A MIXED SUPERVISED LEARNING FRAMEWORK FOR TARGET SOUND DETECTION

Dongchao Yang[1], *Helin Wang*[1], *Wenwu Wang*[2], *Yuexian Zou*[1]*

[1]ADSPLAB, School of ECE, Peking University, Shenzhen, China
[2]Center for Vision, Speech and Signal Processing, University of Surrey, UK

ABSTRACT

Target sound detection (TSD) aims to detect the target sound from mixture audio given the reference information. Previous works have shown that TSD models can be trained on fully-annotated (frame-level label) or weakly-annotated (clip-level label) data. However, there are some clear evidences show that the performance of the model trained on weakly-annotated data is worse than that trained on fully-annotated data. To fill this gap, we provide a mixed supervision perspective, in which learning novel categories (target domain) using weak annotations with the help of full annotations of existing base categories (source domain). To realize this, a mixed supervised learning framework is proposed, which contains two mutually-helping student models (*f_student* and *w_student*) that learn from fully-annotated and weakly-annotated data, respectively. The motivation is that *f_student* learned from fully-annotated data has a better ability to capture detailed information than *w_student*. Thus, we first let *f_student* guide *w_student* to learn the ability to capture details, so *w_student* can perform better in the target domain. Then we let *w_student* guide *f_student* to fine-tune on the target domain. The process can be repeated several times so that the two students perform very well in the target domain. To evaluate our method, we built three TSD datasets based on UrbanSound and Audioset. Experimental results show that our methods offer about 8% improvement in event-based F-score as compared with a recent baseline.

Index Terms— Target sound detection, audioset, weakly supervised, mixed supervised learning

1. INTRODUCTION

In target sound detection (TSD) [1], one aims to recognize and localize the target sound source within a mixture audio given a reference audio, e.g. detecting the dog bark sound within a street. TSD can be applied to numerous potential fields [2–4], such as species migration monitoring and large-scale multimedia indexing. Sound event detection (SED) [5] is a similar task with TSD, and a lot of works have been done for SED [6–11]. However, SED aims to classify and localize all pre-defined events (*e.g.*, dog barking, man speaking) within an audio clip, which significantly limits the flexibility to detect unseen classes. Different with SED, TSD only focuses on detecting the event that we interest. TSD does not require pre-defined set of categories, therefore, it can be easily extended to sound detection from an open set. In a recent work, a target sound detection network (TSDNet) [1] is presented, where a conditional network is used to generate sound-discriminative embedding which is then used as the reference information to guide a detection network for the detection of the target sound from the mixture audio. TSDNet provides a good detection performance when training data is fully-annotated, e.g. the onset and offset time of the target sound are provided in the annotations. However, collecting large-scale fully-annotated data is time-consuming and labor-intensive. Weakly supervised TSD is an effective method to reduce the reliance on fully-annotated data, but the performance tends to degrade significantly [1].

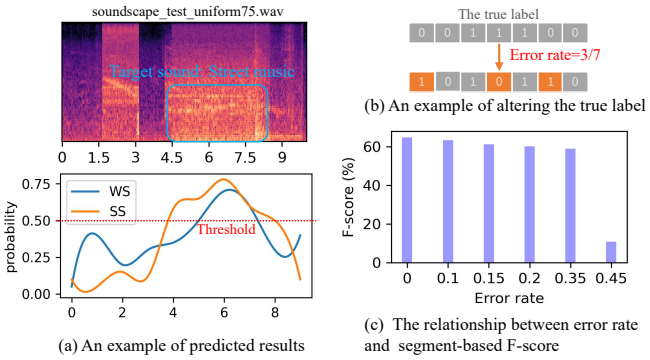

Figure 1: (a) shows the predicted results generated by weakly (WS) and strongly supervised (SS) learning. (b) shows an example of altering the true label to false label. (c) shows the influence of the error rate of the frame-level label when we train TSDNet model [1] on the URBAN-TSD dataset.

