DATA-EFFICIENT ACOUSTIC SCENE CLASSIFICATION WITH PRE-TRAINED CP-MOBILE

Technical Report

David Nadrchal, Aida Rostamza, Patrick Schilcher

Students at Johannes Kepler Universität Linz, Linz, Austria
{k12213656, k12237081, k12222369}@students.jku.at

ABSTRACT

This report presents our submission for Task 1: Data-Efficient Low-Complexity Acoustic Scene Classification in the DCASE2024 challenge. Drawing inspiration from the top-ranked system in the 2023 edition, our approach is based on a Knowledge Distillation training routine: we employ ensembles of fine-tuned CP-ResNet and PaSST as teachers for each subset, with a modified version of the CP-Mobile baseline model serving as the student. A key improvement in our methodology is pre-training the student on both AudioSet and the corresponding training subset before knowledge distillation, which significantly enhances its performance. To improve device generalization, we use various data augmentation techniques, including Freq-MixStyle, Device impulse response augmentation, FilterAugment, frequency masking, and time rolling. Our results demonstrate substantial improvements in test accuracy compared to the baseline system, validating the effectiveness of our approach for each subset.

Index Terms— CP-Mobile, Knowledge distillation, CP-ResNet, Device Impulse Response augmentation, Freq-MixStyle, AudioSet pre-training, Acoustic scene classification, Frequency masking, Time rolling

1. INTRODUCTION

Acoustic Scene Classification (ASC) systems aim to categorize audio recordings into predefined scene classes. This field has gained prominence through the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge which takes place annually since 2016 [1]. In this report, we describe our submission for Task 1 of DCASE 2024 edition [2] which concerns urban acoustic scene classification utilizing the TAU Urban Acoustic Scenes dataset [3]. The task has gradually evolved over the years [4, 5] to address real-world problems such as device mismatch or low-complexity constraints, emphasizing model applicability to portable devices.

This year's challenge introduces a new scenario with limited labeled data availability. Participants must design systems that maintain high prediction accuracy with limited training data across five scenarios: 5%, 10%, 25%, 50%, and 100% of the full training set. Systems must be trained solely on the specified subset and allowed external resources [6]. Furthermore, the rules specify a maximum of 128 kB of parameters and a ceiling of 30 million multiply-accumulate operations (MMACs) per inference of a one-second audio clip.

Our approach builds on insights from the previous year's top-ranked system [7]. It involves training a CP-Mobile [7] model by distilling the knowledge [8] from a CP-ResNet [9] and PaSST [10]. We utilize Device Impulse Response augmentation [11], FilterAugment [12] and Freq-MixStyle [13] to handle generalization to unseen devices. Additionally, we pre-trained the student on AudioSet, leading to a substantial performance improvement.

2. FEATURE EXTRACTION & DATA AUGMENTATION

2.1. Dataset

The development dataset employed for this challenge is the TAU Urban Acoustic Scenes 2022 Mobile development dataset (TAU22) [3]. This dataset encompasses recordings from 12 European cities, capturing 10 distinct acoustic scenes using 4 real devices. Additionally, synthetic data for 11 mobile devices was generated based on the original recordings. The development set is restricted to audio recorded by three real devices (A, B, and C) and six simulated devices (S1-S6).

TAU22 retains the same content as the TAU Urban Acoustic Scenes 2020 Mobile development dataset (TAU20) [14], but the 10-second audio clips from TAU20 have been segmented into 1-second fragments, resulting in ten times more files. This segmentation significantly increases the difficulty of the prediction task. The dataset comprises 230,350 audio clips, each with a duration of 1 second and a label indicating the acoustic scene. All audio files are in a single-channel, 44.1 kHz, 24-bit format.

2.2. Preprocessing

Audio for the CP-Mobile model is resampled to 32 kHz and processed to Mel spectrograms with 256 frequency bins. The Short Time Fourier Transformation (STFT) uses a window size of 96 ms and a hop size of 16 ms. As observed by the authors of the top-ranked system of 2023 [7], increasing the window size from 64 ms to 96 ms and employing a 4096-point Fast Fourier Transform (FFT) results in a marginal performance improvement compared to the configuration detailed in [4].

