KNOWLEDGE DISTILLATION FROM TRANSFORMERS FOR LOW-COMPLEXITY ACOUSTIC SCENE CLASSIFICATION

Florian Schmid[1,2]*, Shahed Masoudian*[2]*, Khaled Koutini*[2]*, Gerhard Widmer*[1,2]

[1]Institute of Computational Perception (CP-JKU),[2]LIT Artificial Intelligence Lab,
Johannes Kepler University Linz, Austria
florian.schmid@jku.at, shahed.masoudian@jku.at, khaled.koutini@jku.at

ABSTRACT

Knowledge Distillation (KD) is known for its ability to compress large models into low-complexity solutions while preserving high predictive performance. In Acoustic Scene Classification (ASC), this ability has recently been exploited successfully, as underlined by three of the top four systems in the low-complexity ASC task of the DCASE'21 challenge [1] relying on KD. Current KD solutions for ASC mainly use large-scale CNNs or specialist ensembles to derive superior teacher predictions. In this work, we use the *Audio Spectrogram Transformer* model *PaSST*, pre-trained on Audioset, as a teacher model. We show how the pre-trained *PaSST* model can be properly trained downstream on the *TAU Urban Acoustic Scenes 2022 Mobile development dataset* [2] and how to distill the knowledge into a low-complexity CNN student. We study the effect of using teacher ensembles, using teacher predictions on extended audio sequences, and using Audioset as an additional dataset for knowledge transfer. Additionally, we compare the effectiveness of Mixup and Freq-MixStyle to improve performance and enhance device generalization. The described system achieved rank 1 in the Low-complexity ASC Task of the DCASE'22 challenge [3][1].

Index Terms— Patchout Spectrogram Transformer, Mixup, Freq-MixStyle, Knowledge Distillation

1. INTRODUCTION

Acoustic Scene Classification (ASC) aims to build a model that accurately predicts a scene label given an audio recording. Convolutional Neural Networks (CNNs) are well-established models that dominate the field of ASC [1, 4–6]. Low-complexity solutions are of increasing interest, making the ASC systems runnable on edge devices. Model compression techniques to reduce complexity include *Parameter Pruning* [7–9], *designing efficient network architectures* [10, 11] and *Knowledge Distillation (KD)* [12–14]. KD recently showed promising results for ASC, with three out of the top four systems applying KD in the low-complexity ASC task of the DCASE'21 challenge [1].

This paper focuses on distilling knowledge from Patchout faSt Spectrogram Transformers (PaSST) [15] into low-complexity CNNs, the winning approach of the DCASE'22 low-complexity ASC task [3]. The *TAU Urban Acoustic Scenes 2022 Mobile development dataset* [2] and the low-complexity constraints of the DCASE'22 challenge define an ASC scenario involving various challenges:

- C1 Limits on the number of parameters and the number of multiply-accumulate operations.
- C2 Scenes are recorded using different recording devices, resulting in a domain shift.
- C3 1-second audio snippets are provided compared to 10-second snippets in the ASC task of the DCASE'21 challenge [1]. Shorter audio files contain less information and are more challenging to classify.

The system studied in the following tackles C1 by using KD with PaSST [15] models as a teacher and a low-complexity CNN as a student. Regarding C2, the domain shift caused by different recording devices is counteracted by *Freq-MixStyle* [16], a modified version of *MixStyle* [17], operating on frequency statistics. The 1-second audio snippets described in C3 result from splitting the 10-second snippets of the *TAU Urban Acoustic Scenes 2020 Mobile development dataset* [2] into ten pieces. The provided segment identifiers allow for reassembling the complete 10-second files. We train all models on randomly cropped 1-second pieces from the reassembled 10-second files to increase the diversity in the training data. Reassembling the 10-second files additionally allows for obtaining superior teacher predictions for KD, as described later.

The related work regarding the system's building blocks is described in Section 2. Section 3 introduces student and teacher architectures and the PaSST downstream training. Section 4 describes how student and teacher are combined in a KD setup and sets the stage for the results presented in Section 5. The paper is concluded in Section 6.