In this paper, we consider TSD with mixed supervision, which learns novel sound categories (target domain) using weak annotations with the help of full annotations of the existing base sound categories (source domain). Under this setting, we can use a small-scale fully-annotated dataset (e.g. URBAN-SED [12]) to complement a large-scale weakly-annotated dataset (e.g. Audioset [13]). To achieve this, we propose a novel mixed supervised learning framework, which includes two mutually-helping student models (*f_student* and *w_student*), which are trained by fully- and weakly-annotated data, respectively. The proposed method involves three novel aspects. Firstly, the *f_student* learned from fully-annotated data has better ability in capturing detailed information than the *w_student*. As Figure 1 (a) shows, the model trained on weakly-annotated data fails to locate the event boundary: it only focuses on the most distinct part and misses the boundary information. Thus, we propose a frame-level knowledge distillation (KD) strategy to transfer the knowledge from *f_student* to *w_student*, which makes *w_student* able to capture more details, *e.g.* boundary information.

* Corresponding Author: zouyx@pku.edu.cn

This paper was partially supported by Shenzhen Science & Technology Research Program (No: GXWD20201231165807007-20200814115301001; No: JSGG20191129105421211) and NSFC (No: 62176008).

However, it is hard to transfer all of the knowledge from *f_student* to *w_student*. Thus, we propose to directly apply *f_student* to the target domain with the guidance of *w_student*. Specifically, *w_student* is used to produce frame-level pseudo labels for *f_student*. This strategy inspired by an interesting phenomenon that even if there are some errors in the frame-level labels, the detection performance remains stable or decreases only slightly (when the error rate lower than 0.35), as Figure 1 (c) shows. The process of mutually helping can be repeated several times so that the two students perform very well in the target domain. Lastly, we found that the mismatch between source and target data distribution tends to affect significantly the performance of transfer learning, *e.g.* URBAN-SED [12] and Audioset [14]. Thus, we propose an adversarial training strategy to solve the domain mismatch problem. To evaluate our method, we built two small-scale fully-annotated datasets and a large-scale weakly-annotated dataset based on URBAN-SED and Audioset. Experimental results show that two small-scale fully annotated datasets could significantly improve the performance on large-scale weakly-annotated dataset.

2. RELATED WORK

Many methods [15–22] have been proposed to utilize both fully- and weakly-annotated data to train the SED model. However, previous methods assume that fully- and weakly- annotated data belong to the same set of pre-defined categories. Our method aims to use the existing base categories with full annotations to facilitate the recognition of novel categories with weak labels. This setting is more realistic for the reason that a small-scale fully-annotated dataset with few categories is, in practice, easier to create, as compared with a large-scale fully-annotated dataset of many categories.

3. TSD DATASETS

We built two TSD datasets based on Audioset [14], which includes 94126 training clips and 16118 test clips, from 456 different classes. One is the large-scaled TSD dataset, named as L-TSD dataset, by choosing 192 different classes from Audioset. The other is the small-scaled fully-annotated TSD dataset (S-TSD), by choosing 51 different classes from Audioset. L-TSD includes two types of annotated samples, that is, fully-annotated (*i.e.* L-TSD-strong) and weakly-annotated (*i.e.* L-TSD-weak) samples. There is no common class between L-TSD and S-TSD datasets. The process of building the datasets is similar to the one described in [1]. The mixture audio signals come from Audioset. If there are N sound events in the mixture audio, we can generate N positive samples. We also generate $N/2$ negative samples, which do not contain the target sound. To prepare reference audio, we select the audio clips for each category directly from the Audioset training set i.e. those that do not contain interference from other events. In total, L-TSD includes 490,336 training and 83,334 test clips, while S-TSD contains 26,247 training and 5113 test clips. In addition, we choose the URBAN-TSD dataset [1] as another small-scale fully-annotated dataset.

4. PROPOSED METHOD

Figure 2 shows our proposed mixed supervised learning framework. The core idea of the framework is that the two students can teach each other iteratively. One of the student models is trained on fully-annotated data, we name it as *f_student*. The other model is trained on weakly-annotated data, and we name it as *w_student*. The two students have the same structure, while the only difference is that *w_student* has a linear softmax pooling layer [10].

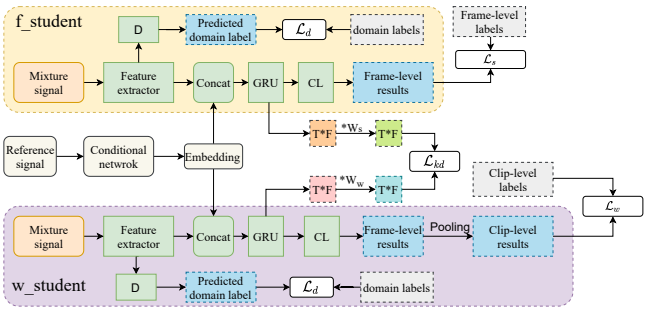

Figure 2: The architecture of the mixed supervised learning framework. CL denotes classification layers, which includes two fully-connected layers and one softmax function. D denotes the discriminator, which consists of three convolutional layers and one fully-connected layer.

