INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Fine-tuning Audio Spectrogram Transformer with Task-aware Adapters for Sound Event Detection

Kang Li[1], *Yan Song*[1], *Ian McLoughlin*[1,2], *Lin Liu*[3], *Jin Li* [1], *Li-Rong Dai*[1]

[1]National Engineering Research Centre of Speech and Language Information Processing, University Of Science And Technology Of China, Hefei, China.
[2]ICT Cluster, Singapore Institute of Technology, Singapore.
[3]iFLYTEK Research, iFLYTEK Co. Ltd, Hefei, China.

likang0311@mail.ustc.edu.cn, songy@ustc.edu.cn

Abstract

In this paper, we present a task-aware fine-tuning method to transfer Patchout faSt Spectrogram Transformer (PaSST) model to sound event detection (SED) task. Pretrained PaSST has shown significant performance on audio tagging (AT) and SED tasks, but it is not optimal to fine-tune the model from a single layer as the local and semantic information have not been well exploited. To address this, we first introduce task-aware adapters including SED-adapter and AT-adapter to fine-tune PaSST for SED and AT task respectively, and then propose task-aware fine-tuning to combine local information from shallower layer with semantic information from deeper layer, based on task-aware adapters. Besides, we propose the self-distillated mean teacher (SdMT) to train a robust student model with soft pseudo labels from teacher. Experiments are conducted on DCASE2022 task4 development set, the EB-F1 of 64.85% and PSDS1 of 0.5548 are achieved which outperform previous state-of-the-art systems.

Index Terms: sound event detection, transformer, task-aware, fine-tune, mean teacher

1. Introduction

Sound Event Detection (SED) is a task of determining both the categories and timestamps of multiple overlapped events within a given audio clip. It has wide applications for real world systems including smart home devices [1] and automatic surveillance [2]. Access to large corpora with strongly-labeled sound events is expensive and difficult in engineering applications, weakly-supervised SED task has been set up by DCASE challenges [1] to evaluate the progress of SED research.

In the past DCASE challenges, due to the shortage of available labeled data, semi-supervised learning (SSL) based SED methods have drawn increasing research interest. Among different SSL methods [3, 4, 5], mean teacher (MT) [5] has achieved significant SED performance. Other SSL methods such as interpolation consistency training (ICT) [6], shift consistency training (SCT) [7], and confident mean teacher (CMT) [8] have been proposed to exploit unlabeled data efficiently. Designing extra audio tagging (AT) model or branch [7, 9, 10] to guide the SED model learning has shown helpful for SED performance. Convolutional Recurrent Neural Network (CRNN) [11] is commonly used as backbone to perform frame-level feature extraction and context modeling for SED. To improve the feature extraction ability of basic convolution, stronger convolution blocks with attention mechanism have been proposed such as event-aware module [12], strip pooling based attention module [13], frequency dynamic convolution (FDConv) [14] and multi-dimensional frequency dynamic convolution [8] (MFD-Conv). In [15], the Conformer, a convolution-augmented Transformer architecture, is introduced for modeling both local and global context information.

