

Data-Efficient Low-Complexity Acoustic Scene Classification via Distilling and Progressive Pruning

Bing Han¹, Wen Huang¹, Zhengyang Chen¹, Anbai Jiang², Pingyi Fan², Cheng Lu³, Zhiqiang Lv⁴,
Jia Liu^{2,4}, Wei-Qiang Zhang², Yanmin Qian^{1,*}

¹ Shanghai Jiao Tong University, Shanghai, China

² Tsinghua University, Beijing, China

³ North China Electric Power University, Beijing, China

⁴ Huakong AI Plus Company Limited, Beijing, China
{hanbing97, yanminqian}@sjtu.edu.cn

Abstract—The goal of the acoustic scene classification (ASC) task is to classify recordings into one of the predefined acoustic scene classes. However, in real-world scenarios, ASC systems often encounter challenges such as recording device mismatch, low-complexity constraints, and the limited availability of labeled data. To alleviate these issues, in this paper, a data-efficient and low-complexity ASC system is built with a new model architecture and better training strategies. Specifically, we firstly design a new low-complexity architecture named Rep-Mobile by integrating multi-convolution branches which can be reparameterized at inference. Compared to other models, it achieves better performance and less computational complexity. Then we apply the knowledge distillation strategy and provide a comparison of the data efficiency of the teacher model with different architectures. Finally, we propose a progressive pruning strategy, which involves pruning the model multiple times in small amounts, resulting in better performance compared to a single step pruning. Experiments are conducted on the TAU dataset. With Rep-Mobile and these training strategies, our proposed ASC system achieves the state-of-the-art (SOTA) results so far, while also winning the first place with a significant advantage over others in the DCASE2024 Challenge.

Index Terms—data-efficient, low-complexity, acoustic scene classification, progressive pruning, knowledge distillation

I. INTRODUCTION

Acoustic scene classification (ASC) [1] is a basic audio processing task that identifies and classifies audio signals into predefined environmental sound scenes such as airports, parks and urban streets. Due to the limitations of application scenarios, ASC systems frequently face problems such as mismatch between recording devices, constraints due to low model complexity, and a scarcity of labeled data, especially when aiming for real-time processing and deployment on resource-constrained devices [2].

In order to build low-complexity ASC systems, many excellent works have emerged. In terms of lightweight model architecture design, researchers predominantly focus on harnessing the power of convolutional neural networks (CNNs). Conventional methods for developing CNN-based ASC systems typically rely on stacking multiple two-dimensional (2D)

kernels [3], [4], whereas more recent studies [5], [6] have investigated the use of one-dimensional (1D) kernels as a promising alternative. In terms of training strategies, some researchers aim to enhance the robustness of models through data augmentation [7], [8], while others directly incorporate adversarial strategies into training [9]. In addition, common model compression methods including pruning [10] and distillation [11] have also been explored in ASC for further reducing computational complexity. Although these works have made good progress in building low-complexity ASC models, there has been no previous exploration of system construction methods in limited training data, and the exploration of distillation and pruning methods is also preliminary.

In this paper, to achieve the highest possible accuracy with limited parameter, computational and training data, we mainly design our system from the aspects of model design, training strategies, model pruning, and distillation on the basis of previous works. The main contributions of this paper can be summarized:

- For building low-complexity ASC model, we propose Rep-Mobile by integrating multi-branches convolution which can be reparameterized at the inference.
- We explore and compare the performance of teacher models with different architecture for knowledge distillation in the absence of data, and find that transformer-based models are more data-efficient.
- We propose progressive pruning strategy, which involves pruning the model multiple times in small amounts. It effectively reduces the parameter gap between the models before and after pruning, while resulting in better performance compared to the single step pruning method.

II. APPROACHES

In this section, we will give a detailed description on our solution for a data-efficient low-complexity ASC system, including the architecture of Rep-Mobile, knowledge distillation with ensemble teachers and progressive pruning strategy.

A. Rep-Mobile Architecture

To improve the performance of the low-complexity model for ASC task, we develop the architecture of Rep-Mobile.