2. RELATED WORK

2.1. ASC Architectures

In the past years, Convolutional Neural Networks (CNNs) became the most prominent solution to process spectrograms and dominated previous DCASE challenges [1, 4–6]. Restricting the receptive field of CNNs, known as Receptive Field Regularization, was shown to be particularly well suited for ASC tasks [18, 19].

Inspired by Vision Transformers (ViT) [20], transformers capable of processing spectrograms to solve audio tasks have been proposed recently. In this regard, Audio Spectrogram Transformers (AST) [21], pre-trained on computer vision tasks, have been adapted to the audio domain and achieved state-of-the-art results on Audioset [22]. Image [20, 23] and spectrogram [15, 21] transformers extract overlapping patches with a certain stride from the input image and add a positional encoding. Patchout faSt Spectrogram Transformer (PaSST) [15] disentangles frequency- and time-encodings, simplifying downstream training on shorter audio clips. PaSST additionally introduces *Patchout*, a mechanism that drops parts of the input sequence to improve generalization and reduce

[1]Source Code:https://github.com/CPJKU/cpjku_dcase22

memory and computational complexity in the quadratically scaling attention layers.

2.2. Knowledge Distillation in ASC

In KD [12–14], the aim is to compress the knowledge of a possibly large number of complex teacher models into a low-complexity student model without significant performance loss. The student minimizes a weighted sum of hard label loss and distillation loss, as shown in Eq. 1. The distillation loss matches student and teacher predictions based on the Kullback-Leibler divergence between soft targets: $L_{\mathrm{DIST}} = D_{\mathrm{KL}}(q_{\mathrm{teacher}}||q_{\mathrm{student}})$. By minimizing the distillation loss, the student learns to mimic the teacher.

$$L_{\mathrm{TOTAL}} = L_{\mathrm{LABEL}} + \lambda L_{\mathrm{DIST}} \tag{1}$$

The soft targets q are created from the logits z by computing $q = \mathrm{softmax}(z/T)$ with a specific temperature T. Raising the temperature T creates a softer distribution and allows the student to exploit the rich similarity information between classes predicted by the teacher compared to the hard labels [12].

In the low-complexity task of the DCASE'21 challenge [1], the top two systems included KD based on a pre-trained CNN teacher network [4] and a large two-stage fusion CNN teacher model [5].

Jung et al. [24] showed that learning from soft targets in a teacher-student setup has a beneficial effect as one-hot labels do not reflect the blurred decision boundaries between different acoustic scenes. Teacher superiority is achieved by using multiple audio segments of the same scene recorded at different locations to generate teacher predictions.

Another popular idea to generate superior teacher predictions is to use ensembles of specialist models. In this regard, ensembles of device-experts [25], ensembles of models trained on different audio representations [26], and ensembles of specialist models for confusing pairs of acoustic scenes [27] have been studied and successfully applied.

2.3. Mixup and Freq-MixStyle

Mixup [28] constructs virtual training samples by linearly mixing two existing samples and their labels. In particular, the coefficients of the convex sample and label combinations are drawn randomly from a Beta distribution, its shape specified by a parameter α. Mixup has been shown to improve generalization on ASC tasks before [29].

MixStyle [17] is introduced to enhance domain generalization by mixing channel-wise statistics of images. However, the device-style in spectrograms primarily resides in the frequency-wise statistics. To enhance generalization across recording devices, an adapted version of MixStyle, *Freq-MixStyle*, is proposed in [16]. It proceeds by normalizing each frequency band and denormalizing it with mixed frequency coefficients of two different samples. *Freq-MixStyle* is guided by two parameters: α determines the shape of the Beta distribution used to randomly draw mixing coefficients, and p specifies the probability of whether it is applied to a batch or not.