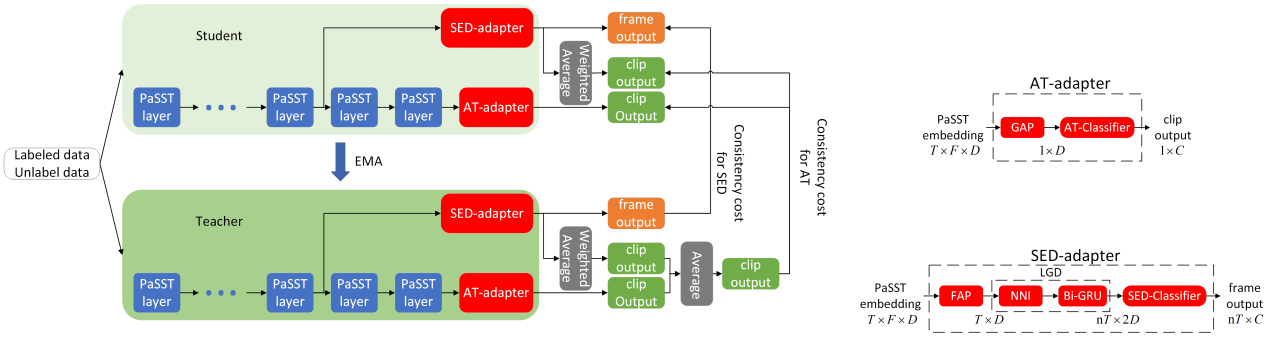
In DCASE2022, several works were proposed to exploit external large-scale weakly-labeled AudioSet [16] data. One way is pretraining SED models on AudioSet, for example, the forward-backward CRNN (FB-CRNN) and Bi-directional CRNN (Bi-CRNN) [17] are firstly pretrained on AudioSet, then with self-training method, the FB-CRNN is firstly fine-tuned for AT task, then the Bi-CRNN is fine-tuned for SED task with the weak pseudo label from FB-CRNN, which achieves the first rank in DCASE2022 task4. Another way is fine-tuning pretrained AT models for SED task, for example, Xiao [18] used RNNs to context model the output of the pretrained audio neural network (PANN) [19] and audio spectrogram transformer (AST) [20]. As it is not optimal to directly use the output from pretrained AT model, in our previous work (*i.e.*, AST-SED) [21], the Encoder-Decoder block consisting of frequency-wise transformer encoder (FTE) and local GRU decoder (LGD) is proposed to effectively fine-tune AST for SED, it extracts discriminative temporal representation with self-attention, and further produces a high-temporal-resolution representation, which is beneficial for SED task. We introduce the SED-adapter as shown in Figure 1(b) to fine-tune PaSST (the improved AST model) [22] for SED, where only the LGD block is used as the frequency-wise attention has been considered in PaSST. The PaSST-SED model (*i.e.*, PaSST with SED-adapter) achieves higher performance compared with AST-SED due to the well-pretrained PaSST.

However, it is still not optimal to fine-tune PaSST with only attaching the SED-adapter to a single layer as AST-SED do, as the semantic and local information have not been well combined and exploited, and the two types of information may not be the strongest synchronously at the same layer. Besides, the hard strong-label in the training data may be noisy as the timestamps are hard to determined [23], the noisy-label issue is deserved to be explored further.

In this paper, we propose a task-aware fine-tuning method, as shown in Figure 1(a), to exploit the local and semantic information efficiently, based on the task-aware adapters as shown in Figure 1(b). Specifically, the SED-adapter is attached to shallower layer of pretrained PaSST to exploit the local information and the AT-adapter is attached to deeper layer to exploit the semantic information, which is a better way to fine-tune PaSST for SED task. We further propose the self-distillated mean teacher (SdMT), as shown in Figure 3, to train a robust vice-student model with soft knowledge distillation [24], which can reduce the adverse impact of noisy labels. Extensive experiments have been conducted on the DCASE2022 challenge

[1]https://dcase.community/challenge2022/task-sound-event-detection-in-domestic-environments

10.21437/Interspeech.2023-1174



(a) Task-aware fine-tuning (b) Task-aware adapter

Figure 1: *(a) Task-aware fine-tuning: a SED-adapter is attached to shallower layer to exploit local information and a AT-adapter is attached to deeper layer to exploit global semantic information for efficient fine-tuning. (b) Task-aware adapters: the SED-adapter is used to fine-tuning for SED task, the AT-adapter is used to fine-tuning for AT task. GAP and FAP denotes global average pooling and frequency-wise average pooling respectively. LGD denotes local GRU decoder [21], where the NNI denotes nearest neighbour interpolation. For simplicity, the classification loss is not indicated in the figure.*

task4 development set to evaluate the proposed task-aware fine-tuning and SdMT. Specifically, a performance of 64.85% Event-based F1-score (EB-F1) and 0.5548 Polyphonic Sound detection Score scenario1 (PSDS1) is achieved, significantly outperforming the 59.60% and 0.5140 of the previous state-of-the-art AST-SED.

