CONFIDENCE-AWARE ENSEMBLE KNOWLEDGE DISTILLATION FOR LOW-COMPLEXITY ACOUSTIC SCENE CLASSIFICATION

Technical Report

Sarang Han, Dong Ho Lee*, Min Sik Jo,*
Eun Seo Ha, Min Ju Chae, Geon Woo Lee†,

Department of AI Software, Chosun University, Gwangju 61452, Republic of Korea
6002tkfkd@chosun.ac.kr, gghj0201@chosun.kr,
{mswd81, murru8989, minju9642, geonwoo}@chosun.ac.kr

ABSTRACT

We propose a confidence-aware ensemble knowledge distillation method for acoustic scene classification under low-complexity and limited-data settings. Our approach utilizes heterogeneous teacher models—BEATs, and EfficientAT—fine-tuned on the DCASE 2025 Task 1 dataset, to guide the training of a lightweight student model, TFSepNet. To improve over naive ensemble distillation, we introduce a confidence-weighted strategy that emphasizes reliable teacher outputs. Experimental results show improved generalization on unseen devices and domains, outperforming single-teacher and uniform ensemble baselines.

Index Terms— Acoustic Scene Classification (ASC), Knowledge Distillation (KD), Ensemble Learning, Confidence Weighting, Lightweight Neural Networks

1. INTRODUCTION

Acoustic Scene Classification (ASC) aims to automatically infer the surrounding environmental context, such as subway stations, parks, and shopping malls, from real-world audio signals [1]. ASC has emerged as a key enabler for smart city infrastructure, public safety, and context-aware services [2, 3]. For deployment in practical scenarios, ASC models must operate with minimal latency and computational overhead on resource-constrained platforms, including mobile devices, IoT nodes, and edge computing systems [1].

To address these deployment-oriented constraints, benchmark initiatives such as the DCASE Challenge have introduced ASC tasks under limited-resource conditions, utilizing datasets such as TAU Urban Acoustic Scenes. These data sets reflect real-world challenges including device heterogeneity, class imbalance, and domain mismatch, particularly through evaluation on unseen devices, thus promoting the development of models with robust generalization capability.

Traditional studies have used a variety of model compression techniques, such as compact CNN architectures [4, 5], pruning, quantization, and knowledge distillation (KD) [6] to reduce model complexity. Among these, KD has emerged as a scalable and effective framework for transferring high-level semantic representations from large-scale teacher models to compact student architectures [7], enabling efficient performance without significantly increasing model size or inference cost [8].

*These authors contributed equally to this work.
†Corresponding author.

However, conventional KD approaches often rely on a single-teacher model, which can introduce inductive bias and degrade generalization to mismatched domains or unseen conditions. To mitigate this limitation, ensemble KD approaches have been explored, in which multiple teacher models are used to provide diverse and complementary knowledge. However, naive averaging of output logits can obscure class boundaries due to inconsistencies across teacher predictions.

In this work, we propose a confidence-aware ensemble KD framework for ASC. Specifically, we adopt two heterogeneous teacher models—BEATs and EfficientAT—each pretrained and fine-tuned on the DCASE 2025 dataset. A lightweight student model, TFSepNet, is trained using a selective distillation process, where soft predictions from teachers are aggregated using a confidence-based weighting mechanism. This design enables the student to prioritize reliable and informative outputs, thereby enhancing generalization across device types and acoustic domains. Furthermore, we extend TFSepNet with an attention module conditioned on embedded device information, enabling the model to explicitly capture and infer device-specific characteristics. The experimental results on the DCASE 2025 Task 1 development set confirm that the proposed approach offers improved performance under constrained training conditions and heterogeneous deployment scenarios.

2. PROPOSED METHOD

This section outlines the overall architecture and training methodology for ASC under low-complexity constraints. The proposed system enables the transfer of semantically rich and complementary acoustic representations from multiple large-scale pretrained teacher models into a tiny student network via a confidence-aware ensemble knowledge distillation approaches.

The system consists of two key components: (1) a set of heterogeneous teacher models—each individually fine-tuned on the ASC task—to provide diverse supervisory cues with distinct inductive biases; and (2) a confidence-weighted distillation mechanism that adaptively aggregates the teachers' outputs based on their predictive reliability, thereby allowing the student model to selectively capture high-quality information during training.

