INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

An Effective Mutual Mean Teaching based Domain Adaptation Method for Sound Event Detection

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

[1]National Engineering Laboratory for 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

zx980216@mail.ustc.edu.cn, {songy, ivm, lrdai}@ustc.edu.cn, linliu@iflytek.com

Abstract

In this paper, we present a novel mutual mean teaching based domain adaptation(MMT-DA) method for sound event detection(SED) task, which can effectively exploit synthetic data to improve the SED performance. Existing methods simply treat the synthetic data as strongly-labeled data in semi-supervised learning (SSL) framework. Benefiting from the strong labels of synthetic data, superior SED performance can be achieved. However, a distribution mismatch between synthetic and real data raises an evident challenge for domain adaptation (DA). In MMT-DA, convolutional recurrent neural networks(CRNN) learned from different datasets (*i.e. total data*:real+synthetic, and *real data*) are exploited for DA. Specifically, mean teacher method using CRNN is employed for utilizing the unlabeled real data. To compensate the domain diversity, an additional domain classifier with gradient reverse layer(GRL) is used for training a mean teacher for *total data*. The student CRNNs are mutually taught using the soft predictions of unlabeled data obtained from different teachers. Furthermore, a strip pooling based attention module is exploited to model the inter-dependencies between channels and time-frequency dimensions to exploit the structure information. Experimental results on Task4 of DCASE2020 demonstrate the ability of the proposed method, achieving 52.0% F1-score on the validation dataset, which outperforms the winning system's 50.6%.

Index Terms: sound event detection, domain adaptation, mutual mean teaching, semi-supervised learning

1. Introduction

Sound event detection (SED) is the task of detecting both the onset and offset of a sound event. It has wide applications for real-world systems including smart home devices [1], and automatic surveillance [2]. 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] based approaches have achieved state-of-the-art performance in DCASE2018 [6]. MT methods consist of a teacher which is updated as an ensemble of the students to generate targets for SSL, and a consistency cost employed as a regularization term, as shown in Figure 1(a).

More recently, synthetic dataset with strong labels has been proposed for training in SED tasks [7]. Compared to the real audio, which is generally weakly labeled or unlabeled, synthetic data can be strongly labeled and have accurate time-stamps. These has been shown to be important factors for achieving better SED performance [8, 9, 10]. However, these methods ignore the domain gap between synthetic and real audio data. Yang *et al.* [11] thus proposed a domain adaptation model to mitigate the distribution mismatch between synthetic and real data, as shown in Figure 1(b). It is an elegant approach which combines the SSL and DA methods using a network, as shown in Figure 1(c). However, it was shown in [12] that network parameters learned from DA and SSL are quite different. In fact, networks learned via DA are typically dominated by the synthetic data, whereas those learned using SSL are biased towards the real data. Thus, the two networks can provide complementary views of unlabeled data. We believe this can be exploited by co-training [13, 14].

In this paper, we present an effective mutual mean teaching based domain adaptation (MMT-DA) method to explore this, as shown in Figure 1(d). Specifically, we decompose the cross-domain SED task into two sub-problems, *i.e.* DA and SSL. Two CRNNs are used to solve DA and SSL tasks respectively. Mutual mean teaching is performed to guide the student model to utilize the unlabeled dataset. Furthermore, a strip pooling based attention module is designed to explicitly model inter-dependencies between channels and time-frequency domain. Compared with 'squeeze-and-excitation' [15], where time-frequency information is discarded by global average pooling, time or frequency information is reserved in the strip pooling based attention module. This provides an attention mechanism able to adaptively re-calibrate the output feature map. We evaluate the effectiveness of the proposed methods through extensive experiments and ablation studies using DCASE2020 challenge Task 4 benchmarks. The F1-score of 52% achieved by the proposed system on the validation dataset outperforms the previous winning system's 50.6%, clearly highlighting the effectiveness of the proposed methods.

2. Previous works on SED

In this section, we will introduce previous works on sound event detection, mainly focusing on the mean-teacher (MT) based framework in Figure 1(a), and the adversarial training based domain adaptation (DA) approach in Figure 1(b).

