HYU SUBMISSION FOR DCASE 2025 TASK 1: LOW-COMPLEXITY ACOUSTIC SCENE CLASSIFICATION USING REPARAMETERIZABLE CNN WITH CHANNEL-TIME-FREQUENCY ATTENTION

Technical Report

Seung-Gyu Han[1], *Pil Moo Byun*[2], *Joon-Hyuk Chang*[1,2,*]

[1] Artificial Intelligence Semiconductor Engineering, Hanyang University, Seoul, Republic of Korea
[2] Artificial Intelligence, Hanyang University, Seoul, Republic of Korea,
{sghan99, fordream0309, jchang}@hanyang.ac.kr

ABSTRACT

This paper presents the Hanyang University team's submission for the DCASE 2025 Challenge Task 1: Low-Complexity Acoustic Scene Classification with Device Information. The task focuses on developing compact and efficient models that generalize well across both seen and unseen recording devices, under strict constraints on model size and computational cost. To address these challenges, we propose Rep-CTFA, a lightweight convolutional neural network that integrates two key design elements: (1) reparameterizable convolutional blocks with learnable branch scaling coefficients, and (2) a Channel-Time-Frequency Attention (CTFA) module. In addition, we explore input resolution variation by adjusting the hop length and number of mel bins to control time-frequency granularity. Knowledge distillation from a PaSST-based teacher ensemble is used to guide the training of the student model, improving generalization. Finally, we adopt a device-aware fine-tuning scheme that updates lightweight classification heads per device while keeping the shared backbone intact.

Index Terms— Acoustic scene classification, Low-complexity, Reparameterization, Learnable branch scaling coefficients, CTFA, Knowledge distillation

1. INTRODUCTION

Acoustic Scene Classification (ASC) involves identifying the environment in which an audio recording was captured, such as a street, park, or shopping mall [1, 2]. Recent advances in deep learning have significantly improved the performance of ASC systems [3, 4, 5]. However, deploying such models on resource-constrained edge devices, such as mobile phones and embedded systems, remains a challenging problem. These devices often have limited memory and processing capacity and must operate under strict latency and energy constraints [6]. Furthermore, ASC systems must remain robust to domain shifts caused by differences in device characteristics.

To address these challenges, the DCASE 2025 Task 1 [7, 8] focuses on low-complexity acoustic scene classification with device information. Participants are required to design models that operate under fixed complexity limits—defined in terms of parameter count and multiply-accumulate operations—and generalize well across both seen and unseen devices. Moreover, the task simulates real-time inference inference conditions by limiting the input to a short audio segment, thereby increasing the difficulty of the problem.

In response to these requirements, we propose a compact and efficient ASC model named Rep-CTFA. The model integrates two key architectural innovations: (1) reparameterizable convolutional blocks that decouple the training and inference architectures to allow efficient deployment [9], and (2) a Channel-Time-Frequency Attention (CTFA) mechanism, inspired by multi-channel speech enhancement [10], designed to improve spectro-temporal feature extraction. These components work in synergy to enhance model capacity while maintaining low inference complexity.

We also examine the impact of input resolution on model performance. By varying the hop length and number of mel bins in the feature extraction process, we obtain multiple configurations with different time-frequency resolutions [4, 5]. Each configuration offers complementary strengths in capturing short-duration acoustic cues.

To improve generalization to diverse devices, we apply knowledge distillation [11, 12] using an ensemble of large teacher models based on the PaSST architecture [13]. In addition, we adopt a device-specific fine-tuning strategy, inspired by the DCASE baseline [7], which uses device labels to adapt lightweight heads while keeping a shared backbone fixed. This enables specialization for seen devices without compromising model compactness.

These techniques enable our model to perform effectively under both complexity constraints and device variability.

2. METHOD

2.1. Model Architecture

Our proposed architecture, Rep-CTFA, is designed to maximize representational capacity during training while maintaining lightweight and efficient performance during inference. It integrates two core components: reparameterizable convolutional blocks with branch scaling and an attention mechanism tailored to spectro-temporal dynamics.

2.1.1. Reparameterizable Convolutional Backbone

Our model builds upon the Rep-Mobile architecture [9, 14], which extends the idea of reparameterizable convolution to mobile-efficient designs. During training, each convolutional block con-

*Corresponding author

Table 1: Comparison of different systems

System	Hyper-parameter			Complexity		Accuracy (%)	
	Hop	Mel	Freq-mask.	MACs	Parameters	General	Device-specific
S1	500	256	48	18,758,084	62,578	57.41	58.85
S2	500	400	75	29,302,844	62,578	58.17	59.55
S3	300	256	48	29,512,940	62,578	58.29	59.81
S4	368	320	60	29,578,644	62,578	57.62	59.38

sists of multiple branches enabling richer feature extraction and improved optimization. At inference time, these branches are fused into a single equivalent 3x3 convolution, resulting in a simplified and efficient structure well-suited for deployment on resource-constrained devices.

This reparameterization strategy enables the models to leverage the high representational capacity of a multi-branch design during training, while ensuring compliance with the strict complexity constraints—both in parameter count and multiply-accumulate operation (MACs).