* Yanmin Qian is corresponding author

This work was supported in part by China NSFC projects under Grants 62122050 and 62071288, in part by Shanghai Municipal Science and Technology Commission Project under Grant 2021SHZDZX0102.

Like prior works on structural reparameterization [12]–[14], the train- and inference-time architecture of Rep-Mobile is different. The blocks of Rep-Mobile are based on CP-Mobile [15] which is built by enhancing MobileNet [16] with receptive field regularization techniques [3]. Furthermore, we introduce trivial over-parameterization branches which provide further accuracy gains.

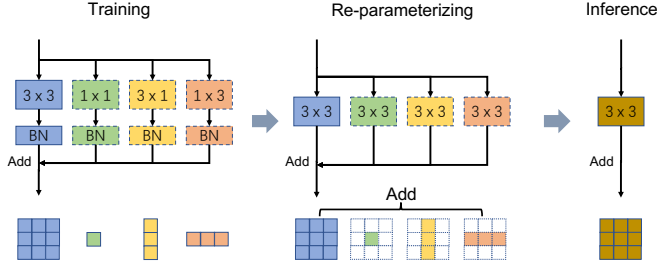


Fig. 1. The reparameterization process of CNN block in Rep-Mobile. During training, multiple branches are used to enhance modeling ability, and multiple branches are merged through reparameterization without increasing computational complexity during the inference. BN denotes BatchNorm layer.

The entire module design is shown in Figure 1. During the training phase, we introduce additional branches of other shapes to the depth-wise convolution modules. The addition of a 1×1 convolution kernel aims to introduce more linear transformations, while the 3×1 and 1×3 convolutions can extract the features separately in the frequency- and time-domains, whose similar designs have been validated as effective in other works [17], [18]. After the training is completed, reparameterization techniques can be used to equivalently transform multiple branches by expanding the kernels with different shape, and then ultimately merge them into the main branch. Specifically, for an additional convolutional layer of kernel size $K_0 \times K_1$, input channel dimension C_{in} and output channel dimension C_{out} , the weight matrix is indicated as $\mathbf{W} \in \mathbb{R}^{C_{out} \times C_{in} \times K_0 \times K_1}$ and bias is indicated as $\mathbf{b} \in \mathbb{R}^D$. The first step of reparameterization is to fill in the same shape using zero values according to the shape of the main branch, denoted as $\mathbf{W}'$ and $\mathbf{b}'$. Then the next step is to perform an equivalent transformation to eliminate the BatchNorm (BN). For a BN layer, it contains accumulated mean μ , standard deviation σ , scale γ and bias β . Since convolution and BN at inference are linear operations, they can be folded by: $\hat{\mathbf{W}} = \mathbf{W}' \times \frac{\gamma}{\sigma}$ and $\hat{\mathbf{b}} = (\mathbf{b}' - \mu) \times \frac{\gamma}{\sigma} + \beta$. Following these steps, Rep-Mobile model does not have any branches which will bring additional calculation cost at inference.

B. Knowledge Distillation with Ensemble Teachers

To enhance the performance of low-complexity models in the ASC task, we employ the widely adopted knowledge distillation framework [19], which emphasizes replicating the final predictions of the teacher model. The knowledge transfer process consists of two key steps, beginning with the pretraining phase. At this stage, we mainly consider two different model architectures, including CNN (i.e., BC-ResNet [6], CP-ResNet and CP-Mobile [3]) and Transformer (i.e., PaSST [20]), and

pretrain or finetune them on the training set. These models will be fused together to form the ensemble teacher model for subsequent knowledge distillation.

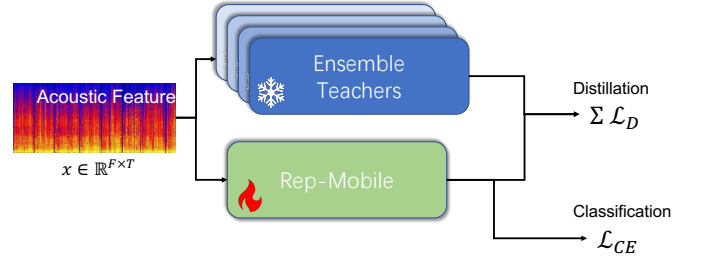


Fig. 2. Illustration of knowledge distillation with ensemble teachers to low-complexity Rep-Mobile. Note that snowflake represents that the parameters of teachers are frozen, while the flame icon represents that the student model are updated with gradient.