2. Proposed Methods

In this section, we briefly introduce the baseline model, where the SED-adapter is attached to single layer of PaSST to fine-tune and mean teacher (MT) is also used as SSL method, and then analyze its shortcomings. We then present the proposed: (1) task-aware fine-tuning with both SED-adapter and AT-adapter, (2) self-distillated mean teacher (SdMT).

2.1. Baseline: fine-tune PaSST with SED-adapter and MT

As shown in Figure 1(b), the SED-adapter consists of three blocks: (1) frequency-wise average pooling (FAP) to extract a frame-level representation, (2) local GRU decoder (LGD) to produce a high-temporal-resolution representation, (3) SED classifier to produce frame-level SED output. The up-sampling ratio in the LGD is 10. With attaching SED-adapter, the pretrained PaSST is fine-tuned for SED task as shown in Figure 1(a) (in the baseline, the AT-adapter is not used), the model, termed as PaSST-SED, is fine-tuned with the loss function determined as follows,

$$L_{baseline} = L^{sed}_{BCE,frame} + \lambda^{sed}_1 L^{sed}_{BCE,clip} + \lambda^{sed}_2 L^{sed}_{MSE,frame} + \lambda^{sed}_3 L^{sed}_{MSE,clip} \quad (1)$$

where $L^{sed}_{BCE,frame}$ denotes frame-level classification BCE loss for strongly-labeled data, $L^{sed}_{BCE,clip}$ denotes clip-level classification BCE loss for weakly-labeled data, $L^{sed}_{MSE,frame}$ and $L^{sed}_{MSE,clip}$ denote frame-level and clip-level teacher-student consistency MSE loss for unlabeled data respectively. Referenced to [14], the weight λ^{sed}_1, λ^{sed}_2, λ^{sed}_3 is set to 0.5, 2, 2 respectively. The clip-level output y_{clip} is a weighted average from frame-level output y_{frame} with linear-softmax pooling [25],

$$y_{clip} = \sum_{i=0}^{T} y^2_{frame,i} / \sum_{i=0}^{T} y_{frame,i} \quad (2)$$

where i denotes th i^{th} frame. The teacher model is an exponential moving average (EMA) from student model.

Although the PaSST-SED outperforms AST-SED, it is still not optimal as the model is only fine-tuned from a single layer and richer semantic information is not fully exploited. In the next subsection, we will detail the proposed task-aware fine-tuning where the PaSST is fine-tuned from two layers with two types of adapters, and we further present the self-distillated mean teacher (SdMT) where a vice-student is trained with soft knowledge distillation to reduce the adverse impact of noisy labels.

2.2. Improved fine-tuning methods

2.2.1. Task-aware fine-tuning: fine-tune PaSST with both SED-adapter and AT-adapter

Before introducing task-aware fine-tuning, we first introduce the AT-adapter, as shown in Figure 1(b), the AT-adapter consists of: (1) Global average pooling (GAP) to extract a clip-level representation, (2) AT classifier to produce a clip-level output. With AT-adapter, the pretrained PaSST is fine-tuned for AT task as shown in Figure 1(a) (SED adapter is not used), the mean teacher method is also used, and the loss function is determined as follows,

$$L_{AT} = L^{at}_{BCE,clip} + \lambda^{at}_1 L^{at}_{MSE,clip} \quad (3)$$

where the weight λ^{at}_1 is set to 1, $L^{at}_{BCE,clip}$ denotes classification BCE loss for weakly-labeled data and $L^{at}_{MSE,clip}$ denotes clip-level teacher-student consistency MSE loss for unlabeled data.