4.1. Network Structure

Conditional network. The conditional network aims to extract a sound-discriminative embedding vector from the reference audio. Similar to the previous work [1], we adopt a VGG-like convolutional neural network (CNN) model [23] for the conditional network.

Detection network. Similar to the previous work [1], the network is composed of 5 convolutional layers, 1 Bi-GRU layer, and 2 fully-connected layers. Given the mel-spectrogram of the mixture audio $\boldsymbol{x} \in \mathcal{R}^{T \times F}$, where T and F denote the number of frames and the dimension of frame. The detection network aims to predict frame-level probabilities

$$\hat{p}_i = \mathbb{P}(Y = k | X = x_i, \boldsymbol{e}; \phi) \tag{1}$$

where ϕ denotes the trainable parameters of the detection network, $\boldsymbol{e}$ denotes the embedding obtained from the conditional network and x_i denotes the i-th frame of the mixture audio $\boldsymbol{x}$. For *f_student*, given the ground-truth label $p_i \in \{0, 1\}$ for each frame, which can be optimized by minimizing the binary cross entropy (BCE) loss:

$$\mathcal{L}_s = \sum_{i=1}^{t} (-p_i \log \hat{p}_i - (1 - p_i) \log(1 - \hat{p}_i)) \tag{2}$$

where t indicates the number of frames. The difference between *f_student* and *w_student* is that the latter needs a pooling layer to get the clip-level prediction. Thus, a LinSoft pooling layer [10] is added after the last layer of the *f_student*. *w_student* aims to predict a clip-level probability $\hat{P} = f_{LSP}(\hat{p}_1, \hat{p}_2, ..., \hat{p}_t)$ where $f_{LSP}(\cdot)$ denotes the LinSoft pooling function. Given the clip-level ground-truth label $P \in \{0, 1\}$, the BCE loss is applied as the loss function:

$$\mathcal{L}_w = -P \log \hat{P} - (1 - P) \log(1 - \hat{P}) \tag{3}$$

4.2. Two-student Learning

In this part, we introduce the details of how to enable *w_student* and *f_student* to help each other.

Frame-level knowledge distillation. According to formula (3), we

can see that *w_student* makes a decision on the whole audio clip. Compared to *f_student*, *w_student* is limited in capturing the detailed information of the sound events. To address this issue, we propose to first train *f_student* on a small-scale fully-annotated dataset (*i.e.* source domain), and then transfer its knowledge to *w_student*, so that *w_student* can get better performance on weakly-annotated dataset (*i.e.* target domain). Specifically, we first train the *f_student* model on the source dataset with strong labels. After that, we train the *w_student* model on the source data with weak labels. We then treat the trained *f_student* model as a teacher, to generate a frame-level feature representation. As a result, *w_student* may capture more detailed information, due to the frame-level class-agnostic knowledge distillation. More specifically, we can train *w_student* with the following objective function,

$$\mathcal{L}_{w\_kd} = \mathcal{L}_w + \mathcal{L}_{kd} \tag{4}$$

$$\mathcal{L}_{kd} = ||\boldsymbol{F}_s \cdot \boldsymbol{W}_s - \boldsymbol{F}_w \cdot \boldsymbol{W}_w||_2 \tag{5}$$

where $\boldsymbol{F}_s$ and $\boldsymbol{F}_w$ denote the feature map of the GRU layer of the two models, and $\boldsymbol{W}_s$ and $\boldsymbol{W}_w$ denote the transformation matrix.

