DYNACP: DYNAMIC PARALLEL SELECTIVE CONVOLUTION IN CP-MOBILE UNDER MULTI-TEACHER DISTILLATION FOR ACOUSTIC SCENE CLASSIFICATION

Technical Report

Yuandong Luo[1]*, Hongqing Liu*[1]*, Liming Shi*[1]*, Lu Gan*[2]

[1]School of Communications and Information Engineering
Chongqing University of Posts and Telecommunications, Chongqing, China
hongqingliu@cqupt.edu.cn
[2]College of Engineering, Design and Physical Science, Brunel University, London, U.K.

ABSTRACT

This report introduces the acoustic scene classification (ASC) architecture submitted by the Chongqing University of Posts and Telecommunications – Audio Lab (CQUPT-AUL) for DCASE 2025 Task 1. The architecture is a lightweight and efficient network structure, termed as DynaCP. Built upon CP-Mobile, DynaCP dynamically selects between dilated convolutions with pooling or depthwise convolutions with pooling at different network layers, thereby enhancing multi-scale feature representation with minimal computational overhead, while also alleviating the issue of information sparsity caused by dilated convolutions. To improve classification accuracy, a multi-teacher knowledge distillation approach is employed using pre-trained models of DYMN and MN. Experimental results demonstrate that DynaCP achieves competitive performance while maintaining low computational complexity.

***Index Terms*—** DynaCP, Knowledge distillation, Lightweight, Acoustic scene classification, DCASE 2025

1. INTRODUCTION

The acoustic scene classification (ASC) task of DCASE 2025 Task 1 aims to classify 1-second audio clips into one of ten predefined acoustic scene categories [1]. As in previous editions, the challenge emphasizes model design under conditions of limited labeled data and low computational complexity, requiring participants to develop compact models with reduced parameter count and inference cost.

This year's challenge introduces several key changes that further increase the difficulty while also opening up new opportunities for performance improvement. Participants are restricted to using only 25% of the training dataset, encouraging the adoption of data-efficient strategies such as pre-training or transfer learning. Additionally, recording device information is now provided for both the development and evaluation sets, enabling device-specific fine-tuning and improving system performance in realistic deployment scenarios, where the recording device is typically known. Participants are allowed to train separate models for different devices to fully exploit this information.

To ensure deployability on resource-constrained platforms, strict constraints are imposed on model size and computational complexity. Specifically, the total number of model parameters must not exceed 128 kB (including all values, even zeros), and the maximum inference complexity is capped at 30 million multiply-accumulate operations (MMACs), reflecting the computational requirements of typical edge devices.

Under these constraints, we propose a compact, efficient, and deployable ASC system based on a modified version of CP-Mobile [2]. In details, to enhance multi-scale feature extraction while maintaining low computational cost, we integrate either dilated convolutions with pooling or depthwise convolutions with pooling at different network layers. The resulting model is named DynaCP, where "Dyna" stands for dynamic selection, indicating our configuration of the most suitable convolution-pooling combinations at each layer.

To further improve the accuracy of the classification, we employ multi-teacher knowledge distillation [3] using two pre-trained models, DyMN and MN as teachers [4, 5, 6]. These models are variants of MobileNet, respectively, both of which are CNN-based and have shown excellent performance in audio classification tasks, achieving notable results in past DCASE challenges [7, 8]. The architectural consistency between the student and teacher models facilitates effective knowledge transfer. Moreover, we introduce a two-stage distillation strategy to better preserve and transfer knowledge from the teacher models. Experimental results demonstrate that the proposed model achieves competitive performance while satisfying all hardware constraints, highlighting its potential for real-world deployment.

2. DATA PREPROCESSING AND AUGMENTATION

2.1. Data Preprocessing

For the student model DynaCP, audio clips were sampled at 44.1 kHz and each 1-second segment was converted into a log-Mel spectrogram with 256 frequency bins, using a Hann window of length 3072 and hop size 500. The STFT was computed with a window length of 4096. Notably, the log-Mel spectrogram was based on an effective sampling rate of 32,000 Hz, simulating lower bandwidth while retaining the original audio quality. DynaCP contains only 61.6K parameters and requires 28,938,900 MACs for inference on 1-second audio, meeting the competition requirements.