The overview of distillation is illustrated in Figure 2. The input acoustic feature is a log-mel spectrogram $x \in \mathbb{R}^{F \times T}$. Then the features x will be fed into pretrained K ensemble teacher models and the corresponding predictions $y_{t,k}$ are obtained, serving as the teacher logits for the knowledge distillation process. For the student model, the acoustic feature x is also used as input and y_s is the corresponding output. Two combined losses $\mathcal{L} = \lambda \mathcal{L}_{CE} + (1 - \lambda) \frac{1}{K} \sum_k \mathcal{L}_{D,k}$ are used for gradient optimization at the same time, where λ is the factor to balance them. Among them, $\mathcal{L}_{CE}$ stands for classification loss, which is calculating the cross entropy (CE) between y_s and the groundtruth $\hat{y}$. $\mathcal{L}_{D,k}$ represents knowledge distillation loss, which is achieved by calculating the KL divergence between output distributions of all teacher models and students:

$$\mathcal{L}_{D,k} = \text{KLD}(\delta(y_s), \delta(y_{t,k})) \quad (1)$$

where $\delta(y) = \text{Softmax}(y/\tau)$ is a sharpening operation with temperature factor $\tau = 0.1$.

In addition, to reduce the training cost, following [21], we efficiently distill the ensemble large teacher models by storing the teacher's logits as well as their respective augmentation method on disk. As a result, during the distillation, the forward process of the teacher models does not need to be recomputed, significantly reducing computational overhead.

C. Progressive Pruning

Currently model pruning is a very commonly used compression technique [22]. It aims to reduce model complexity by removing less important parameters, such as weights or neurons, while preserving the performance. However, dropping too many parameters at once can easily lead to parameter gaps and a significant degradation in performance. To alleviate this issue, we propose a progressive pruning strategy which breaks down the pruning process from a single step to multiple steps. It repeats the “pruning-finetuning” process, and reduces the model size gradually. The goal is to create a more efficient model with fewer parameters and lower computational costs, making it ideal for resource-constrained environments like mobile or edge devices. The basic pruning algorithm we adopt

is the linear pruning whose sparsity ratio increases linearly during each pruning rounds, provided by NNI toolkit [23].

III. EXPERIMENTS SETUP

A. Datasets

The experiments are conducted with the *TAU Urban Acoustic Scene 2022 Mobile* dataset [24], [25] following the setup in [26], which is a widely recognized benchmark for the task of low-complexity ASC. It provides one-second audio snippets with a sampling rate of 44.1 kHz in single-channel from ten distinct acoustic scenes. The recordings in this dataset were captured by various mobile devices across multiple cities worldwide, introducing challenges to the generalization ability of ASC model.

B. Training Configuration

For the acoustic features, all audio segments are down-sampled to 32k Hz. Short-Time Fourier Transform (STFT) is employed to extract time-frequency representations, with a window size of 3072 and a hop size of 512. Following the STFT, a Mel-scaled filter bank with 256 frequency bins and 4096 FFT is applied to transform the spectrograms into a Log-Mel spectrograms. To ensure a fair comparison, all models, including the baseline systems, are trained for 150 epochs using the Adam optimizer with Pytorch Framework. The batch size is fixed at 128. A warmup strategy is employed to facilitate faster convergence and stabilize training. Specifically, the learning rate is linearly increased from 0 to 0.01 in the first 5 epochs and then gradually reduced to 0 over the remaining epochs following a cosine annealing schedule.

C. Data Augmentation

Data augmentation has a crucial impact on model performance, and we have adopted the following data augmentation strategies: (1) Roll Audios: Waveforms are randomly rolled over time with a maximum shift of 125ms for diversity. (2) SpecAug [27]: Random masking in the frequency band will be applied with a maximum size of 48. (3) Freq-MixStyle [28], [29]: Freq-MixStyle (FMS) is a frequency-wise version of the original MixStyle [30] that operates on the channel dimension. α and p of Freq-MixStyle are respectively set to 0.3 and 0.7.