We evaluate the performance of different PaSST layers on SED and AT task with SED-adapter and AT-adapter, respectively. As illustrated in Figure 2, the deeper layer achieves better AT performance, but the *layer10* achieves the best SED performance. We infer that more global semantic information exists in the last layer but the local information, useful for SED task, is insufficient in the last layer since the PaSST is pretrained on AudioSet for AT task

Motivated by this, in the proposed task-aware fine-tuning, as shown in Figure 1(a), the PaSST model is fine-tuned with two adapters, where the SED-adapter is attached to *layer10* for exploiting local information and AT-adapter is attached to

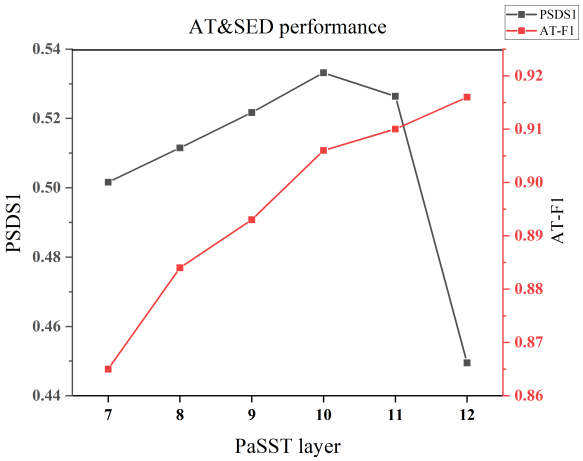


Figure 2: *AT and SED performance of different PaSST layers. AT-F1 is used to evaluate the AT performance, PSDS1 is used to evaluate the SED performance.*

layer12 for exploiting more global semantic information. Mean teacher is also used and the clip-level teacher-student teaching is changed: the clip-level output of teacher is an ensemble from two adapters, which is more accurate and effective for teaching. The loss function determined in the task-aware fine-tuning is as follows,

$$L_{task-aware} = L_{baseline} + \lambda_{AT} L_{AT} \tag{4}$$

where $L_{baseline}$ and L_{AT} are same as Eqn. (1) and Eqn. (3) respectively, λ_{AT} is set to 2.

2.2.2. SdMT: self-distillated mean teacher

In the proposed SdMT as shown in Figure 3, same as mean teacher, the main-student is trained with classification loss L_{class} for labeled data and the teacher-student consistency loss L_{cons} for unlabeled data. The teacher model is an EMA from student model. The detailed training loss of main-student is the same as Eqn. (1) or Eqn. (4), which depends on if using task-aware fine-tuning or not. However, different from main-student, the vice-student is trained with only the soft pseudo label to learn more knowledge from teacher. The training stage is termed as soft distillation. For comparison, we also evaluate the hard distillation.

Specifically, in the **soft distillation**, for all data including labeled and unlabeled, given logits of the vice student and teacher, the MSE loss is minimized,

$$\begin{aligned} L_{kd,soft} =& MSE(\delta(z_{s,frame}), \delta(z_{t,frame}/\tau)) \\ &+ \lambda_{clip} MSE(\delta(z_{s,clip}), \delta(z_{t,clip}/\tau)) \end{aligned} \tag{5}$$

where $z_{s,frame}$, $z_{t,frame}$, $z_{s,clip}$, $z_{t,clip}$ denotes student frame-level logits, teacher frame-level logits, student clip-level logits, teacher clip-level logits respectively, δ denotes sigmoid activation function, and the temperature τ is set to 1.

In the **hard distillation**, the teacher prediction is firstly converted to hard pseudo label using a threshold of 0.5, then the BCE loss between student and teacher is minimized,

$$\begin{aligned} L_{kd,hard} =& BCE(y_{s,frame}, \hat{y}_{t,frame}) \\ &+ \lambda_{clip} BCE(y_{s,clip}, \hat{y}_{t,clip}) \end{aligned} \tag{6}$$

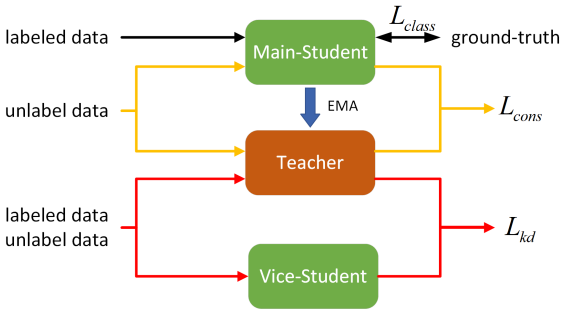


Figure 3: *Self-distillated mean teacher (SdMT). The main-student is trained with classification L_{class} and consistency loss L_{cons}, then the knowledge is transferred from teacher to vice-student with soft distillation (L_{kd}, training with soft label).*

where $y_{s,frame}$, $y_{s,clip}$, $\hat{y}_{t,frame}$, $\hat{y}_{t,clip}$ denote student frame-level prediction, student clip-level prediction, teacher frame-level pseudo label and teacher clip-level pseudo label respectively. The weight λ_{clip} is set to 0.5 in Eqn. (5) and Eqn. (6).