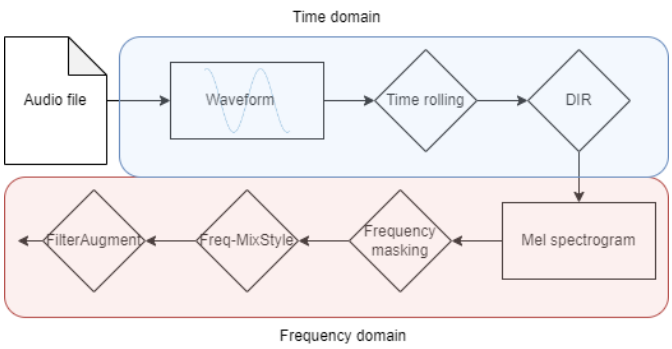
For the teacher models, we adhered to the AudioSet [15] pre-training configuration of PaSST [10]. This involved utilizing a window size of 25 ms and a hop size of 10 ms to generate Mel spectrograms with 128 frequency bins. In the case of CP-ResNet [9], the audio was downsampled to 22.05 kHz. Mel spectrograms are computed using a hop size of approximately 9 ms, a window size of 23 ms, and 256 Mel bins.

2.3. Data augmentations

In order to mitigate overfitting, especially with a relatively small dataset of 5% split, we implemented a diverse range of data augmentation techniques. These augmentations played a crucial role in enhancing the generalization capability of our models.

- SpecAugment [16] is a widely adopted technique for audio data augmentation that applies time and frequency masking to input spectrograms. In our experiments, frequency masking was found to be particularly beneficial. We applied masking up to 48 frequency bins, which significantly improved model robustness.
- FilterAugment [12] is a more sophisticated variant of SpecAugment. Unlike the traditional masking approach, FilterAugment applies frequency-specific weighting to simulate the effects of various impulse responses encountered in different environments.
- Freq-MixStyle [13] is an adaptation of the MixStyle augmentation [17] adjusted for audio data. MixStyle enhances model robustness to domain shifts by normalizing input features using the mean and standard deviation of other samples within the same batch, leveraging the observation that instance-wise statistical moments encapsulate style information [18]. Freq-MixStyle focuses on the frequency dimension, which is crucial for audio data, and we apply it to a batch with a probability of 70%. Mixing coefficients are drawn from a Beta distribution with $\alpha = 0.6$.
- Device Impulse Response (DIR) augmentation [11] involves convolving the input recordings with impulse responses from 66 different vintage microphones. This technique is designed to enhance the model's ability to generalize across recordings from various devices. We apply DIR augmentation to a sample with a probability of 70%.
- Time-rolling involves shifting a prefix/suffix of a randomly sampled length (up to 0.1 seconds) to the other end of the input signal. This augmentation, computed in the time domain, helps to simulate variations in the temporal alignment of the audio data.

Figure 1: Data preprocessing and augmentations



3. ARCHITECTURES

3.1. TEACHER MODELS: PaSST and CP-ResNet

Audio spectrogram transformer models, such as PaSST, excel in capturing the global context of an audio clip due to their purely self-attention-based architecture. Previous studies have demonstrated that PaSST serves as an effective teacher model for low-complexity CNNs [19]. The Patchout faSt Spectrogram Transformer (PaSST) [10] is a complex, self-attention-based model pre-trained on AudioSet and comprising 85 million parameters. This pre-trained model can be fine-tuned to achieve state-of-the-art performance across multiple downstream tasks, including ASC. PaSST models have consistently proven to be excellent teachers for low-complexity CNNs [19].

Similarly, CP-ResNet [9], a receptive-field regularized CNN (RFR-CNN), incrementally builds local features over a spatially restricted area. CP-ResNet has shown significant success in ASC in prior DCASE ASC challenges. By utilizing both PaSST and CP-ResNet as teacher models, we seek to further diversify the predictions within the ensemble, leveraging their unique strengths to improve overall performance [11].

3.2. STUDENT MODEL: CP-Mobile

We use the default CP-Mobile architecture [7] provided as the baseline system. Its architecture is described in Table 1.

Block	Unbatched input shape	Parameters
Input Convolution	[1,256,65]	2,456
CPM-S Block 1	[32, 64, 17]	4,992
CPM-D Block 2	[64, 64, 17]	4,992
CPM-S Block 3	[32, 64, 17]	4,992
CPM-T Block 4	[32, 64, 9]	6,576
CPM-S Block 5	[56, 32, 9]	15,112
CPM-T Block 6	[56, 32, 9]	20,968
Final layer	[104, 32, 9]	1,060