3. MODEL SPECIFICATIONS

3.1. Student Model: Compact RFR-CNN

The student model is a Receptive Field Regularized [18, 19] Convolutional Neural Network (RFR-CNN) and is based on *CP_ResNet* which performed well in previous editions of the DCASE ASC challenge [1, 2, 6, 29, 30]. The initial width is reduced to $W = 32$ channels, and a grouping of 2 is applied to the penultimate block to obtain a compact model. The max-pooling layers (indicated by P_f) are adapted to perform pooling only over the frequency dimension to account for the shorter audio clips (1-second). This allows us to downscale spectrogram dimensionality while preserving temporal information. Finally, $C = 36$ channels are cut from the final residual block of the network to conform to the low-complexity limitations. Table 1 summarizes the architecture of the CNN used as the student.

Table 1: Low-complexity Student Model realized by a compact CP-ResNet Architecture.

WIDTH	GROUPING	BLOCK	CONFIG
W		INPUT	5×5, P
W	1	R	3×3, 1×1, P_f
W	1	R	3×3, 3×3, P_f
		LINEAR	$W \rightarrow 2\dot{W}$
$2 \times W$	2	R	3×3, 3×3
		LINEAR	$2W \rightarrow 4\dot{W}$
$4 \times W - C$	1	R	3×3, 1×1
CLASSIFIER $4 \times W - C \rightarrow 10$ CLASSES			
GLOBAL MEAN POOLING			

P: 2×2 MAX POOLING.
P_f: 2×1 MAX POOLING OVER THE FREQUENCY DIMENSION.
R: RESIDUAL, THE INPUT IS ADDED TO THE OUTPUT

3.2. Teacher Model: PaSST

The main criterion for selecting teacher models is the accuracy of the predictions to reflect class similarity structures. The choice of PaSST transformer models as teachers is motivated by their high performance on downstream tasks – for instance, achieving an accuracy of 76.3% [15] on the *TAU Urban Acoustic Scenes 2020 Mobile development dataset* [2] –, and their ability to recognize fine-grained acoustic events after being pre-trained on the 527 classes of Audioset [22]. The PaSST models selected for downstream training on the *TAU Urban Acoustic Scenes 2022 Mobile development dataset* [2] are pre-trained on Audioset and extract patches of size 16x16 with a stride of 10. We use a *structured Patchout* of 6 *only* on the frequency dimension, which means that 6 frequency bands are dropped at random during training. This is an important countermeasure to prevent overfitting on the downstream dataset [3]. Due to the short length of the audio clips, we do not apply Patchout across the time dimension. The experimental setup is the same as provided in the system's technical report [31].

Table 2 compares a PaSST Baseline model with *Mixup* and *Freq-MixStyle* augmentation techniques. *Freq-MixStyle* is only applied to the raw spectrograms. While the effect on real devices is limited, *Mixup* and *Freq-MixStyle* substantially improve over the Baseline on simulated and unseen devices. In particular, on unseen devices, *Freq-MixStyle* significantly outperforms *Mixup*, underlining its superior device generalization capabilities and leading to an overall performance gain compared to *Mixup*. *Ensemble* denotes averaged logits of five PaSST models trained with different Freq-MixStyle configurations. The most significant performance gain of

Method	Real Devices				Simulated Devices				Unseen Devices				Overall
	A	B	C	**Real**	S1	S2	S3	**Sim**	S4	S5	S6	**Unseen**	
PaSST Baseline	72.00	63.29	67.59	67.63	58.13	56.61	58.25	57.66	57.05	57.62	53.67	56.11	60.46
+ Mixup	**72.65**	62.86	**68.04**	**67.85**	59.24	**57.30**	58.81	58.45	57.99	58.24	55.07	57.10	61.13
+ Freq-MixStyle	72.14	**63.55**	67.32	67.68	**59.69**	57.24	**59.99**	**58.97**	**59.03**	**58.65**	**56.98**	**58.22**	**61.64**
Ensemble	73.70	64.07	68.09	68.62	61.30	58.24	60.79	60.11	60.39	60.03	58.76	59.73	62.82