2.1. MT based framework for SED

As shown in Figure 1(a), the MT framework, adopts two CRNNs with the same architecture (namely, the teacher and the student network). The CNN part contains 5 convolutional blocks, whose architecture is the same as baseline convblock in [16], followed by 2 Bi-GRU layers and a localization module with linear softmax pooling [17]. In the MT approach, the student model is trained by back propagation of supervised loss from labeled data and teacher-student consistency loss, while

http://dx.doi.org/10.21437/Interspeech.2021-281

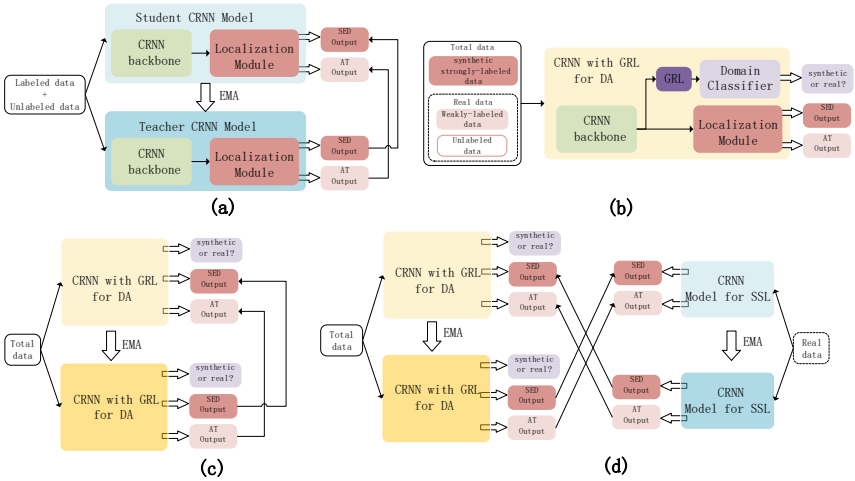

Figure 1: *(a) Mean teacher(MT) based framework for semi-supervised SED, where AT is shorthand for audio tagging. (b) adversarial training domain adaptation (DA) method with gradient reverse layer(GRL) [11], (c) MT-DA system combining (b) with (a), (d) our proposed mutual mean-teaching based domain adaptation (MMT-DA) method.*

the teacher model is updated, much more slowly, by the exponential moving average(EMA) of the student parameters. The consistency loss $L_{consist}$ is defined as the expected distance between the output of the teacher model $f_{\theta'}$ and the student model f_{θ}:

$$\begin{aligned} L_{consist} = \; & J_{mse}(f_{\theta}^{sed}(\mathbf{x}_u), f_{\theta'}^{sed}(\mathbf{x}_u)) \\ & + J_{mse}(f_{\theta}^{at}(\mathbf{x}_u), f_{\theta'}^{at}(\mathbf{x}_u)) \end{aligned} \tag{1}$$

where $\mathbf{x}_u$ denotes unlabeled data, $f^{sed}(\cdot)$ and $f^{at}(\cdot)$ are SED and AT outputs respectively, and J_{mse} is shorthand for mean squared error (MSE).

2.2. Adversarial Training based DA method for SED

As presented in Figure 1(b), an adversarial training based DA method [11] is proposed to mitigate the distribution mismatch between synthetic and real data to improve SED performance when using synthetic data. Specifically, a domain-classifier is trained to discriminate when an input audio comes from the synthetic dataset or the real dataset, whereas the feature extractor is trained to deceive the domain classifier to match feature distributions. Let θ_F, θ_D denate the parameters for the CRNN backbone and domain discriminator respectively, then the overall adversarial learning objective functions are:

$$\hat{\theta}_F = \arg\min_{\theta_F} L_S + L_W - w_d \cdot L_D \tag{2}$$

$$\hat{\theta}_D = \arg\min_{\theta_D} L_D \tag{3}$$

where L_S, L_W, L_D denate strong binary cross entropy (BCE) loss, weak BCE loss and domain BCE loss respectively and w_d weights the domain loss component. In the implementation of adversarial learning, a gradient reversal layer (GRL) [18] is applied to flip the gradient between CRNN backbone and domain classifier with respect to L_D. During the backpropagation though, the GRL takes the gradient from the subsequent level, changes its sign and passes it to the preceding layer.

Since Yang *et al.* [11] simply treated unlabeled data as real data, the SED or AT pseudo label for the unlabeled data was not explored. As shown in Figure 1(c), a straightforward implementation of GRL to the MT framework leads to the MT-DA method, which can achieve an SED performance improvement, but our proposed MMT-DA in Figure 1(d) is even more effective than MT-DA.

3. Proposed Methods

As aforementioned, the proposed MMT-DA method uses 2 different CRNNs to exploit DA subtask from synthetic data to real data, and semi-supervised SED subtask for real data respectively. Mutual mean teaching (MMT) is then employed to guide student to exploit the unlabeled data. Meanwhile, a strip pooling based attention (SP_A) module is proposed to further improve SED performance of the MMT-DA system. Details of our proposed methods will be described, respectively, in the following subsections.

3.1. MMT-DA based method for sound event detection

As shown in Figure 1(d), the CRNN models with and without GRL (left and right side respectively) are aiming at a DA subtask, and a semi-supervised SED subtask on real domain data respectively. To achieve this, the input data for these two models will differ: both synthetic and real data are fed into the CRNNs with GRL, whereas only real data is fed into the CRNNs without GRL.