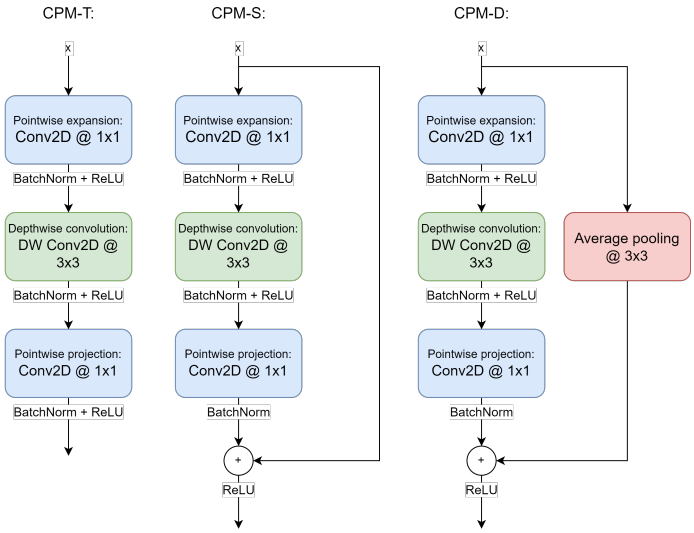
Table 1: Network architecture

The core technique of CP-Mobile is CPM block, a computationally efficient alternative for the classical convolutional layer. Each CPM block consists of three stages: 1) Pointwise expansion 2) Depthwise convolution and 3) Pointwise projection (as depicted in Figure 2). Each of those stages is implemented as a standard covolutional block that includes Batch normalization and ReLU activation.

The CPM blocks are categorized into three types:

1. Transition blocks (T): used to expand the channel dimension. No residual connection is used.
2. Standard blocks (S): number of input and output channels is the same. Residual connection is used.
3. Spatial downsampling blocks (D): number of input and output channels is the same. Residual connection with strided average pooling is used.

Figure 2: Types of CPM blocks



The sequence of CPM blocks is preceded by two classical convolutional blocks (convolution, normalization, activation) and succeeded by a 2D convolution, batch normalization and adaptive average pooling. As the activation function, we use ReLU all across our model.

4. KNOWLEDGE DISTILLATION

Knowledge distillation (KD) [8] is a proven technique for compressing large, complex machine learning models (referred to as teacher models) into smaller, more efficient models (referred to as student models) while maintaining robust performance. The teacher model is a large, high-performing model trained to achieve high accuracy. It generates "soft targets" using a temperature-adjusted softmax function. These soft targets are probability distributions over classes that convey nuanced class similarities beyond traditional hard labels. The temperature parameter (τ) in the softmax function controls the sharpness of these distributions, with higher temperatures producing softer, more informative distributions. The student model is trained using both the soft targets from the teacher model and the standard one-hot encoded labels. This dual training approach enables the student model to capture both explicit label information and the more generalized, nuanced patterns present in the teacher's outputs. The training of the student model employs a combination of two loss (as detailed in Equation 2): the hard label loss (L_l) and the distillation loss (L_{kd}). The hard label loss (L_l) typically uses cross-entropy loss, while the distillation loss (L_{kd}) is computed as the Kullback-Leibler (KL) divergence between the teacher's and student's outputs.

λ is a weight that balances the contributions of the hard label loss and the distillation loss. The distillation loss L_{kd} is defined as:

$$L_{DIST} = D_{KL}(q_{\text{student}} || q_{\text{teacher}}) \quad (1)$$

q_{student} and q_{teacher} are the soft targets of the student and teacher models, respectively. This balance is crucial to ensure that the student model learns both the precise labels and the generalized knowledge from the teacher model.

$$\text{Loss} = \lambda L_l(\delta(z_S), y) + (1-\lambda)\tau^2 L_{kd}(\delta(z_S/\tau), \delta(z_T/\tau)) \quad (2)$$

where z_S and z_T are the logits of the student and teacher models, respectively, and y represents the hard labels. The factor τ^2 ensures that the magnitudes of the gradients produced by the soft targets scale appropriately, maintaining the relative contributions of the hard and soft targets even when the temperature used for distillation is modified.

A common strategy to enhance KD is ensembling the teacher models [11]. An ensemble approach often results in more robust and generalized student models by integrating diverse insights from multiple teachers. The teacher outputs are usually aggregated via averaging. However, in our experiments, the Bayesian Ensemble Averaging (BEA) [20] which replaces averaging with probabilistic sampling, turned out to work slightly better.

5. PRE-TRAINING OF THE STUDENT MODEL

Before passing the student model to knowledge distillation, we subjected it to two pre-training procedures, first on the AudioSet and the second on the respective train split.

AudioSet, comprising over 2 million human-labeled 10-second sound clips, offers a diverse and comprehensive resource for training and evaluating audio recognition models across 527 distinct sound categories [15]. For pre-training the CPM student model on it, we use the training routine described in [21].

Additionally, we train the student model on the corresponding training split and only after that we pass it into KD.