Fig. 1 Fine-tuning pipeline for teacher models and the confidence-aware training strategy for the student model. The proposed framework leverages shared log-Mel spectrograms and soft targets derived from multiple fine-tuned teachers. This design

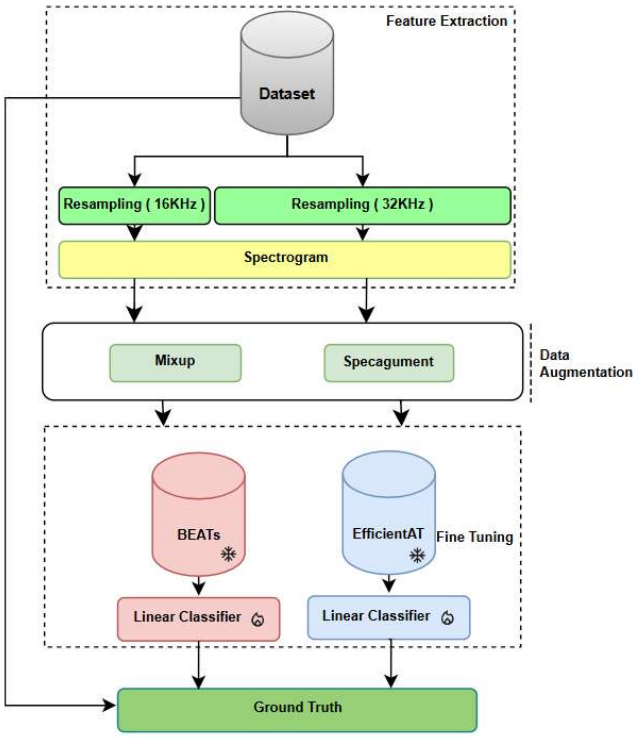

Figure 1: Fine-tuning pipeline for teacher models.

enables efficient knowledge transfer under data-constrained conditions, while maintaining competitive classification performance and low computational overhead. Such characteristics allow the proposed system suitable for real-world deployment on resource-limited platforms, including mobile and edge devices.

2.1. Feature Extraction

The input to the system comprises 1-second mono audio clips originally sampled at 44.1 kHz. To ensure compatibility with the architectural requirements of each model, the waveform is resampled to model-specific sampling rates: 16 kHz for BEATs and the student model (TFSepNet), and 32 kHz for EfficientAT. This pre-processing step ensures temporal-spectral consistency while aligning with the internal resolution of each model's feature extraction pipeline.

The resampled waveform is subsequently transformed into a log-Mel spectrogram via a Short-Time Fourier Transform (STFT), computed using a 40 ms window length, 20 ms hop size, and a 2048-point FFT. A bank of 40 Mel-scale filters is then applied to capture perceptually salient time-frequency characteristics essential for acoustic scene understanding.

These log-Mel spectrograms serve as the shared input representation across all teacher and student models within the distillation processing. Specifically, both BEATs and EfficientAT are fine-tuned on the DCASE 2025 Task 1 dataset using these inputs, allowing each model to adapt its high-level representations to the target task. The same spectrogram format is employed to train the student model, TFSepNet, ensuring consistent input semantics across the system.

Following fine-tuning, each teacher generates temperature-scaled soft predictions for every training dataset. These soft labels are then aggregated through the proposed confidence-aware ensemble distillation approach, which selectively emphasizes predictions with higher reliability based on a confidence score. This selective transfer facilitates efficient and robust learning of the student model, particularly under data-scarce and domain-mismatched conditions.

2.2. Fine-Tuning for Teacher Models

The proposed framework employs two heterogeneous teacher models—BEATs and EfficientAT—each selected to introduce distinct architectural priors and inductive biases. BEATs is transformer-based model pretrained on an imbalanced 2M-sample subset of AudioSet, whereas EfficientAT is a convolutional model trained on the balanced 20K-sample version of the same corpus.

To align the pretrained representations with the target ASC task, each teacher model is fine-tuned on the DCASE 2025 Task 1 dataset. During this process, only the classification head is updated while the feature extraction backbone remains frozen. This design enables effective domain adaptation while minimizing overfitting, thereby preserving the general-purpose acoustic features acquired during large-scale pretraining.

As shown in Fig. 1, the fine-tuning process includes pre-processing steps such as resampling and spectrogram generation, followed by data augmentation to improve robustness. After fine-tuning, each teacher outputs soft class probabilities scaled by a temperature factor. These soft predictions are then aggregated using a confidence-based mechanism, wherein outputs from more reliable teachers are emphasized.