Pseudo Supervised Training. The idea of pseudo supervised training strategy is motivated by an interesting observation, i.e. even if there are some errors in the frame-level labels, the detection performance remains stable or decreases only slightly. This means that we could use the noisy frame-level labels as our training target. In this paper, we propose to use *w_student* to produce noisy frame-level labels (i.e. pseudo labels), and then use the pseudo labels to re-train *f_student*, as follows

$$\hat{p}_i^w = \mathbb{P}(Y = k|X = x_i, \boldsymbol{e}; \phi_w), \hat{p}_i^s = \mathbb{P}(Y = k|X = x_i, \boldsymbol{e}; \phi_s) \tag{6}$$

$$\mathcal{L}_{re\_s} = \sum_{i=1}^{t}(-\hat{p}_i^w \log \hat{p}_i^s - (1 - \hat{p}_i^w) \log(1 - \hat{p}_i^s)) \tag{7}$$

Adversarial Training. In our experiments, we found that the mismatch between source and target data distribution could significantly degrade the performance of the mixed supervised learning framework. For example, if we choose the URBAN-TSD as the source dataset, and the L-TSD-weak as the target dataset, the performance will decrease substantially. This is because there is domain mismatch between URBAN-TSD and L-TSD-weak datasets [24, 25], i.e. *f_student* and *w_student* are first trained on the URBAN-TSD dataset but tested on L-TSD-weak dataset. To solve the domain mismatch problem, we propose a domain adversarial training strategy that aims to learn a common subspace shared by both the source and target domains, which enables all domains to have the same data distribution in the feature space. Specifically, inspired by GAN [26] and DANN [27], we make use of the adversarial relationship between modules Feature extractor (F) and Discriminator (D) to learn domain-invariant features in the feature space. To achieve this, we add an adversarial loss when we train *f_student* and *w_student*, as follows

$$\mathcal{L}_d = ||D(\boldsymbol{z}) - \boldsymbol{d}||_2^2 \tag{8}$$

$$\mathcal{L}_{d\_tsd} = \mathcal{L}_{tsd} - \lambda_d * \mathcal{L}_d \tag{9}$$

where $\boldsymbol{z}$ denotes the intermediate feature produced by F, and $\boldsymbol{d}$ denotes the domain label. The domain label is defined as $\boldsymbol{d} = [1, 0]^T$ (which stands for the source domain) or $\boldsymbol{d} = [0, 1]^T$ (target domain). $\mathcal{L}_d$ denotes the domain classification loss of the discriminator, $\mathcal{L}_{tsd}$ denotes the detection loss. $\mathcal{L}_{d\_tsd}$ denotes the training objective function, which minimizes the detection loss and meanwhile maximizes the domain classification loss. The parameter λ_d controls the trade-off between $\mathcal{L}_{tsd}$ and $\mathcal{L}_d$. In our experiments, λ_d is set to 0.2 empirically based on the validation set.

Iterative Training Strategy. According to the previous description, we can use frame-level KD to transfer the knowledge from *f_student* to *w_student*, and use the pseudo supervised training to transfer the knowledge from *w_student* to *f_student*. Intuitively, the process can be repeated several times. Thus, an iterative training strategy is proposed. The whole algorithm is summarized in Algorithm 1.

Algorithm 1 Two-Student Learning

Input:
The source dataset D_s and the target dataset D_t
Output: *f_student* and *w_student* model
1: Training *f_student* on D_s using formula (2) and (9)
2: Training *w_student* on D_s using formula (4) and (9)
3: Retraining *w_student* on D_t using formula (3) and (9)
4: Retraining *f_student* on D_t using formula (7) and (9)
5: While True:
Retraining *w_student* on D_t, using formula (4)
Retraining *f_student* on D_t using formula (7)
If no improvement: break;
6: **return** *f_student* and *w_student*;

5. EXPERIMENTS

5.1. Datasets

In this section, we introduce the source dataset and target dataset used in our experiments. The source (*resp.*, target) dataset is fully-annotated (*resp.*, weakly-annotated).

Source Dataset. We first use the S-TSD dataset as the source dataset, which is a small-scale fully-annotated dataset based on Audioset [14]. In addition, we choose 10-category URBAN-TSD [1] as another source dataset, which includes two similar categories as in the L-TSD dataset: *dog_bark* and *gun_shot*.

Target Dataset. We take the L-TSD-weak dataset as the target dataset. The details were given in Section 3.

5.2. Experimental Setups

Conditional Network. We use the pre-trained PANNs [23] model to initialize the conditional network, and then fix it in the training process.