The effectiveness of such pre-training has been demonstrated throughout our experiments. Utilizing a pre-trained student significantly enhances the performance of the student model during the KD phase. This boost can be attributed to the fact that the pre-trained student model already possesses a foundational understanding of the data distribution. When the knowledge from one or more teacher models is distilled into such a student, the inherent data familiarity accelerates the adaptation and integration of new knowledge. This rapid assimilation allows the student to more effectively mimic the teacher models. Moreover, the existing competence of the pre-trained student may enable it to not only replicate but potentially surpass the performance of the teacher models.

6. EXPERIMENTS

6.1. Experimental setup

We trained our models using the provided train and validation splits. We aim to have every improvement compared to the baseline performance supported by a substantial increase in average performance in at least four experiments. Unfortunately, this was not always feasible because of the high computational demands of some of the used techniques (especially knowledge distillation with PaSST) and our limited computing power.

6.2. Knowledge Distillation

For knowledge distillation, we utilized an ensemble of teacher logits, aggregating the logits from four CP-ResNet teachers and in the case of System 2 also a PaSST. The distillation process involved setting the temperature parameter τ to 2 to soften the logits and applying a distillation loss with a weight of 0.02 (see Equation 2).

6.3. Training procedure

To address the challenge of generalization with a limited dataset, we leveraged pretraining on AudioSet for both the teacher and student models.

Each CPM model was initially pretrained on AudioSet and then fine-tuned using our available subset with specific augmentations. The training regime included training the models for 150 epochs with a batch size of 256. We utilize the Adam optimizer [22] and a cosine learning rate scheduler was employed to dynamically adjust the learning rate during training.

We use a weight decay of 0.001, frequency masking of up to 48 frequency points, time rolling of up to 0.1 seconds and linear FilterAugment augmenting from 3 to 6 mel bands in the range of -6 to 6 dB. Some of the hyperparameters were, however, adjusted for different architectures and training setups. Table 2 lists their configuration:

Training	lr	DIR p	FMS p	FMS α
CPM	0.005	0.6	0.6	0.4
CP-ResNet	0.001	0.7	0.7	0.6
PaSST	0.00001	0.6	0.4	0.4

Table 2: Varying hyper-parameters for different training setups (for student model we use the same hyperparameters both for pertaining on TAU22 and for KD)

6.4. Submissions

Based on our results we put together the following three submission systems:

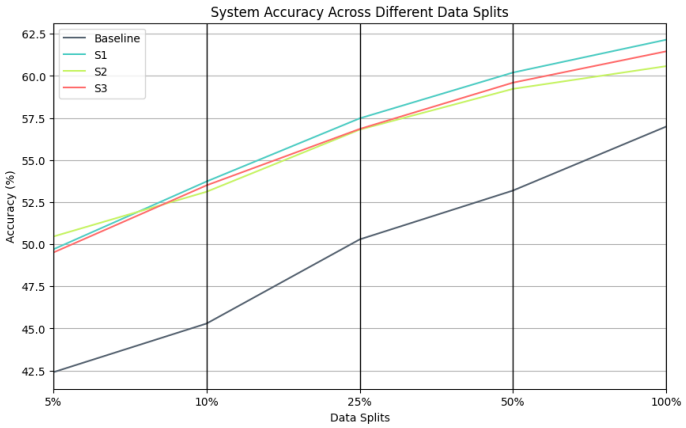
1. **S1**: this system contains only techniques that have been consistently demonstrated to yield good results: DIR, Freq-MixStyle, frequency masking, time rolling, ensemble of three CP-ResNets (aggregated using BEA) as a teacher and pre-training the student model on both AudioSet and on the corresponding train split before passing it to knowledge distillation.
2. **S2**: This system is specialized for the largest splits and adds PaSST (pre-trained on AudioSet and then finetuned on our data) to the teacher ensemble. For CP-ResNets, we used BEA and applied mean aggregation to integrate the results of BEA and the PaSST logits. Other than that, this system is identical with **S1**.
3. **S3**: this system uses classical averaging of teacher logits in KD (i.e. no BAE). Other than that this system is identical with **S1**.

In the Table 3 and in the Figure 3, we show our final validation accuracies of each of the systems for each of the splits:

System	5%	10%	25%	50%	100%
Baseline	42.4	45.29	50.29	53.19	56.99
S1	49.70	53.73	57.48	60.20	62.15
S2	50.46	53.12	56.81	59.23	60.58
S3	49.51	53.50	56.85	59.60	61.46

Table 3: Final results

Figure 3: Validation accuracies for different systems and subset



7. CONCLUSION

In this report, we presented our approach for Task 1: Data-Efficient Low-Complexity Acoustic Scene Classification in the DCASE2024 challenge. Inspired by the DCASE2023 winning model [7], our method leveraged ensembles of fine-tuned CP-ResNet and PaSST as teacher models, with the CP-Mobile model serving as the student. A critical enhancement in our methodology was the pre-training of the baseline model prior to down-stream training on the respective Task 1 data split, which significantly boosted its performance.