The teacher models also adopted the same preprocessing pipeline to maximize the effectiveness of knowledge distillation. Both DyMN and MN were pre-trained on the AudioSet dataset and subsequently fine-tuned on our dataset. Notably, due to the inclusion of device information in this challenge, MN and DyMN

were first trained on data without device labels, producing four fine-tuned versions each, for a total of eight models. The fine-tuning logic involved controlling whether time-frequency masking and roll augmentation strategies were enabled. Then, they were further trained on data that included device information, resulting in $(4 + 4) \times 6 = 48$ device-specific models. Accordingly, we designed the corresponding distillation strategy: the student model was first distilled using data without device information, followed by six additional distillations using data from each of the six devices separately, aiming to enhance adaptation to different recording conditions.

2.2. Data Augmentation

To enhance the generalization and robustness of the model, we employed various data augmentation techniques during training. These include SpecAugment strategies [9], where the frequency masking width is set to 48 and time masking is not used (maximum masking width set to 0), simulating the effects of missing frequency bands or device variations. Given a spectrogram $S \in R^{T \times F}$, a segment of frequency bins is masked out with a maximum width of F_{mask}. The operation is

$$S'(t,f) = \begin{cases} 0 & \text{if } v \le f < v + F_{\text{mask}}, \\ S(t,f) & \text{otherwise}, \end{cases} \tag{1}$$

where v is a randomly chosen starting frequency index.

Additionally, we introduced the MixStyle augmentation strategy [10], which mixes the statistical distributions of feature maps to generate new distributions. This method activates with a probability of $p = 0.8$, and the mixing weights follow a Beta distribution with $\alpha = 0.4$, thereby improving the model's adaptability to different recording conditions.

Given two training samples (x_i, y_i) and (x_j, y_j), the Mixup operation is formulated by

$$\begin{aligned} &\lambda \sim \text{Beta}(\alpha, \alpha), \\ &\hat{x} = \lambda x_i + (1 - \lambda) x_j, \end{aligned} \tag{2}$$

where $\hat{x}$ denotes the mixed output.

Meanwhile, we applied a pre-emphasis filter to enhance high-frequency components and introduced spectral diversity by randomly perturbing the cutoff frequencies f_{min} and f_{max} of the Mel filter banks.

3. MODEL STRUCTURE AND TRAINING METHODOLOGY

3.1. Overall Design Pipeline

The overall design pipeline is illustrated in Figure 1 and Figure 2. It covers the knowledge distillation process as well as the interactions among DynaCP, DyMN, and MN models, with a detailed depiction of the two-stage distillation procedure.

In Figure 1, eight pre-trained teacher models are fine-tuned using the training set, which consists of 25% of the total dataset without device information. The logits output by these teacher models are summed to generate soft labels, which are then used to perform distillation on an untrained DynaCP student model. The purpose of using an ensemble of teachers is to reduce the output fluctuation of individual teacher models and thereby improve the efficiency and stability of the distillation process. Through this stage of training, we obtain an initial student model, denoted as *base*.

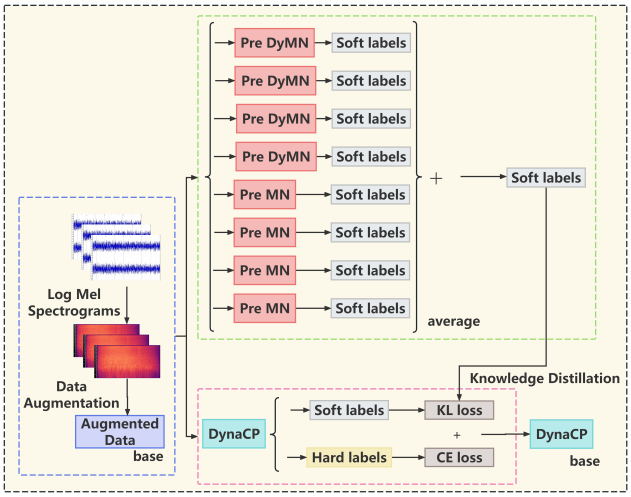

Figure 1: Distillation architecture without device information.

In Figure 2, data corresponding to six different devices (a, b, c, s1, s2, s3) are extracted from the *base* training set to form six separate sub-training sets. Each sub-training set is named according to its associated device. For each sub-training set, we fine-tune four MN and DyMN models. Notably, these models have already been pre-finetuned using the base training set prior to this stage. Subsequently, each sub-training set is used to further distill the base student model. This process is repeated six times, once for each sub-training set. Finally, seven fine-tuned models are obtained, labeled as DynaCP_a, DynaCP_b, DynaCP_c, DynaCP_s1, DynaCP_s2, DynaCP_s3, and DynaCP_base, respectively.