Table 2: PaSST downstream training on the *TAU Urban Acoustic Scenes 2022 Mobile development dataset* [2] following the official test split: Device-wise comparison between Baseline, Mixup ($\alpha = 0.3$), Freq-MixStyle ($\alpha = 0.4, p = 0.4$) and an ensemble of five PaSST models trained with different Freq-MixStyle configurations. The provided accuracies (%) are averaged over three runs and the last 10 epochs of training. The devices are grouped according to real devices (**Real**: A, B, C), the seen, simulated devices (**Sim**: S1, S2, S3) and the unseen, simulated devices (**Unseen**: S4, S5, S6).

the ensemble occurs with unseen devices, underlining the ensemble's robustness.

4. KNOWLEDGE DISTILLATION FROM PASST TO CNN

Knowledge Distillation (KD) [12–14] is the concept used to transfer the knowledge of the well-performing, large PaSST transformer from Section 3.2 to the low-complexity RFR-CNN introduced in Section 3.1. PaSST derives its superiority from its size and the pre-training on Audioset [22]. Hence, in a KD setup, PaSST implicitly passes its semantically rich understanding of acoustic scenes gained from the 527 classes of Audioset to the low-complexity student model. We limit the knowledge transfer to the logits and leave experiments on mimicking embeddings as an interesting future research direction. In Section 5, we experiment with KD and probe for a positive performance impact of the following three variations:

- **Teacher Ensemble**: We ensemble five PaSST models with different Freq-MixStyle configurations by averaging their logits. Freq-MixStyle trained models with different values for α and p tend to have different strengths on real, simulated, and unseen devices, leading to a robust and well-performing ensemble, as shown in Table 2.
- **Superior Teacher**: Given that the full 10-second audio files can be reassembled from the 1-second pieces, the distillation loss can be additionally based on teacher predictions for the full 10 seconds. This way, the student has to match superior teacher predictions while having only access to one-tenth of the input sequence. Eq. 2 presents the adapted loss calculation, adding the new distillation loss $L_{\text{DIST\_SUP}}$ with its corresponding weight λ_{SUP}.

$$L_{\text{TOTAL}} = L_{\text{LABEL}} + \lambda L_{\text{DIST}} + \lambda_{\text{SUP}} L_{\text{DIST\_SUP}} \quad (2)$$

- **Distillation on Out-of-Domain Dataset**: In addition to KD on the *TAU Urban Acoustic Scenes 2022 Mobile development dataset* [2], we experiment with transferring the knowledge using Audioset [22]. For each batch, we sample a batch of Audioset samples of the same size, generate teacher predictions, and calculate the distillation loss. The procedure results in a total loss calculated as shown in Eq. 3.

$$L_{\text{TOTAL}} = L_{\text{LABEL}} + \lambda(L_{\text{DIST}} + L_{\text{DIST\_AUDIOSET}}) \quad (3)$$

When computing student and teacher predictions in the context of Mixup, the same audio snippets are mixed using the same mixing coefficients. Freq-Mixstyle is applied to student and teacher spectrograms independently, which forces the student to match the teacher's soft targets in the context of different device-styles.

4.1. Experimental Setup

Preprocessing: The raw audio signal is down-sampled using a sampling rate of 32 kHz. Spectrograms are generated by applying Short Time Fourier Transformation with a window size of 2048 and an overlap of 744 (approximately 36%) in case of the students and a window size of 800 with an overlap of 320 (40%) for the teachers. A Mel-scaled filter bank is applied to create spectrograms with 256 and 128 mel bins for students and teachers, respectively. The applied preprocessing matches the teacher's pre-training, while, for the student, a higher frequency resolution proved beneficial.
Training: All students in the KD framework are trained with a batch size of 64 for a total amount of 750 epochs, where the models only process one-tenth of the available data each epoch because of the random 1-second cropping. Adam optimizer with a specific learning rate schedule is applied. The learning rate is exponentially increasing to 1×10^{-3} until epoch 150 and linearly decreasing from epoch 250 until epoch 650, dropping to a value of 5×10^{-6}.