The resulting ensembled soft labels are utilized as supervisory signals for training the student model. This selective distillation process forms the core of the proposed confidence-aware ensemble KD framework.

2.3. Device-Information Module

To incorporate device-specific information into the student model, TFSepNet is augmented with an attention-based device-information module. This device-info module takes as input a categorical indicator of the recording device, comprising a total of 10 classes—including 9 known device types and an additional unknown class. These device classes are first mapped into continuous representations via an embedding layer, which is subsequently fed into a self-attention layer.

The output of the self-attention layer is a feature vector of dimension 1024, aligned with the frequency axis of the log-Mel spectrogram. This vector is broadcasted across all time frames and element-wise multiplied with the spectrogram, thereby modulating spectral representations based on device characteristics. The resulting vector can be interpreted as a learnable proxy for device-specific impulse response, enabling the model to adapt its internal representation to device-induced variations. Both the embedding and attention layers are jointly optimized during training.

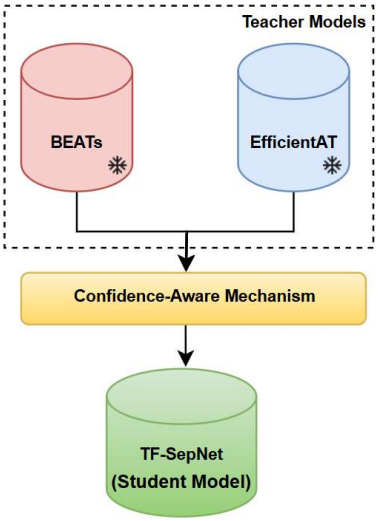

Figure 2: Training pipeline of the student model using fine-tuned teacher models.

2.4. Student Model Training

The student model, TFSepNet [4], augmented with the device-information module, is trained using a confidence-aware ensemble knowledge distillation. Instead of uniformly averaging teacher outputs, the proposed method assigns confidence-based adaptive weights to each teacher's prediction, allowing the student to selectively emphasize more reliable supervisory signals during training.

As depicted in Fig. 2, soft predictions are generated by fixed teacher models—BEATs and EfficientAT—whose parameters remain frozen throughout the distillation phase. These confidence-weighted soft targets serve as supervisory signals for training the student model in a fully supervised manner. This selective distillation mechanism enables the student to suppress noisy or ambiguous outputs and effectively integrate complementary knowledge from multiple heterogeneous teachers.

The TFSepNet with device-information model is selected as the student for its favorable trade-off between computational efficiency and classification performance, rendering it well-suited for low-resource ASC scenarios. The resulting model demonstrates robust generalization across mismatched acoustic conditions and diverse device types, even under limited training data and stringent computational constraints.

3. RESULT

3.1. Experimental Setup

3.1.1. Dataset

DCASE 2025 Challenge Task 1 [13] is built upon the TAU Urban Acoustic Scenes 2022 Mobile dataset, previously used in the 2022–2024 challenges. It consists of 1-second, single-channel, 24-bit audio clips sampled at 44.1 kHz, spanning 10 acoustic scene categories. To simulate multi-device conditions, impulse responses from real devices are convolved with audio recorded using device A, generating synthetic devices S1–S10 that reflect the spatial and compression characteristics of real hardware.

The dataset is partitioned into a development set and an evaluation set. The development set includes 64 hours of audio from real devices (A, B, C) and simulated devices (S1–S6), with S4–S6 held out for testing to evaluate generalization. Only 25% of the development data—along with device and city metadata—is accessible for training. The evaluation set contains five unseen devices (D, S7–S10) and recordings from two previously unseen cities. Scene labels are withheld during evaluation; device IDs are available at inference, while city labels remain hidden. The known/unknown device distribution is balanced between development and evaluation phases.

3.1.2. Data Augmentation

Data augmentation plays a critical role in ASC, particularly under limited supervision. To enhance the diversity and generalizability of training data, we employ a composite augmentation pipeline that integrates MixUp, MixStyle, SpecAug, FilterAug, frame-level time shifting, and DirAug. These techniques collectively increase the robustness of the model by simulating various acoustic and channel conditions.