In the following sections, we will provide detailed descriptions of the student model and teacher model.

3.2. Student Model: DynaCP

DynaCP is a lightweight dynamic architecture model based on the CP-Mobile baseline. The core innovation lies in modifying the inverted residual block of CP-Mobile, referred to as CPBlock, by replacing the standard depthwise convolution with a dynamic combination of dilated convolution and pooling or depthwise convolution and pooling. We name this improved module DPCPBlock (Dynamic Parallel Convolution and Pooling Block). In DPCPBlock, "D" represents either a standard depthwise convolution (depthwise convolution) or a dilated convolution, while "P" denotes the pooling operation. To distinguish between different types of convolutions, when using a standard depthwise convolution, the module is denoted by DPCPBlock(1); if using a dilated convolution with a dilation factor r, it is denoted by DPCPBlock(r), where r is the dilation rate. This design aims to enhance the feature extraction capabilities of depthwise convolution while mitigating the information sparsity caused by dilated convolution through parallel pooling branches.

Furthermore, DPCPBlock supports three different residual structures depending on the stride and channel dimensions. Therefore, combining the two dynamic structures with three residual forms, DPCPBlock supports a total of $2 \times 3 = 6$ variant structures, depcited in Figure 3. For example, DPCPBlock(1)_1 indi-

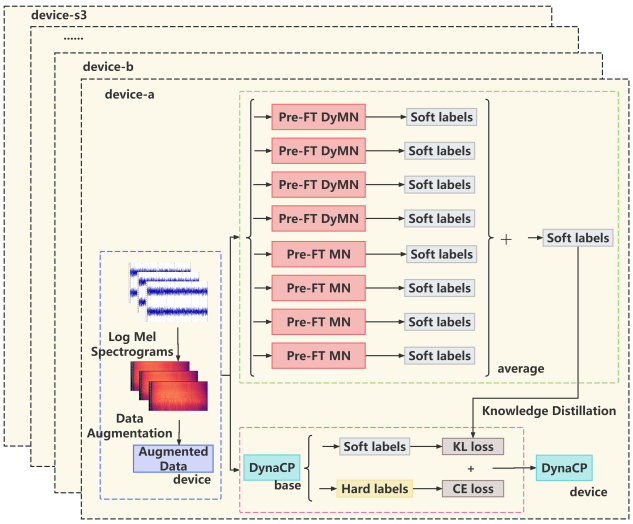

Figure 2: Multi-device information distillation architecture.

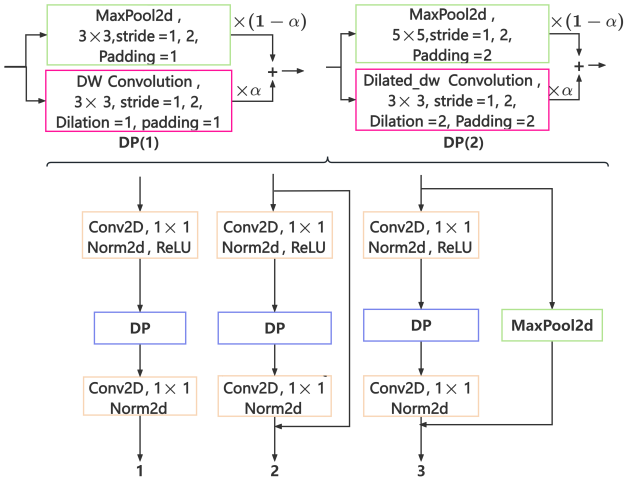

Figure 3: The various dynamic structures of the DPCPBlock, with attention to the fact that the pooling in Residual 3 maintains the same structure as the pooling in DP.

cates the use of a standard depthwise convolution with the first type of residual connection, while DPCPBlock(2)_2 corresponds to a dilated convolution combined with the second type of residual connection. These modules are flexibly configured based on input conditions, thereby enhancing the overall adaptability and flexibility of the model.

The overall architecture of DynaCP is summarized in Table 1.