To address the challenges of limited availability of training data and generalization across recording devices, we employed a variety of data augmentation techniques. These included Freq-MixStyle, Device Impulse Response (DIR) augmentation, FilterAugment, frequency masking, and time rolling. These augmentations played a vital role in enhancing the robustness and generalization capability of our models.

As shown in Table 3, our experimental results demonstrated substantial improvements in test accuracy compared to the baseline model.

8. ACKNOWLEDGEMENT

This project has been realized as a part of "Machine Learning and Audio: A Challenge" course at JKU Linz. We are grateful to the team of Professor Gerhard Widmer at the Institute of Computational perception for giving us this opportunity to work on a real-world research project. Especially, we would like to thank the course organizers, Florian Schmid and Paul Primus, for their supervision.

9. REFERENCES

[1] DCASE Community, "Dcase community information," https://dcase.community/community_info, 2024, accessed: 2024-06-03.

[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," 2024.

[3] T. Heittola, A. Mesaros, and T. Virtanen, "TAU Urban Acoustic Scenes 2022 Mobile, Development dataset," Mar.

2022. [Online]. Available: https://doi.org/10.5281/zenodo.6337421

[4] DCASE Community, "Dcase 2023 challenge: Task - low-complexity acoustic scene classification," https://dcase.community/challenge2023/task-low-complexity-acoustic-scene-classification, 2023, accessed: 2024-06-03.

[5] ——, "Dcase 2022 challenge: Task - low-complexity acoustic scene classification," https://dcase.community/challenge2022/task-low-complexity-acoustic-scene-classification, 2022, accessed: 2024-06-03.

[6] ——, "Dcase 2024 challenge: Task - data-efficient low-complexity acoustic scene classification," https://dcase.community/challenge2024/task-data-efficient-low-complexity-acoustic-scene-classification, 2024, accessed: 2024-06-03.

[7] F. Schmid, T. Morocutti, S. Masoudian, K. Koutini, and G. Widmer, "Cp-jku submission to dcase23: Efficient acoustic scene classification with cp-mobile," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023)*, 2023, pp. 161–165.

[8] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015.

[9] K. Koutini, H. Eghbal-zadeh, and G. Widmer, "Receptive field regularization techniques for audio classification and tagging with deep convolutional neural networks," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 1987–2000, 2021.

[10] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 2753–2757.

[11] T. Morocutti, F. Schmid, K. Koutini, and G. Widmer, "Device-robust acoustic scene classification via impulse response augmentation," in *31st European Signal Processing Conference, EUSIPCO 2023, Helsinki, Finland, September 4-8, 2023*. IEEE, 2023, pp. 176–180.

[12] H. Nam, S. Kim, and Y. Park, "Filteraugment: An acoustic environmental data augmentation method," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*. IEEE, 2022, pp. 4308–4312.

[13] B. Kim, S. Yang, J. Kim, H. Park, J. Lee, and S. Chang, "Domain generalization with relaxed instance frequency-wise normalization for multi-device acoustic scene classification," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 2393–2397.

[14] T. Heittola, A. Mesaros, and T. Virtanen, "TAU Urban Acoustic Scenes 2020 Mobile, Development dataset," Mar. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3670167

[15] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*. IEEE, 2017, pp. 776–780.

[16] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 2613–2617.

[17] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with mixstyle," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[18] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *CoRR*, vol. abs/1607.08022, 2016.

[19] F. Schmid, S. Masoudian, K. Koutini, and G. Widmer, "Knowledge distillation from transformers for low-complexity acoustic scene classification," in *Proceedings of the 7th Workshop on Detection and Classification of Acoustic Scenes and Events 2022, DCASE 2022, Nancy, France, November 3-4, 2022*, M. Lagrange, A. Mesaros, T. Pellegrini, G. Richard, R. Serizel, and D. Stowell, Eds. Tampere University, 2022.

[20] J. Xu, S. Li, A. Deng, M. Xiong, J. Wu, J. Wu, S. Ding, and B. Hooi, "Probabilistic knowledge distillation of face ensembles," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 2023, pp. 3489–3498.

[21] F. Schmid, K. Koutini, and G. Widmer, "Efficient large-scale audio tagging via transformer-to-cnn knowledge distillation," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*. IEEE, 2023, pp. 1–5.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.