Assume $f_{\theta_{DA}}$, $f_{\theta'_{DA}}$, $f_{\theta_{SSL}}$, $f_{\theta'_{SSL}}$ represent the student teacher CRNN with GRL for the DA subtask, and the student teacher CRNN without GRL for the SSL subtask for real data, respectively. Given an unlabeled sample, a soft pseudo label is generated by the teacher models, and mutual mean teaching is employed by the MSE loss as follows:

$$\begin{aligned} L_{MMT} = \; & J_{mse}(f_{\theta_{SSL}}^{sed}(\mathbf{x}_u), f_{\theta'_{DA}}^{sed}(\mathbf{x}_u)) \\ & + J_{mse}(f_{\theta_{SSL}}^{at}(\mathbf{x}_u), f_{\theta'_{DA}}^{at}(\mathbf{x}_u)) \end{aligned} \tag{4}$$

$$\begin{aligned} L_{MMT'} = \; & J_{mse}(f_{\theta_{DA}}^{sed}(\mathbf{x}_u), f_{\theta'_{SSL}}^{sed}(\mathbf{x}_u)) \\ & + J_{mse}(f_{\theta_{DA}}^{at}(\mathbf{x}_u), f_{\theta'_{SSL}}^{at}(\mathbf{x}_u)) \end{aligned} \tag{5}$$

Then the total loss for SSL and DA models are combined:

$$L_{SSL\_total} = L_W + w(t) \cdot L_{MMT} \tag{6}$$

$$L_{DA\_total} = L_S + L_W + w_d \cdot L_D + w(t) \cdot L_{MMT'} \quad (7)$$

where L_S, L_W, L_D denote strong BCE loss for synthetic data, weak BCE loss for weakly-labeled data and domain BCE loss, and $w(t)$ is the ramp-up weight for current epoch t.

3.2. Strip Pooling based attention(SP_A) module

The strip pooling based attention module is applied after each convolutional block to provide an effective attention mechanism to adaptively re-calibrate the feature map. Before describing the formulation of our strip pooling based attention module, we first briefly introduce the basic strip pooling method.

3.2.1. Strip Pooling

Given an intermediate feature map $\mathbf{U} \in \mathbb{R}^{C\times T\times F}$, there are two types of strip pooling, namely temporal and frequency strip pooling, the output $\mathbf{Y}^T \in \mathbb{R}^{C\times F}$ after temporal strip pooling is calculated as:

$$\mathbf{Y}^T(c,f) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{U}(c,t,f) \quad (8)$$

Similarly, the output $\mathbf{Y}^F \in \mathbb{R}^{C\times T}$ after frequency strip pooling can be written as:

$$\mathbf{Y}^F(c,t) = \frac{1}{F}\sum_{f=1}^{F}\mathbf{U}(c,t,f) \quad (9)$$

Compared with global average pooling (GAP), temporal or frequency information is retained in the strip frequency or temporal pooling processes, and can be further exploited to provide temporal or frequency attention.

3.2.2. Strip Pooling based attention module

The temporal strip pooling based attention(TSP_A) module is presented in Figure 2. Firstly, as described in eqn. (8), temporal strip pooling is used on an input feature map $\mathbf{U}$ to obtain output $\mathbf{Y}^T$.

Then, the attention map $\mathbf{M} \in \mathbb{R}^{C\times F}$is calculated via several non-linear operators:

$$\mathbf{M} = \sigma(\mathbf{W}_2(\delta(\mathbf{W}_1(\mathbf{Y}^T)))) \quad (10)$$

where $\mathbf{W}_1 \in \mathbb{R}^{C\times\frac{C}{r}}$, $\mathbf{W}_2 \in \mathbb{R}^{\frac{C}{r}\times C}$ denote 1×1 convolutional layers, $\sigma(\cdot)$ and $\delta(\cdot)$ denote sigmoid and ReLU respectively, and r is the reduction rate.

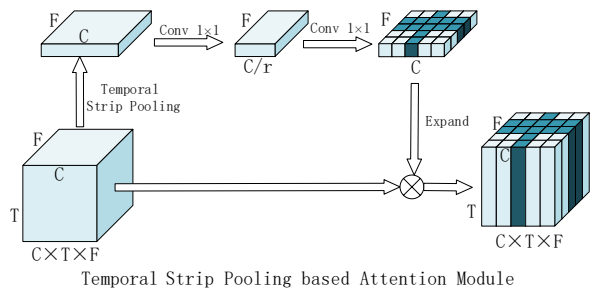

Figure 2: *Proposed strip pooling based attention module.*

Finally, the input feature map is multiplied by the attention map $\mathbf{M}$:

$$\mathbf{U}' = \mathbf{M} \otimes \mathbf{U} \quad (11)$$

where $\otimes$ denotes element-wise multiplication, and during multiplication, the 2D attention map should be expanded to match the shape of $\mathbf{U}$. Note that here we provide a formula for the temporal strip pooling based attention module – the frequency strip pooling formula is similar and can easily be deduced.