D. Evaluation

To evaluate data-efficiency, five pre-defined subsets [26] will be utilized, each progressively reducing the available training data to 100%, 50%, 25%, 10%, and 5% of the recordings. The distribution of acoustic scenes, cities, and recording devices remains consistent across all subsets. The smaller subsets are fully included in the larger ones, corresponding to the idea of progressively collecting more data.

IV. RESULTS AND ANALYSIS

A. Comparison with other models

To demonstrate the low-complexity of our proposed model Rep-Mobile, we select multiple classic models as baseline for comparison. And results with parameters and MACs are

TABLE I
EVALUATION RESULTS ON THE TEST SET UNDER 100% TRAINING DATASET. ACC. DENOTES THE TOP-1 ACCURACY (%) ON THE TEST SET. MACs (MULTIPLY ACCUMULATE OPERATIONS) INDICATES THE COMPUTATIONAL COSTS PER INFERENCE. PARAM. REPRESENTS THE NUMBER OF PARAMETERS. BASE CHANNEL OF REP-MOBILE IS 32.

Models	# Param.	MACs	Acc.
ResNet34 [31]	7.42M	4901M	58.32
ECAPA-TDNN [32]	6.45M	367M	51.28
MobileNet [33]	3.91M	89M	58.22
BC-ResNet [6]	23K	29M	55.80
CP-ResNet [15]	94K	29M	57.57
CP-Mobile [15]	126K	29M	56.15
TF-SepNet [17]	126K	29M	57.00
Rep-Mobile	126K	29M	58.59

showed in Table I. For a more comprehensive comparison, we firstly provide the comparison results of ResNet [31], [34], ECAPA-TDNN [32], MobileNet [33], [35] which are powerful audio encoders widely used in other tasks. Although these models achieve relatively high accuracies, the accompanying large computational and parameter loads can bring serious real-time and storage issues. As for the recently popular models with low-complexity design, we align the computational complexity by adjusting the width of them, in order to make more consistent comparisons. BC-ResNet [6] uses 1D and 2D CNN features together for better efficiency and won 1st place in DCASE2021 challenge. It has a smaller number of parameters, but the computational complexity is relatively higher and the performance is relatively worse. CP-ResNet and CP-Mobile [15] are modified CNN models by restricting the receptive field [15]. And TF-SepNet [17] separates the feature processing along the time and frequency dimensions to reduce computation cost. Through comparison, it can be concluded that our proposed Rep-Mobile has achieved the state-of-the-art (SOTA) currently under similar computational and parameter requirements, which also proves the effectiveness of reparameterization for building low-complexity ASC models.

B. Evaluation on the Multi-branch kernels

TABLE II
ABLATION STUDY OF MULTI-BRANCH KERNELS. RESULTS ARE GIVEN WITH ACCURACY (%) UNDER 100% TRAINING DATA. W/O DENOTE WITHOUT. THE MACs AND CHANNELS OF MODELS IS 29M AND 32.

Kernel	-	w/o 3x3	w/o 1x1	w/o 3x1 and 1x3
Acc.	58.59	57.06	57.88	57.47

Table II shows the impacts of multi-kernels within the branches in Rep-Mobile blocks. Noted that experiments are conducted under 100% training data. When we remove any branch, the performance of the model decreases accordingly. This also indicates that convolution kernels of different shapes with prior knowledge can indeed enhance the modeling ability of the model. Moreover, benefited by reparameterization technology, this performance gain does not come with additional computational complexity.

C. Evaluation on the Knowledge Distillation with Ensemble

TABLE III

EVALUATION ACCURACY (%) RESULTS OF CNN AND TRANSFORMER-BASED TEACHER MODELS WITH DIFFERENT TRAINING DATA SIZES. TRAN. DENOTES TRANSFORMER BASED MODELS. PT MEANS PRETRAINED ON AUDIOSET [36]. AVG. IS SHORT FOR AVERAGE.