3. Experiments Setup

3.1. Dataset

For evaluation, we employ the DCASE2022 task4 development set (DESED) [26] which is composed of training and validation datasets. The training dataset contains: 1578 weakly-labeled clips, 3470 strongly-labeled clips, 10000 synthetic-strongly-labeled clips, and 14412 unlabeled in-domain clips. The validation dataset consists of 1168 strongly-labeled clips.

3.2. Feature Extraction

A 32kHz audio input waveform is first converted into 128-dimensional log Mel spectrogram features with a window size of 25ms and frame shift of 10ms. As a result, each 10-second sound clip is transformed into a 2D time-frequency representation with a size of (1000×128), then it shares same normalization as [22]. Frequency mask [27] is used for data augmentation with identical parameters to [22].

3.3. Experimental Settiongs

The model is trained over 20 epochs with the AdamW [28] optimizer, and a ratio of 1:1:2:2 for strong, synthetic-strong, weak and unlabeled data is used for each batch. Learning rates (lr) are set to 5e-6, 1e-4 for pre-trained PaSST and the task-aware adapters. During training, the lr is constant for the first 10 epochs, then reduced with exponential-down schedule to 5e-7,1e-5 for the last 10 epochs. When using SdMT, the main-student and teacher are firstly trained over 20 epochs, then the vice-student is trained over another 20 epochs with the aforementioned settings. For backend processing, median filter and weak prediction masking [29] are used. In median filter, the filter time length of event is presented as *category (time length)* as follows: Alarm_bell_ringing (0.32s), Blender (0.71s), Cat (0.32s), Dishes (0.32s), Dog (0.32s), Electric_shaver_toothbrush (4.29s), Frying (3.91s), Running_water (3.14s), Speech (0.32s), Vacuum_cleaner (1.09s). Event-Based F1-score (EB-F1) [30] and Polyphonic Sound detection Score scenario1 (PSDS1) [31] are used to evaluate fine-grained SED performance, and all event types share a threshold of 0.5 to obtain hard predictions

Table 1: *Comparison of model performance on DCASE 2022 DESED. Our implementation is based on the codebase from [14].*

Model	EB-F1,%	PSDS1
CRNN	50.50	0.4006
FDY-CRNN [14]	51.56	0.4256
SK-CRNN [12]	52.77	0.4004
Ensembled PANNs-RNN [18]	N/A	0.4450
Ensembled AST-RNN [18]	N/A	0.4590
BiCRNN (Winner) [17]	57.30	0.5050
AST-SED [21]	59.60	0.5140
PaSST-SED (ours)	**64.85**	**0.5548**

Table 2: *Performance of the proposed methods on DCASE 2022 DESED.* † *denotes our baseline.*

Model	Task-aware fine-tuning	SSL	EB-F1, %	PSDS1
PaSST-SED†	✗	MT	62.26	0.5345
PaSST-SED	✓	MT	64.37	0.5435
PaSST-SED	✗	SdMT	62.62	0.5364
PaSST-SED	✓	SdMT	**64.85**	**0.5548**

for calculating EB-F1.

4. Results and Discussion

In this section, we firstly compare the fine-tuned PaSST-SED with previous SOTA models, and show the respective contributions of task-aware fine-tuning and SdMT. Then we separately evaluate the task-aware fine-tuning and SdMT with different configurations.