3.1.3. Optimization and Quantization

The student model, TFSepNet with the device-information module, is trained for 150 epochs using the AdamW optimizer with an initial learning rate of 1e-4. To improve convergence, stochastic gradient descent with warm restarts is applied throughout the training process. The batch size is set to 512. Following training, post-training static quantization is performed using the Intel Neural Compressor, converting the model weights to the INT8 data type. This quantization step significantly reduces memory footprint and inference cost, enabling efficient deployment on edge devices.

3.2. Performance Evaluation

We evaluated the performance of four systems using a consistent student architecture, TF-SepNet. The experiments varied in terms of data augmentation strategies and whether or how knowledge distillation from teacher models was applied.

System 1: The student model was trained with a comprehensive augmentation pipeline consisting of MixUp ($\alpha = 0.3$), MixStyle ($\alpha = 0.4, p = 0.8$), SpecAugmentation (mask size = 0.2, $p = 1.0$), FilterAugmentation (step filter, ± 4 dB), additive noise (SNRs: 10, 20 dB), frequency masking (1/16), time masking (1/10, 1/20), frame shift (pooling factor 2), and DIR-Aug ($p = 0.4$). No knowledge distillation was applied.

System 2: The student model was trained using the same augmentation settings as in System 1. Knowledge distillation was performed using an ensemble of five BEATs teacher models—one trained with SpecAugmentation probability $p = 0.8$ and four with $p = 0.5$.

System 3: The student model was trained using the same augmentation pipeline as in System 1, excluding time masking. Knowledge distillation was conducted using two teacher models: one BEATs model trained with SpecAugmentation $p = 0.8$, and one EfficientAT model.

System 4: The student model configuration was the same as in System 1, and the teacher ensemble was the same as in System 3.

Table 1: Performance comparison between 4 different systems

	General		Device-Specific	
	log loss	Accuracy	log loss	Accuracy
System1	1.2340	55.04	1.2286	55.06
System2	1.3182	51.30	1.3215	51.36
System3	1.2532	54.10	1.2580	54.01
System4	1.3141	51.25	1.3213	50.97

4. CONCLUSIONS

This study proposed a confidence-aware ensemble knowledge distillation framework for Acoustic Scene Classification (ASC), aimed at transferring complementary knowledge from multiple pretrained teacher models—BEATs and EfficientAT—into a compact and computationally efficient student model, TFSepNet. The framework leverages architectural diversity and inductive bias variation among teachers to enable the student to learn semantically rich and generalizable acoustic representations.

Experimental results demonstrate that each teacher model contributes distinct discriminative information, and that this diversity can be effectively exploited through ensemble distillation. The confidence-based weighting mechanism selectively emphasizes informative predictions during training, allowing the student to benefit from robust supervision without incurring additional computational overhead.

5. REFERENCES

[1] Y. Cai, P. Zhang, and S. Li, "Tf-sepnet: An efficient 1d kernel design in cnns for low-complexity acoustic scene classification," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 821–825.

[2] F. Schmid, P. Primus, T. Heittola, A. Mesaros, I. Martín-Morató, K. Koutini, and G. Widmer, "Data-efficient low-complexity acoustic scene classification in the dcase 2024 challenge," *arXiv preprint arXiv:2405.10018*, 2024.

[3] 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," 2025. [Online]. Available: https://arxiv.org/abs/2505.01747

[4] Y. Cai, P. Zhang, and S. Li, "Tf-sepnet: An efficient 1d kernel design in cnns for low-complexity acoustic scene classification," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 821–825.

[5] D. Nadrchal, A. Rostamza, and P. Schilcher, "Data-efficient acoustic scene classification with pre-trained cp-mobile," DCASE2024 Challenge, Tech. Rep, Tech. Rep., 2024.

[6] F. Schmid, S. Masoudian, K. Koutini, and G. Widmer, "Cp-jku submission to dcase22: Distilling knowledge for low-complexity convolutional neural networks from a patchout audio transformer," *Tech. Rep., Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2022.

[7] B. Han, W. Huang, Z. Chen, A. Jiang, X. Chen, P. Fan, C. Lu, Z. Lv, J. Liu, W.-Q. Zhang, *et al.*, "Data-efficient acoustic scene classification via ensemble teachers distillation and pruning," 2024.

[8] H. Truchan, T. H. Ngo, and Z. Ahmadi, "Ascdomain: Domain invariant device-adversarial isotropic knowledge distillation convolutional neural architecture," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.