Models	5%	10%	25%	50%	100%	Avg.
CP-ResNet [15]	44.71	50.39	56.35	59.25	61.71	54.48
CP-Mobile [15]	43.80	50.12	54.71	58.95	60.90	53.70
BC-ResNet [6]	40.50	45.73	52.77	55.46	59.31	50.75
PaSST [20]	50.90	53.18	56.98	58.86	60.20	56.02
PaSST-PT [20]	50.82	53.46	56.76	58.94	60.07	56.01
Ensemble CNN	49.11	54.72	59.93	63.56	66.23	58.71
Ensemble Tran.	52.29	54.90	58.17	60.26	61.44	57.41
Ensemble All	54.38	58.90	63.19	66.38	68.25	62.22

In order to explore the data-efficiency ability of models with different structures for teacher models in knowledge distillation, we compare the performance of common CNNs (CP-ResNet [6], CP-Mobile [15] and BC-ResNet) and Transformers (PaSST [20]) structured networks under different training data size, and the corresponding results are shown in Table III. It should be noted that for CNN based models, we adjusted the width of them to maximize their performance for better teacher models. From the results, it can be seen that both CNNs and Transformers achieve comparable results when there is sufficient data (25%-100%), while Transformers perform better when training data is limited (5% and 10%), demonstrating better data-efficiency ability. Meanwhile, we also compare whether to use AudioSet [36] for pre-training (PaSST vs PaSST-PT), and find that the performance gain of PaSST is not obtained by pre-training with more data, but by the architectural advantages of Transformer itself.

TABLE IV

EVALUATION ACCURACY (%) RESULTS OF KNOWLEDGE DISTILLATION WITH DIFFERENT ENSEMBLE TEACHER MODELS. AVG. IS SHORT FOR AVERAGE. STUDENT MODEL IS REP-MOBILE WITH 32 CHANNELS.

Teachers	5%	10%	25%	50%	100%	Avg.
Baseline w/o KD	43.41	46.76	51.22	54.04	58.59	50.80
Ensemble CNN	44.48	48.90	54.75	57.90	60.65	53.34
Ensemble Tran.	49.06	52.65	57.27	58.65	61.13	55.75
Ensemble All	46.83	51.55	56.94	59.45	61.12	55.18

In Table IV, we also adopted different ensemble methods to explore the effectiveness of different models when distilling knowledge to Rep-Mobile. By comparing with the baseline, using larger models for knowledge distillation on low-complexity models can effectively improve the model performance, and different teacher models exhibit different characteristics. Although the CNN ensemble teacher model has shown great superiority compared to the Transformer teacher, it has not shown corresponding advantages in knowledge distillation for the student model, especially for the setups with less training data. This also indicates that the selection of teacher models is very important in knowledge distillation, and Transformer-based systems is more appropriate in this task.

D. Distillation with Progressive Pruning

TABLE V

EVALUATION ACCURACY (%) RESULTS OF KNOWLEDGE DISTILLATION WITH PRUNING. SGL DENOTES THE SINGLE STEP PRUNING, AND PROG DENOTES THE PROGRESSIVE PRUNING. THE NUMBER OF CHANNELS REPRESENTS THE WIDTH OF THE MODEL, WHICH REFLECTS THE SIZE OF PARAMETERS.

Channels	Prune	5%	10%	25%	50%	100%	Avg.
32	-	49.06	52.65	57.27	58.65	61.13	55.75
64		49.61	54.48	58.48	60.93	63.47	57.39
96		50.84	55.51	59.10	62.19	65.25	58.58
64→32	Sgl	50.89	53.92	58.43	60.42	62.24	57.18
96→32		50.29	54.29	57.37	59.45	62.30	56.74
96→64→32	Prog	51.38	55.26	58.54	60.07	62.30	57.51

In addition, we have explored pruning strategies for building low-complexity models and showed the results in Table V. It should be noted that these results are obtained by distilling from Transformer to Rep-Mobile with different model size. From the analysis of the results, the strategy of training a larger model (64 or 96 channels) first and then pruning and compressing to smaller one (channels 32) can effectively improve the capacity of small model, by comparing with method of directly distilling to smaller one. Our proposed progressive pruning (channels from 96 to 64 then to 32), using multiple and small compression strategies, can effectively reduce the gap and alleviate the performance degradation caused by pruning.