4.1. Performance of the proposed methods

As shown in Table 1, the PaSST-SED model, with task-aware fine-tuning and SdMT, achieves EB-F1 of 64.85% and PSDS1 of 0.5548, significantly outperforming the previous SOTA models such as AST-SED and BiCRNN. This demonstrates that a general PaSST model, pretrained for AT task, can be well transferred for SED task with task-adapters and improved fine-tuning methods such as the proposed task-aware fine-tuning and SdMT.

As shown in Table 2, task-aware fine-tuning and SdMT both lead to improvement, but only using SdMT just achieves limited gain, it may be that the task-aware fine-tuning helps train a robust teacher with more accurate soft labels to guide the vice-student learning. Without task-aware fine-tuning and SdMT, the baseline model (*i.e.*, PaSST-SED with MT) also outperforms previous models with achieving the EB-F1 of 62.26% and PSDS1 of 0.5345 because of the well-pretrained transformer model and effective SED-adapter.

Table 3: *Evaluation of different clip-level teachers in the task-aware fine-tuning.*

clip-level teacher	EB-F1, %	PSDS1
Fusion	**64.37**	**0.5435**
SED-adatper	62.96	0.5357
AT-adatper	64.16	0.5428
Independent	63.16	0.5369

Table 4: *Evaluation of the soft and hard distillation in the SdMT.*

Model	Distillation type	Loss	EB-F1, %	PSDS1
main-student	-	BCE&MSE	64.37	0.5435
vice-student	soft	MSE	**64.85**	**0.5548**
vice-student	hard	BCE	64.30	0.5434

4.2. Evaluation of task-aware fine-tuning method

In this subsection, we evaluate different clip-level AT teachers in the task-aware fine-tuning method. As shown in Table 3, the "Fusion" AT teacher achieves the best performance with EB-F1 of 64.37% and PSDS1 of 0.5435, the ensemble of clip-level prediction of SED-adapter and AT-adapter is more accurate to guide the student learning by exploiting unlabeled data more efficiently. The "AT-adapter" is a suboptimal teacher, as the performance just decreases slightly compared with "Fusion" but outperforms "SED-adapter" by a large margin, which shows the AT-adapter helps exploit richer semantic information from deeper PaSST layer to guide the student learning in clip-level. It is interesting that the "Independent" teacher, where the two adapters in the teacher model guide the two adapters in the student model independently, also outperforms the baseline, which will be explored in the future.

4.3. Evaluation of SdMT method

We compare the soft and hard distillation where the vice-student in SdMT is trained with only soft or hard pseudo label from teacher. As shown in Table 4, soft distillation achieves the better PSDS1 and EB-F1 compared with hard distillation, and also outperforms the main-student which demonstrates that soft label is of rich information to guide the student learning. The hard distillation achieves no improvement compared with the main-students, it may be that the event is hard to be determined, and a fixed threshold of 0.5 is not optimal, which results in noisy labels. It is also shown that the PSDS1 is improved more appreciably compared with EB-F1 in the soft distillation. One possible explanation is that the EB-F1 is calculated with hard prediction determined by one fixed threshold of 0.5, but the PSDS1 is calculated with a set of thresholds, training with soft label may be more appropriate for PSDS1.

5. Conclusion

This paper presents an improved SED method based on pretrained PaSST model. Specifically, the SED-adapter and AT-adapter are first introduced, and based on the adapters, the task-aware fine-tuning is proposed to efficiently transfer pretrained PaSST to SED task where the SED-adapter is attached to shallower layer of PaSST to exploit local information and AT-adapter is attached to deeper layer to exploit semantic information. Besides, the SdMT is proposed to train a robust student with only soft labels to learn more knowledge from teacher. Experimental results on DCASE2022 task4 demonstrate the effectiveness of the task-aware fine-tuning and SdMT, the pretrained PaSST is well fine-tuned to outperform previous state-of-the-art models. In the future, we aim to propose more effective fine-tuning method to better exploit pretrained AT models and study how to effectively train SED systems with soft labels.

6. References

[1] A. Southern, F. Stevens, and D. Murphy, "Sounding out smart cities: Auralization and soundscape monitoring for environmental sound design," *J. Acoustical Soc. America*, vol. 141, no. 5, pp. 3880–3880, 2017.