5. RESULTS

A summary of the results for the RFR-CNN student model, as presented in Section 3.1, is shown in Table 3. The results are categorized into *Student Baseline* (no KD), *KD Baseline* (*PaSST + Freq-MixStyle* as teacher) and the three KD variations presented in Section 4. In the following, we describe the effect of Freq-MixStyle compared to Mixup, the effect of KD, and the impact of the KD variations.

5.1. Mixup vs. Freq-MixStyle

We investigated different Mixup and Freq-MixStyle configurations and observed that Mixup with $\alpha = 0.4$ and Freq-MixStyle with $\alpha = 0.3$ and $p = 0.4$ yield robust results across a variety of configurations. For *Student Baseline*, Freq-MixStyle outperforms Mixup significantly, achieving the highest performance gains on the unseen device category. In combination with KD, Freq-MixStyle is still slightly superior to Mixup in terms of accuracy but Mixup leads to lower log losses. Freq-MixStyle generalizes much better to unseen devices than Mixup but weakens the performance on real devices.

5.2. Effectiveness of KD

We experiment with three different temperature configurations and adapt the distillation loss weight λ for each temperature. We select the best of the three settings **High (H)** (T=8, λ=800), **Medium (M)** (T=3, λ=100) and **Low (L)** (T=1, λ=50) to be listed in each row of Table 3. KD with Mixup requires high temperatures, while Freq-MixStyle tends to favour low temperature settings.

Method	Configuration					Test Accuracy (%)				Log Loss
	Mixup	Freq-MixStyle	Temp	Teach. Type	AS	Real	Sim	Unseen	Overall	Overall
Student Baseline	✗	✗	-	No	✗	61.97	50.10	40.71	50.92	1.5822
	✓	✗	-	No	✗	62.70	52.48	42.99	52.72	1.4161
	✗	✓	-	No	✗	63.89	56.00	49.98	56.62	1.2344
KD Baseline	✗	✗	H	Single	✗	66.21	57.35	50.14	57.89	1.1316
	✓	✗	H	Single	✗	66.43	58.31	51.32	58.68	1.1063
	✗	✓	L	Single	✗	64.36	58.36	55.12	59.28	1.1431
KD Ensemble	✓	✗	H	Ensemble	✗	66.30	58.65	52.06	59.00	1.0888
	✗	✓	L	Ensemble	✗	64.74	58.59	55.14	59.49	1.1322
KD Superior Teacher	✓	✗	H	Superior	✗	66.53	58.54	51.89	58.98	1.1033
	✗	✓	L	Superior	✗	64.73	58.60	**55.15**	59.49	1.1313
KD Audioset	✓	✗	M	Single	✓	**66.54**	59.09	52.49	59.37	1.0906
	✗	✓	M	Single	✓	64.99	58.50	54.43	59.30	1.0939
	✓	✗	M	Ensemble	✓	66.35	**59.95**	52.99	**59.76**	**1.0794**

Table 3: Results of the low-complexity RFR-CNN student model on the official test split: Accuracies (%) for the device groups **Real** (A, B, C), **Sim** (S1, S2, S3) and **Unseen** (S4, S5, S6), overall accuracy and overall log loss are compared between a baseline using no KD, a baseline using KD and KD variations. Configuration **Temp** is referring to **High (H)** (T=8), **Medium (M)** (T=3) and **Low (L)** (T=1) temperature when calculating teacher and student soft targets, and **AS** indicates the use of Audioset [22] for KD. All results presented are averages of 3 independent runs averaged over the last 10 epochs of training.

KD Baseline outperforms *Student Baseline*, leading to an overall accuracy improvement of 2.66% when comparing the Freq-MixStyle configurations. While the accuracy on real devices improves only slightly, the accuracy on unseen devices increases by 5.14%.