E. Final DCASE2024 Challenge evaluation results

Finally, we provide a brief comparison of the top three results of DCASE2024 challenge evaluation sets in Table VI. With our proposed Rep-Mobile, ensemble transformer distillation and progressive pruning, our system achieves the first place and outperforms other teams with a significant advantage, proving the effectiveness of our strategies.

TABLE VI

FINAL EVAL SCORES AND RANKS OF TASK1 IN DCASE2024 CHALLENGE.

Rank	1 st (Our)	2 nd	3 rd
Avg. Acc. (%)	58.46	57.19	57.15

V. CONCLUSION

In this paper, towards data-efficient low-complexity acoustic scene classification, we propose Rep-Mobile by adding multi-branch kernels and reparameteriz them during the inference process. Furthermore, we provide an analysis on the selection of teacher models for knowledge distillation and find that transformer-based models are more data-efficient and useful for knowledge distillation. Meanwhile, in order to further improve the performance of low-complexity models, we propose a progressive pruning strategy. It can effectively reduce the parameter gap between the models before and after pruning, resulting in better performance compared to the single step pruning. With these methods, we achieved the SOTA results, and won the 1st place in task 1 of DCASE2024 challenge.

REFERENCES

- [1] Daniele Barchiesi, Dimitrios Giannoulis, Dan Stowell, and Mark 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] Irene Martín-Morató, Francesco Paissan, Alberto Ancilotto, Toni Heittola, Annamaria Mesaros, Elisabetta Farella, Alessio Brutti, and Tuomas Virtanen, "Low-complexity acoustic scene classification in dcase 2022 challenge," *arXiv preprint arXiv:2206.03835*, 2022.
- [3] Khaled Koutini, Hamid Eghbal-Zadeh, Matthias Dorfer, and Gerhard Widmer, "The receptive field as a regularizer in deep convolutional neural networks for acoustic scene classification," in *2019 27th European signal processing conference (EUSIPCO)*. IEEE, 2019, pp. 1–5.
- [4] Hu Hu, Chao-Han Huck Yang, Xianjun Xia, Xue Bai, Xin Tang, Yajian Wang, Shutong Niu, Li Chai, Juanjuan Li, Hongning Zhu, Feng Bao, Yuanjun Zhao, Sabato Marco Siniscalchi, Yannan Wang, Jun Du, and Chin-Hui Lee, "A two-stage approach to device-robust acoustic scene classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 845–849.
- [5] Janghoon Cho, Sungrack Yun, Hyoungwoo Park, Jungyun Eum, and Kyuwoong Hwang, "Acoustic scene classification based on a large-margin factorized cnn," *arXiv preprint arXiv:1910.06784*, 2019.
- [6] Byeonggeun Kim, Seunghan Yang, Jangho Kim, and Simyung Chang, "Qti submission to dcase 2021: Residual normalization for device-imbalanced acoustic scene classification with efficient design," *arXiv preprint arXiv:2206.13909*, 2022.
- [7] Tobias Morocutti, Florian Schmid, Khaled Koutini, and Gerhard Widmer, "Device-robust acoustic scene classification via impulse response augmentation," *arXiv preprint arXiv:2305.07499*, 2023.
- [8] Helin Wang, Yuexian Zou, and Wenwu Wang, "SpecAugment++: A hidden space data augmentation method for acoustic scene classification," *arXiv preprint arXiv:2103.16858*, 2021.
- [9] Shayan Gharib, Konstantinos Drossos, Emre Çakir, Dmitriy Serdyuk, and Tuomas Virtanen, "Unsupervised adversarial domain adaptation for acoustic scene classification," *arXiv preprint arXiv:1808.05777*, 2018.
- [10] Laurens Byttebier, Brecht Desplanques, Jenthe Thienpondt, Siyuan Song, Kris Demuyne, and Nilesch Madhu, "Small-footprint acoustic scene classification through 8-bit quantization-aware training and pruning of ResNet models," Tech. Rep., DCASE2021 Challenge, 2021.
- [11] Florian Schmid, Khaled Koutini, and Gerhard Widmer, "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)*. IEEE, 2023, pp. 1–5.
- [12] Xiaohan Ding, Yuchen Guo, Guiguang Ding, and Jungong Han, "Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [13] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun, "Repvgg: Making vgg-style convnets great again," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 13733–13742.