[2] R. Radhakrishnan, A. Divakaran, and A. Smaragdis, "Audio analysis for surveillance applications," in *IEEE WASPAA 2005.*, 2005, pp. 158–161.

[3] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.

[4] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.

[5] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[6] V. Verma, K. Kawaguchi, A. Lamb, J. Kannala, A. Solin, Y. Bengio, and D. Lopez-Paz, "Interpolation consistency training for semi-supervised learning," *Neural Networks*, vol. 145, pp. 90–106, 2022.

[7] C.-Y. Koh, Y.-S. Chen, Y.-W. Liu, and M. R. Bai, "Sound event detection by consistency training and pseudo-labeling with feature-pyramid convolutional recurrent neural networks," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 376–380.

[8] S. Xiao, "Multi-dimensional frequency dynamic convolution with confident mean teacher for sound event detection," *arXiv preprint arXiv:2302.09256*, 2023.

[9] J. Yan, Y. Song, L.-R. Dai, and I. McLoughlin, "Task-aware mean teacher method for large scale weakly labeled semi-supervised sound event detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 326–330.

[10] L. Lin, X. Wang, H. Liu, and Y. Qian, "Guided learning for weakly-labeled semi-supervised sound event detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 626–630.

[11] E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE Trans. Audio, Speech, and Language Proc.*, vol. 25, no. 6, pp. 1291–1303, 2017.

[12] X. Zheng, Y. Song, I. McLoughlin, L. Liu, and L.-R. Dai, "An improved mean teacher based method for large scale weakly labeled semi-supervised sound event detection," in *IEEE ICASSP*, 2021, pp. 356–360.

[13] X. Zheng, Y. Song, L.-R. Dai, I. McLoughlin, and L. Liu, "An effective mutual mean teaching based domain adaptation method for sound event detection." in *Interspeech*, 2021, pp. 556–560.

[14] H. Nam, S.-H. Kim, B.-Y. Ko, and Y.-H. Park, "Frequency Dynamic Convolution: Frequency-Adaptive Pattern Recognition for Sound Event Detection," in *Proc. Interspeech 2022*, 2022, pp. 2763–2767.

[15] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Convolution augmented transformer for semi-supervised sound event detection," in *Proc. Workshop Detection Classification Acoust. Scenes Events (DCASE)*, 2020, pp. 100–104.

[16] J. F. Gemmeke, D. P. 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 *IEEE ICASSP*, 2017, pp. 776–780.

[17] J. Ebbers and R. Haeb-Umbach, "Pre-training and self-training for sound event detection in domestic environments," DCASE, Tech. Rep., June 2022.

[18] S. Xiao, "Pretrained models in sound event detection for dcase 2022 challenge task4," DCASE, Tech. Rep., June 2022.

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

[20] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech 2021*, 2021, pp. 571–575.

[21] K. Li, Y. Song, L.-R. Dai, I. McLoughlin, X. Fang, and L. Liu, "Ast-sed: An effective sound event detection method based on audio spectrogram transformer," *arXiv preprint arXiv:2303.03689*, 2023.

[22] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient Training of Audio Transformers with Patchout," in *Proc. Interspeech 2022*, 2022, pp. 2753–2757.

[23] I. Martín-Morató, M. Harju, P. Ahokas, and A. Mesaros, "Training sound event detection with soft labels from crowdsourced annotations," *arXiv preprint arXiv:2302.14572*, 2023.

[24] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[25] Y. Wang, J. Li, and F. Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in *IEEE ICASSP*, 2019, pp. 31–35.

[26] N. Turpault, R. Serizel, J. Salamon, and A. P. Shah, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," 2019.

[27] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[28] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[29] H. Nam, B.-Y. Ko, G.-T. Lee, S.-H. Kim, W.-H. Jung, S.-M. Choi, and Y.-H. Park, "Heavily augmented sound event detection utilizing weak predictions," *arXiv preprint arXiv:2107.03649*, 2021.

[30] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.

[31] Ç. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulović, "A framework for the robust evaluation of sound event detection," in *IEEE ICASSP*, 2020, pp. 61–65.