Detection network. All the raw audios are down-sampled to 22.05kHz and then Short Time Fourier Transform (STFT) with a window size of 2048 samples are applied, followed by a Mel-scaled filter bank on perceptually weighted spectrogram. This results in 64 Mel frequency bins and around 50 frames per second. When training *f_student* and *w_student* for the first time, the Adam optimizer [28] is used for 100 epochs, with an initial learning rate of 1×10^{-3}. When they are re-trained, the learning rate is set as 1×10^{-4}.

Metrics. We use the segment-based F-score and event-based F-score [29] as the evaluation metrics, which are the most commonly used metrics for detection task.

Table 1: F-score comparison with different supervision strategy on L-TSD test set. SS, WS and MS represent strong, weak and mixed supervision, respectively. F-scores are macro-averaged.

Method	Source dataset	Segment-F score	Event-F score
SS [1]	-	58.57	50.4
WS [1]	-	49.39	39.07
MS (ours)	S-TSD	50.95	47.19
	URBAN-TSD	51.31	47.56

Table 2: Ablation studies on different strategies on L-TSD test set.

Model	KD	PS	AD	Segment-F score	Event-F score
w_student	✗	✗	✗	49.39	39.07
	✗	✗	✓	49.34	39.11
	✓	✗	✗	37.55	39.74
	✓	✗	✓	50.34	41.35
f_student	✗	✓	✓	48.09	43.47
	✓	✓	✗	37.67	45.42
	✓	✓	✓	**51.31**	**47.56**

5.3. Experiments on S-TSD and L-TSD Datasets

In this section, we use S-TSD as the source dataset and L-TSD-weak as the target dataset. We compare our method with strongly supervised (SS) and weakly supervised (WS) methods. For SS-TSD, we directly train *f_student* on the L-TSD-strong dataset with the strong labels. For WS-TSD, we directly train *w_student* on the L-TSD-weak dataset with the weak labels. Note that for mixed supervised (MS) method, we only report the results obtained by *f_student* for the reason that *f_student* performs better than *w_student*. The experimental results are given in Table 1. From this table, we can see that our proposed MS method performs significantly better than the WS method, and performs similarly to the SS method.

5.4. Experiments on URBAN-TSD and L-TSD Datasets

We also conduct experiments by using URBAN-TSD as the source dataset and L-TSD-weak as the target dataset. Table 1 shows the experimental results, and we can see that using URBAN-TSD as the source dataset significantly improves the performance compared with the WS method. Furthermore, by comparing rows 3 and 4, we can find that using URBAN-TSD as the source dataset obtains better performance than using S-TSD as the source dataset. One of the reasons is the categories of URBAN-TSD and L-TSD-weak datasets have overlaps.

5.5. Ablation Studies

By taking URBAN-TSD as the source dataset, we conduct ablation studies to investigate the effectiveness of knowledge distillation (KD), pseudo supervised training (PS) and adversarial (AD) training, with the results shown in Table 2. For the *w_student* model: (1) The first row shows the results for directly training *w_student* on L-TSD-weak (without using any strategy). (2) The second row shows the results of directly training *w_student* on L-TSD-weak while using the AD strategy. We can see that only using the AD strategy does not give improvements, because it only aims to align the data distribution. (3) By comparing rows 2 and 4, we can see that the KD strategy can improve the *w_student*'s ability in capturing detailed information. (4) By comparing rows 3 and 4, we can see the effectiveness of the AD strategy when there is mismatch between the source and target datasets.

For the *f_student* model: (1) By comparing rows 5 and 7, we can see that the KD strategy can improve the performance of *f_student* for the reason that it can improve the performance of *w_student*, and the better *w_student* leads to better *f_student*. (2) By comparing rows 6 and 7, we can find the improvements given by the AD strategy. Lastly, we can see that *f_student* has better performance than *w_student*. A reason could be that the ability of *f_student* in capturing the detailed information is only partially transferred to *w_student* through the KD strategy.

Table 3: Ablation study on the effect of the number of iterations on iterative training strategy.

Iterations	Model	Segment-F score	Event-F score
1	w_student	50.39	40.88
	f_student	50.21	46.99
2	w_student	50.14	41.60
	f_student	51.10	46.96
3	w_student	51.33	44.70
	f_student	50.95	47.19

Influence of the number of iterations. By taking S-TSD as the source dataset, we conduct ablation studies to investigate the influence of the number of iterations on the iterative training strategy, and the results are shown in Table 3. In this work, considering the cost of training time, we only tested three stages. For *f_student*, the segment- and event-based F-scores are increased from 50.21% and 46.99% to 50.95% and 47.19%, respectively. Furthermore, the event-based F-score of *w_student* is increased from 40.88% to 44.7%. It means that the two students can help each other.