- [14] Pavan Kumar Anasalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan, "Mobileone: An improved one millisecond mobile backbone," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 7907–7917.
- [15] Khaled Koutini, Hamid Eghbal-zadeh, and Gerhard Widmer, "Receptive field regularization techniques for audio classification and tagging with deep convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1987–2000, 2021.
- [16] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al., "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.
- [17] Yiqiang Cai, Peihong Zhang, and Shengchen 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.
- [18] Bing Han, Zhengyang Chen, Bei Liu, and Yanmin Qian, "Mlp-svnet: A multi-layer perceptrons based network for speaker verification," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7522–7526.
- [19] Yiqiang Cai, Li Shengchen, and Shao Xi, "Leveraging self-supervised audio representations for data-efficient acoustic scene classification," *arXiv preprint arXiv:2408.14862*, 2024.
- [20] Khaled Koutini, Jan Schlüter, Hamid Eghbal-Zadeh, and Gerhard Widmer, "Efficient training of audio transformers with patchout," *arXiv preprint arXiv:2110.05069*, 2021.
- [21] Heinrich Dinkel, Yongqing Wang, Zhiyong Yan, Junbo Zhang, and Yujun Wang, "Ced: Consistent ensemble distillation for audio tagging," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 291–295.
- [22] Arshdeep Singh, Dhanunjaya Varma Devalraju, and Padmanabhan Rajan, "Pruning and quantization for low-complexity acoustic scene classification," Tech. Rep., DCASE2021 Challenge, 2021.
- [23] Microsoft, "Neural Network Intelligence," 1 2021.
- [24] Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen, "Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions," *arXiv preprint arXiv:2005.14623*, 2020.
- [25] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, "A multi-device dataset for urban acoustic scene classification," *arXiv preprint arXiv:1807.09840*, 2018.
- [26] Florian Schmid, Paul Primus, Toni Heittola, Annamaria Mesaros, Irene Martín-Morató, Khaled Koutini, and Gerhard Widmer, "Data-efficient low-complexity acoustic scene classification in the dcase 2024 challenge," *arXiv preprint arXiv:2405.10018*, 2024.
- [27] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.
- [28] Byeonggeun Kim, Seunghan Yang, Jangho Kim, Hyunsin Park, Juntae Lee, and Simyung Chang, "Domain generalization with relaxed instance frequency-wise normalization for multi-device acoustic scene classification," *arXiv preprint arXiv:2206.12513*, 2022.
- [29] Florian Schmid, Tobias Morocutti, Shahed Masoudian, Khaled Koutini, and Gerhard Widmer, "Cp-jku submission to dcase23: Efficient acoustic scene classification with cp-mobile," Tech. Rep., DCASE2023 Challenge, Tech. Rep. 2023.
- [30] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang, "Domain generalization with mixstyle," *arXiv preprint arXiv:2104.02008*, 2021.
- [31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [32] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuyne, "Ecapadtnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.
- [33] Soonshin Seo and J Kim, "Mobilenet using coordinate attention and fusions for low-complexity acoustic scene classification with multiple devices," Tech. Rep., DCASE2021 Challenge, 2021.
- [34] Hossein Zeinali, Shuai Wang, Anna Silnova, Pavel Matějka, and Oldřich Plchot, "But system description to voxceleb speaker recognition challenge 2019," *arXiv preprint arXiv:1910.12592*, 2019.
- [35] Anbai Jiang, Qijun Hou, Jia Liu, Pingyi Fan, Jitao Ma, Cheng Lu, Yuanzhi Zhai, Yufeng Deng, and Wei-Qiang Zhang, "Thuee system for first-shot unsupervised anomalous sound detection for machine condition monitoring," *Proceedings of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events, Tampere, Finland*, pp. 20–22, 2023.
- [36] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.