to the DCASE 2025 Task 1 evaluation. Both systems adopt CP-Mobile as the student model, trained under the proposed dual-level distillation framework. The submissions differ in their feature-level distillation strategies: S1 uses DFM, while S2 applies SSFM.

Table 2 summarizes the configuration of the submitted systems, including the feature distillation method, KD setup, and model complexity. The CP-Mobile student model used in both systems contains only 61,160 parameters and 17.05M MACs. All inference was performed using float16 precision, resulting in reduced memory usage and faster computation. The overall system design satisfies the DCASE 2025 Task 1 constraints (128kB parameter memory and 30M MACs).

S1 and S2 achieved 58.80% and 59.30% accuracy, respectively, indicating the potential of dual-level distillation under low-resource constraints.

5. CONCLUSIONS

In this report, we propose a dual-level knowledge distillation framework for low-complexity acoustic scene classification, which integrates output-level supervision from an ensemble of teacher models with intermediate feature-level guidance. To facilitate effective feature knowledge transfer, we investigate two distinct strategies: Direct Feature Matching and Self Similarity Feature Matching, with CP-ResNet employed as the feature-level teacher. All models are trained on a constrained 25% subset of the TAU Urban Acoustic Scenes 2022 Mobile dataset. Under this setting, our submission system achieves an accuracy of up to 59.30% on the official development set.

6. 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] 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.

[4] J. Bai, M. Wang, E.-L. Tan, J. J. S. Yeo, J. W. Yeow, S. Peksi, D. Shi, W.-S. Gan, and J. Chen, "Hierarchical acoustic scene classification with knowledge distillation and pre-trained dynamic networks," DCASE2024 Challenge, Tech. Rep., May 2024, technical Report.

[5] S. Yeo, E.-L. Tan, J. Bai, S. Peksi, and W.-S. Gan, "Data-efficient acoustic scene classification using teacher-informed confusing class instruction," DCASE2024 Challenge, Tech. Rep., May 2024, technical Report.

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

[7] J. Gou, B. Yu, S. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, pp. 1789–1819, 2021.

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

[9] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," in *Proc. ICLR*, 2015.

[10] M. Li, T. Wu, L. Xie, X. Jin, D. Liang, L. Chen, Y. Zhan, and G. Song, "Simkd: Semantic similarity-aware representation for knowledge distillation," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 30 157–30 169, 2021.

[11] 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.

[12] 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.

[13] M. Wang and R. Wang, "Ciaic-ASC system for DCASE 2019 challenge task1," DCASE2019 Challenge, Tech. Rep., June 2019, technical Report.

[14] J. Bai, E.-L. Tan, and W.-S. Gan, "Acoustic scene classification using multi-scale features and multi-level predictions," DCASE2018 Challenge, Tech. Rep., November 2018. [Online]. Available: https://dcase.community/documents/challenge2018/technical_reports/DCASE2018_Bai_998.pdf

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

[16] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with mixstyle," in *International Conference on Learning Representations (ICLR)*, 2021.

[17] "Micirp: Microphone impulse response project," https://micirp.blogspot.com/, accessed: 2025-06-13.

[18] M. Wortsman, G. Ilharco, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, J. Shlens, L. Yatziv, D. Ganguli, *et al.*, "Model soup: Aggregating models to improve few-shot generalization," in *International Conference on Machine Learning*, 2022, pp. 23 965–23 982.

[19] R. D. Nathoo, M. Kegler, and M. Stamenovic, "Two-step knowledge distillation for tiny speech enhancement," in *ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 141–10 145.