Table 1: DynaCP Architecture

Input	Operator	Stride	Dilation
256 x 89 x 1	Conv2D@3x3, BN, ReLU6	2 x 2	-
128 x 45 x 8	Conv2D@3x3, BN, ReLU6	2 x 2	-
64 x 23 x 32	DPCPBlock(2)_2	1 x 1	2
64 x 23 x 32	DPCPBlock(2)_3	2 x 1	1
64 x 23 x 32	DPCPBlock(1)_3	1 x 2	1
64 x 12 x 32	DPCPBlock(2)_1	1 x 1	2
32 x 12 x 72	DPCPBlock(1)_3	2 x 1	1
32 x 12 x 72	DPCPBlock(2)_1	1 x 1	2
32 x 12 x 168	Conv2D@1x1, BN	1 x 1	-
32 x 12 x 10	Avg. Pool	-	-

Input format: Frequency Bands x Time Frames x Channels.

3.3. Teacher Model: Efficient-AT

The teacher models are based on MobileNetV3, a lightweight convolutional neural network that employs efficient inverted residual blocks and linear bottleneck structures, reducing computational complexity and the number of parameters through depthwise separable convolutions. The model includes two variants: MN and DyMN. Among them, DyMN introduces a dynamic mechanism to better adapt to diverse input features.

In both stages of knowledge distillation, we use both MN and DyMN models. The reason for selecting four teacher models lies in the fact that the combination of using or not using time-frequency masking and time roll augmentation results in four different configurations. This strategy leverages the structural similarity between MN and DynaCP to achieve stable general feature transfer, while the dynamic design of DyMN enhances adaptation to specific device conditions. By combining the advantages of these two architectures, we effectively improve the robustness and generalization performance of the student model across various recording environments.

4. SUBMISSIONS AND RESULTS

In the test phase, the proposed design was comprehensively compared with the baseline system, demonstrating competitive performance on nearly all metrics, both on the base dataset and the device dataset, as shown in Table 3.

Table 2: Overall Performance Metrics.

Base				
Model	Real	Seen	Unseen	Accuracy
DyMN	64.57	60.37	59.56	61.50
MN	67.06	59.46	57.33	61.28
DyMN + MN	67.84	62.25	60.53	63.54
General	63.65	56.96	56.39	59.00
Device				
Model	Real	Seen	Unseen	Accuracy
DyMN	67.69	61.62	59.58	62.96
MN	69.11	59.06	57.33	61.83
DyMN + MN	70.09	62.13	60.52	64.24
Device-specific	67.50	57.24	56.39	60.37

Table 2 presents the performance of the model and teacher mod-

Table 3: Class-wise and Device-wise Accuracies for Different Models.

Model	Airport	Bus	Metro	Metro Station	Park	Public Square	Shopping Mall	Street Pedestrian	Street Traffic	Tram
General	46.76	74.14	57.04	53.60	76.23	43.74	65.19	38.45	78.32	56.49
Device-specific	52.67	75.39	57.00	53.43	79.33	46.20	66.30	37.34	78.48	57.57

Model	A	B	C	S1	S2	S3	S4	S5	S6
General	69.00	59.12	62.83	56.30	55.21	59.36	57.12	57.09	54.97
Device-specific	71.52	63.74	67.23	56.48	55.82	59.42	57.12	57.09	54.97

els on additional overall metrics, where it also achieves excellent results.

5. REFERENCES

[1] 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

[2] F. Schmid, T. Morocutti, S. Masoudian, K. Koutini, and G. Widmer, "Distilling the knowledge of transformers and CNNs with CP-mobile," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023)*, 2023, pp. 161–165.

[3] H. Fei, X. Li, and J. Jia, "Acoustic scene classification based on multi-teacher knowledge distillation and serfr-CNN," DCASE2023 Challenge, Tech. Rep., May 2023.

[4] F. Schmid, K. Koutini, and G. Widmer, "Dynamic convolutional neural networks as efficient pre-trained audio models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2227–2241, 2024.

[5] ——, "Efficient large-scale audio tagging via transformer-to-cnn knowledge distillation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[6] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, 2020, pp. 56–60. [Online]. Available: https://arxiv.org/abs/2005.14623

[7] 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.

[8] M. Surkov, "Efficient acoustic scene classification using mean-teacher and knowledge distillation," DCASE2024 Challenge, Tech. Rep., May 2024.

[9] J. Han, M. Matuszewski, O. Sikorski, H. Sung, and H. Cho, "Randmasking augment: A simple and randomized data augmentation for acoustic scene classification," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[10] T. Morocutti, F. Schmid, K. Koutini, and G. Widmer, "Device-robust acoustic scene classification via impulse response augmentation," in *2023 31st European Signal Processing Conference (EUSIPCO)*, 2023, pp. 176–180.