5.3. Effectiveness of KD variations

Using the PaSST ensemble from Table 2 as the teacher improves the results slightly but consistently for both Freq-MixStyle and Mixup, in terms of overall accuracies and log losses. However, the performance gain of more than 1 percentage point in terms of overall accuracy that the ensemble gives compared to a single PaSST model cannot be transferred to the student.

We investigate a range of λ_{SUP} values for each temperature setting for the superior teacher variation. The superior teacher predictions on the 10-second snippets are computed offline. In case of Mixup, the superior predictions are mixed accordingly and no Freq-MixStyle is applied to infer the predictions. *KD Superior Teacher* performs similar to *KD Ensemble* using $\lambda_{\text{SUP}} = 3.0$ for the Mixup configuration and $\lambda_{\text{SUP}} = 1.0$ for the Freq-MixStyle configuration.

With *KD Audioset*, Freq-MixStyle is applied to samples from both datasets, while Mixup is not applied to the samples from Audioset since no hard labels are available. The Mixup configuration shows the best results in terms of accuracy across all Mixup settings, while the Freq-MixStyle configuration only leads to a minor accuracy improvement over *KD Baseline*. However, the Freq-MixStyle configuration achieves the lowest log loss across all Freq-MixStyle experiments. Combining Audioset with *KD Ensemble* and using Mixup leads to the highest overall accuracy and the lowest log loss.

As a final investigation, Figure 1 compares the impact of Freq-MixStyle, Mixup and the KD variations on the overall and the unseen device accuracies. The dominating factor for enhanced generalization to unseen devices is Freq-MixStyle, which clearly outperforms Mixup. The KD variations slightly improve over the KD baseline in terms of overall accuracy but have no significant impact on unseen device performance.

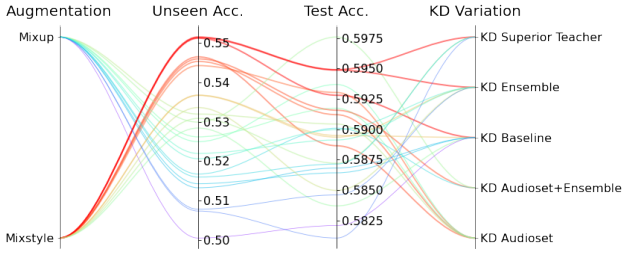

Figure 1: Comparison of the different KD variations and augmentation techniques and their effect on the overall accuracy and the performance on unseen devices. Each line depicts an average over three runs and is colored according to the performance on unseen devices. In addition to the configurations shown in the figure, experiments differ in terms of temperature setting (**High, Medium, Low**).

6. CONCLUSION

In this paper, we distilled the knowledge of a PaSST transformer model into a low-complexity CNN. We showed how the pre-trained PaSST model can be effectively adapted to a downstream task. CNN students that are taught by PaSST models perform significantly better than CNNs learning only from the hard class labels. Based on this, we experiment with three KD variations, including a PaSST teacher ensemble, a superior teacher and KD on Audioset, that show promising performance compared to the KD baseline. To enhance generalization to unseen devices, we compared Mixup with Freq-MixStyle and observed that Freq-MixStyle leads to high accuracy improvements on unseen devices for both PaSST teacher and CNN student models.

7. ACKNOWLEDGMENT

The LIT AI Lab is financed by the Federal State of Upper Austria.

8. REFERENCES

[1] I. Martín-Morató, T. Heittola, A. Mesaros, and T. Virtanen, "Low-complexity acoustic scene classification for multi-device audio: Analysis of DCASE 2021 challenge systems," in *DCASE 2021 Workshop*, 2021.

[2] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions," in *DCASE 2020 Workshop*, 2020, pp. 56–60.

[3] I. Martín-Morató, F. Paissan, A. Ancilotto, T. Heittola, A. Mesaros, E. Farella, A. Brutti, and T. Virtanen, "Low-complexity acoustic scene classification in dcase 2022 challenge," 2022.

[4] B. Kim, S. Yang, J. Kim, and S. Chang, "QTI submission to DCASE 2021: Residual normalization for device-imbalanced acoustic scene classification with efficient design," DCASE2021 Challenge, Tech. Rep., 2021.