Why does pseudo supervised training work? To explain why the pseudo supervised training strategy is effective, we calculate the error rate between the true label and pseudo label (produced by *w_student*) on the L-TSD-strong training set. We found that the average error rate is 27.03%, over the 490,336 audio clips. This is consistent with our empirical results as shown earlier in Figure 1: if the error rate is smaller than 0.35, the performance is similar to that of the case where the true label is taken as the target.

6. CONCLUSIONS

In this paper, we have presented a novel mixed supervised learning framework, which effectively improves the performance of novel categories with the help of a small-scale fully-annotated base categories dataset. In the future, we will apply our method to other tasks, such as SED and object detection. The source code and dataset of this work have been released[1].

7. REFERENCES

[1] D. Yang, H. Wang, Y. Zou, and C. Weng, "Detect what you want: Target sound detection," *arXiv preprint arXiv:2112.10153*, 2021.

[2] J. P. Bello, C. Silva, O. Nov, R. DuBois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, "Sonyc: A system for the

[1] https://github.com/yangdongchao/weakly-target-sound-detection

monitoring, analysis and mitigation of urban noise pollution," *arXiv preprint arXiv:1805.00889*, 2018.

[3] D. Stowell and D. Clayton, "Acoustic event detection for multiple overlapping similar sources," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2015, pp. 1–5.

[4] S. Hershey, S. Chaudhuri, D. Ellis, J. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, *et al.*, "CNN architectures for large-scale audio classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 131–135.

[5] Q. Kong, Y. Xu, I. Sobieraj, W. Wang, and M. D. Plumbley, "Sound event detection and time-frequency segmentation from weakly labelled data," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 777–787, 2019.

[6] H. Dinkel, M. Wu, and K. Yu, "Towards duration robust weakly supervised sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 887–900, 2021.

[7] L. Lin, X. Wang, H. Liu, and Y. Qian, "Specialized decision surface and disentangled feature for weakly-supervised polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1466–1478, 2020.

[8] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "Sound event detection of weakly labelled data with cnn-transformer and automatic threshold optimization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2450–2460, 2020.

[9] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial," *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 67–83, 2021.

[10] 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 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 31–35.

[11] I. Martín-Morató, A. Mesaros, T. Heittola, T. Virtanen, M. Cobos, and F. Ferri, "Sound event envelope estimation in polyphonic mixtures," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 935–939.

[12] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 344–348.

[13] J. Gemmeke, D. 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 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[14] S. Hershey, D. P. Ellis, E. Fonseca, A. Jansen, C. Liu, R. C. Moore, and M. Plakal, "The benefit of temporally-strong labels in audio event classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 366–370.

[15] A. Kumar and B. Raj, "Audio event and scene recognition: A unified approach using strongly and weakly labeled data," in *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 3475–3482.

[16] Y. Liang, Y. Long, Y. Li, and J. Liang, "Joint weakly supervised AT and AED using deep feature distillation and adaptive focal loss," *arXiv preprint arXiv:2103.12388*, 2021.

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

[18] Z. Shi, L. Liu, H. Lin, R. Liu, and A. Shi, "Hodgepodge: Sound event detection based on ensemble of semi-supervised learning methods," *arXiv preprint arXiv:1907.07398*, 2019.

[19] T. K. Chan, C. S. Chin, and Y. Li, "Non-negative matrix factorization-convolutional neural network (nmf-cnn) for sound event detection," *arXiv preprint arXiv:2001.07874*, 2020.

[20] 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 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 356–360.

[21] L. Cances and T. Pellegrini, "Comparison of deep co-training and mean-teacher approaches for semi-supervised audio tagging," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 361–365.

[22] Y. Guan, J. Xue, G. Zheng, and J. Han, "Sparse self-attention for semi-supervised sound event detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 821–825.

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

[24] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[25] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," *Advances in Neural Information Processing Systems*, vol. 19, 2006.

[26] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv preprint arXiv:1406.2661*, 2014.

[27] R. Wang, M. Wang, X.-L. Zhang, and S. Rahardja, "Domain adaptation neural network for acoustic scene classification in mismatched conditions," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 1501–1505.

[28] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

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