2.1.2. Branch Scaling Coefficients

To further enhance flexibility, we introduce learnable branch scaling coefficients. Each convolutional path within a block is associated with a trainable scalar that determines its relative importance. During training, the model learns to emphasize or suppress specific branches, enabling dynamic path weighting. This mechanism improves both optimization stability and generalization, particularly when the number of active parameters is strictly constrained.

2.1.3. CTFA

We integrate a CTFA module into the backbone, adapted from recent advances in multi-channel speech enhancement [10]. This module applies three types of attention sequentially: (1) channel attention to model inter-channel dependencies, (2) temporal attention to capture temporal structures, and (3) frequency attention to emphasize relevant spectral patterns. Combined, these mechanisms enable the model to focus on informative regions of the input while remaining robust to background noise and device variability.

2.2. Data processing

All experiments were conducted on the official 25% subset of the TAU Urban Acoustic Scenes 2022 Mobile development dataset [15].

All audio recordings are resampled to 32kHz and transformed into log-mel spectrograms using a Short-Time Fourier Transform. The baseline preprocessing configuration, adopted from the official DCASE 2025 baseline system [7], uses a hop length of 500 samples and 256 mel-frequency bins. This setup balances time and frequency resolution for general-purpose acoustic scene analysis.

To explore the effect of input resolution under fixed model constraints, we additionally trained three variants by adjusting the hop length and the number of mel bins. One configuration increased the spectral resolution by using 400 mel bins while maintaining the default temporal resolution. Another configuration increased temporal resolution by reducing the hop length to 300 samples. A third variant sought a balanced trade-off using a hop length of 368 and 320 mel bins.

These configurations were intentionally selected to shift the spectro-temporal granularity of the input features while remaining within the allowed MACs budget. Each model variant was trained and submitted independently, in accordance with the competition rule that permits multiple final model submissions.

2.3. Data Augmentation

To improve generalization under limited training data and diverse device domains, we apply a combination of spectrogram-level and waveform-level data augmentation techniques. Each augmentation strategy is designed to improve robustness domain variability.

2.3.1. Frequency MixStyle

We adopt a frequency-wise variant of MixStyle, a domain generalization technique that mixes instance-level feature statistics along the frequency axis [16]. This method implicitly simulates domain shifts across recording conditions by interpolating the mean and variance of spectrogram features from different samples. A mixing probability of 0.7 and an alpha parameters of 0.3 are used.

2.3.2. SpecAugment

SpecAugment [17] is applied in the frequency domain to enhance robustness against frequency-localized distortions and improve generalization. To ensure consistency across different input resolutions, the masking width is adapted to the number of mel bins. Specifically, a maximum frequency mask width of 48 is used for 256-bin inputs, 60 for 320-bin inputs, and 75 for 400-bin inputs. Time masking is disabled to preserve temporal alignment crucial for short-duration audio clips.

2.3.3. Time Rolling

We apply circular shifts in the time domain to the waveform with a maximum offset of 0.125 seconds. This simple augmentation introduces temporal variation without altering the underlying acoustic scene structure, encouraging the model to focus on global scene characteristics rather than local temporal cues.

2.3.4. Device Impulse Response Simulation

To simulate cross-device variability and improve generalization to unseen devices, we augment recordings from the majority device (Device A) by convolving them with impulse responses [18] collected from a variety of consumer devices. With a probability of 0.6, a randomly selected device response is applied to each sample.

Table 2: Class-wise Device-specific accuracy (%) of each system

System	Airport	Bus	Metro	Metro Station	Park	Public Square	Shopping Mall	Street Pedestrian	Street Traffic	Tram	Accuracy
S1	47.47	75.66	57.74	53.40	74.34	45.05	59.80	37.31	76.03	61.69	58.85
S2	48.34	76.40	57.98	53.43	76.23	47.47	60.57	36.73	77.98	60.34	59.55
S3	45.10	74.38	61.58	58.89	75.76	45.12	59.16	40.20	76.94	60.95	59.81
S4	46.72	77.31	61.18	51.55	77.85	46.40	55.99	42.32	77.04	57.36	59.38

Table 3: Device-wise Device-specific accuracy (%) of each system

System	A	B	C	S1	S2	S3	S4	S5	S6	Accuracy
S1	68.39	61.06	63.34	57.03	55.12	61.03	56.79	57.00	49.91	58.85
S2	67.73	60.30	62.46	58.09	58.03	59.82	57.85	57.27	52.57	59.55
S3	68.48	61.03	63.56	58.33	57.70	59.67	58.70	58.70	52.15	59.81
S4	69.21	61.97	62.73	57.27	56.67	60.12	57.58	56.97	50.79	59.38

The impulse responses are sourced from the MicIRP [1], which provides a diverse set of real-world device-specific acoustic signatures.

2.4. Knowledge Distillation

To improve generalization in the low-resource setting, we employ knowledge distillation with an ensemble of teacher models. Each teacher is based on a PaSST[13] architecture pre-trained on the large-scale AudioSet dataset [19]. These models are further fine-tuned on the official 25% subset of the TAU Urban Acoustic Scenes 2022 Mobile development dataset [15] to adapt to the task-specific acoustic domains.