[5] C.-H. H. Yang, H. Hu, S. M. Siniscalchi, Q. Wang, W. Yuyang, X. Xia, Y. Zhao, Y. Wu, Y. Wang, J. Du, and C.-H. Lee, "A lottery ticket hypothesis framework for low-complexity device-robust neural acoustic scene classification," DCASE2021 Challenge, Tech. Rep., 2021.

[6] K. Koutini, F. Henkel, H. Eghbal-zadeh, and G. Widmer, "CP-JKU Submissions to DCASE'20: Low-Complexity Cross-Device Acoustic Scene Classification with RF-Regularized CNNs," DCASE2020 Challenge, Tech. Rep., 2020.

[7] S. A. Janowsky, "Pruning versus clipping in neural networks," *Physical Review A*, vol. 39, no. 12, p. 6600, 1989.

[8] J. Frankle, G. K. Dziugaite, D. Roy, and M. Carbin, "Pruning neural networks at initialization: Why are we missing the mark?" in *ICLR*, 2021.

[9] C. Liu, Z. Zhang, and D. Wang, "Pruning deep neural networks by optimal brain damage," in *INTERSPEECH*, 2014, pp. 1092–1095.

[10] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, 2017.

[11] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *ICML*, vol. 97, 2019, pp. 6105–6114.

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

[13] J. Ba and R. Caruana, "Do deep nets really need to be deep?" in *NIPS*, 2014, pp. 2654–2662.

[14] C. Bucila, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *ACM SIGKDD*, 2006, pp. 535–541.

[15] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," *CoRR*, vol. abs/2110.05069, 2021.

[16] 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," *CoRR*, vol. abs/2206.12513, 2022.

[17] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with mixstyle," in *ICLR*, 2021.

[18] K. Koutini, H. Eghbal-zadeh, and G. 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.

[19] K. Koutini, H. Eghbal-zadeh, M. Dorfer, and G. Widmer, "The Receptive Field as a Regularizer in Deep Convolutional Neural Networks for Acoustic Scene Classification," in *EUSIPCO*, 2019.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[21] Y. Gong, Y. Chung, and J. R. Glass, "AST: audio spectrogram transformer," in *Interspeech*, 2021, pp. 571–575.

[22] 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 *ICASSP*, 2017.

[23] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *ICML 2021*, vol. 139, 2021, pp. 10 347–10 357.

[24] H. Heo, J. Jung, H. Shim, and H. Yu, "Acoustic scene classification using teacher-student learning with soft-labels," in *Interspeech*, 2019, pp. 614–618.

[25] S. Takeyama, T. Komatsu, K. Miyazaki, M. Togami, and S. Ono, "Robust acoustic scene classification to multiple devices using maximum classifier discrepancy and knowledge distillation," in *EUSIPCO*, 2020, pp. 36–40.

[26] L. Gao, K. Xu, H. Wang, and Y. Peng, "Multi-representation knowledge distillation for audio classification," *Multim. Tools Appl.*, vol. 81, no. 4, pp. 5089–5112, 2022.

[27] J. Jung, H. Heo, H. Shim, and H. Yu, "Knowledge distillation in acoustic scene classification," *IEEE Access*, vol. 8, pp. 166 870–166 879, 2020.

[28] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *ICLR*, 2018.

[29] K. Koutini, H. Eghbal-zadeh, and G. Widmer, "CP-JKU submissions to DCASE'19: Acoustic scene classification and audio tagging with receptive-field-regularized CNNs," DCASE2019 Challenge, Tech. Rep., 2019.

[30] K. Koutini, S. Jan, and G. Widmer, "CPJKU Submission to DCASE21: Cross-Device Audio Scene Classification with Wide Sparse Frequency-Damped CNNs," DCASE2021 Challenge, Tech. Rep., 2021.

[31] 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," DCASE2022 Challenge, Tech. Rep., 2022.