To promote diversity among the teachers, we adopt three distinct augmentation strategies during fine-tuning: frequency MixStyle only, device impulse response only, and a combination of both. Their outputs are averaged to form the final soft targets used during distillation. The student model (Rep-CTFA) is optimized using a hybrid loss function that combines cross-entropy with the ground-truth labels and Kullback-Leibler divergence [20] with the ensemble soft targets. A temperature of 0.1 is applied to soften the teacher predictions, and the distillation loss weights set to 0.5 to balance hard and soft supervision.

2.5. Device-specific Fine-Tuning

Since acoustic characteristics can vary significantly across recording devices, device-aware adaptation is critical to achieving robust generalization [21]. Following the official baseline strategy introduced in the [7], we adopt a device-specific fine-tuning approach to further improve performance on seen devices.

After training a unified backbone model using data from all available devices, we fine-tune the model separately for each of the six seen devices (A, B, C, S1, S2, S3). In this phase, a shared feature extractor is frozen, and lightweight device-specific classification heads are trained independently. This modular strategy allows the model to retain generalized acoustic knowledge while specializing in the acoustic nuances of each device. This targeted adaptation not only enhances device-level accuracy but also preserves the overall model compactness, which is essential under the strict model size and MACs constraints defined by the challenge.

[1] https://micirp.blogspot.com/

3. RESULTS

We evaluated four systems with varying input resolutions while keeping the model architecture and parameter count fixed. The primary differences among these systems lie in the hop length, the number of mel-frequency bins, and the frequency masking width used during training. These configurations were carefully selected to explore the impact of time-frequency resolution trade-offs under the MACs constraint of DCASE 2025 Task 1.

Table 1 summarizes each system's configuration and computational complexity, along with their performance in both general and device-specific training modes. Notably, while System S1 (baseline resolution) is the most lightweight, higher-resolution variants (S2–S4) consistently achieve better accuracy.

A breakdown of class-wise, device-specific accuracy is presented in Table 2, illustrating that certain configurations enhance recognition performance for specific scene categories. Table 3 reports device-wise accuracy, highlighting the variation in performance across both seen and unseen devices.

4. 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] Stowell, Dan *et al.*, "Detection and classification of acoustic scenes and events," *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[3] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *IEEE Int. workshop machine learn. signal process.*, 2015, pp. 1–6.

[4] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and D. M. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2880–2894, 2020.

[5] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," in *Proc. Interspeech*, 2021, pp. 571–575.

[6] J.-H. Lee, J.-H. Choi, P. M. Byun, and J.-H. Chang, "Hyu submission for the dcase 2022: Fine-tuning method using

device-aware data-random-drop for device-imbalanced acoustic scene classification," *Detection Classif. Acoust. Scenes Events Challenge Tech. Rep.*, 2022.

[7] Schmid, Florian *et al.*, "Low-complexity acoustic scene classification with device information in the dcase 2025 challenge," *arXiv:2505.01747*, 2025.

[8] "https://dcase.community/challenge2025/."

[9] Han, Bing *et al.*, "Data-efficient low-complexity acoustic scene classification via distilling and progressive pruning," in *Proc. IEEE Int Conf. Acoust., Speech Signal Process.*, 2025.

[10] X. Zeng and M. Wang, "Channel-time-frequency attention module for improved multi-channel speech enhancement," *IEEE Access*, 2025.

[11] K. Koutini, J. Schlüter, H. Eghbal-Zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," *arXiv:2110.05069*, 2021.

[12] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. Journal Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, 2021.

[13] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," in *Proc. Interspeech*, 2022.

[14] F. Schmid, T. Morocutti, S. Masoudian, K. Koutini, and G. Widmer, "Cp-jku submission to dcase23: Efficient acoustic scene classification with cp-mobile," *Detection Classif. Acoust. Scenes Events Challenge Tech. Rep.*, 2023.

[15] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions," in *Proc. Workshop Detection Classif. Acoust. Scenes Events*, 2020.

[16] Kim, Byeonggeun *et al.*, "Domain generalization with relaxed instance frequency-wise normalization for multi-device acoustic scene classification," in *Proc. Interspeech*, 2022.

[17] Park, Daniel S *et al.*, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019.

[18] T. Morocutti, F. Schmid, K. Koutini, and G. Widmer, "Device-robust acoustic scene classification via impulse response augmentation," in *European Signal Process. Conf.*, 2023.

[19] Gemmeke, Jort F *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int Conf. Acoust., Speech Signal Process.*, 2017.

[20] J. M. Joyce, "Kullback-leibler divergence," in *Int. encyclopedia statistical science*, 2011, pp. 720–722.

[21] J.-H. Lee, J.-H. Choi, P. M. Byun, and J.-H. Chang, "Multi-scale architecture and device-aware data-random-drop based fine-tuning method for acoustic scene classification." in *Proc. Workshop Detection Classif. Acoust. Scenes Events*, 2022.