4. Experiments Setup

4.1. Dataset

All experiments are conducted on the DCASE 2020 domestic environment sound event detection (DESED) [7] dataset, which is composed of real soundscapes and synthesized soundscapes. For real soundscapes, data can be divided into 4 subsets; weakly-labeled (1578 clips), unlabeled-in-domain (14412 clips), validation(1168 clips) and evaluation. For synthesized soundscapes, 2060 background audio files from SINS and 1009 foreground audio files from Freesound are provided. Specifically, we synthesized 3190 audio clips using the Scaper toolbox [19], with the reference level set to -35 dB, the polyphony maximum limited to 3, and the foreground-background sound to noise ratio (FBSNR) range is set to 2–30 dB. In our experiments, we utilize weakly-labeled clips, unlabeled in-domain clips and synthesized clips as the training set, and evaluate the performance on the validation set.

4.2. Feature Extraction

The input features used in the proposed system are log-mel spectrograms, which are extracted from the audio signal resampled to 22050 Hz. The log-mel spectrogram uses 2048 STFT windows with a hop size of 431 and 128 Mel-scale filters. As a result, each 10-second sound clip is transformed into a 2D time-frequency representation with a size of (512×128).

4.3. Experimental Settings

In our MMT-DA based system, the two models are trained from scratch, using the Adam optimizer [20], where the maximum learning rate is set to 0.001, and the total training epochs are set to 100. Specifically, the weight for domain loss w_d is set to 0.05 and maximum value for MMT weight $w(t)$ is set to 1. We apply a ramp-up for both the learning rate and the MMT weight $w(t)$ during the first 20 epochs. For backend processing, an adaptive median filter is used, where the filter size for each event category is selected as 1/3 of its average occurrence time.

Event based macro-F1 is used as the main metric For SED tasks. The experimental results are all evaluated by the sed_eval toolbox [21]. Onsets are evaluated with a collar tolerance of 200ms. Tolerance for offsets is computed per event as the maximum of 200ms or 20% of event length.

5. Results and Discussion

We separately evaluate our proposed methods on DCASE2020 Task 4 dataset, including: (1) the MMT-DA based training method, and (2) the strip pooling based attention (SP_A) module.

To demonstrate the effectiveness of our methods, we also compare the performance of our MMT-DA based method with other techniques presented in Figure 1(a)(b)(c), including the MT based method, DA based method and the MT-DA based

method respectively. As shown in Table 1, the MT based method on real or total data can achieve F1-scores of 35.0% and 45.2% respectively, which indicates that adding strongly-labeled synthetic data to the MT system indeed leads to a substantial performance gain. Besides this, the performance of the DA based system achieved an F1 score of 45.5%, which can be further improved to 46.4% by the MT-DA based system. Compared with the previous system, our proposed MMT-DA based method can significantly boost performance to 49.4%, while applying the temporal strip pooling based attention(TSP_A) module in our MMT-DA system increases SED performance to 52.0%, which outperforms the winning system of the DCASE2020 challenge (also shown). A more detailed ablation analysis now follows.

Table 1: *Comparison of system and module performance.*

System	Model	Macro F1, %
MT (real)	CRNNbase	35.0
MT (total)	CRNNbase	45.2
DA	CRNN+GRL	45.5
MT-DA	CRNN+GRL	46.4
MMT-DA	2 models	49.4
MMT-DA	2 models+TSP_A	**52.0**
MT (Top1 [10])	Conformer	50.6
DA (Top2 [11])	CRNN+GRL	48.3

5.1. Evaluation of our MMT-DA based method

In Table 2, we first see that the MT system on real data and the DA system with total data (real and synthetic), achieved a lower F1-score of 35.0% and a higher F1-score of 45.5% respectively. We then experimented with different settings to explore the effective of MMT-DA by using the pretrained fixed teacher model or using soft labels with MSE loss or hard labels with BCE loss. We evaluated with fixed pretrained teachers, where "Base to GRL" refers to the CRNN baseline model with Macro-F1 score of 35.0% in the MT system being used to provide pseudo labels for unlabeled data, to guide the CRNN model with GRL layer in the DA system. "GRL to Base" refers to the contrary setting.

In our experiments, when using soft labels with MSE loss, the "Base to GRL" or "GRL to Base" systems both provides a performance gain over the systems without teachers, and also see the two separate systems are complementary to each other. Since synthetic data is used in the DA system, and the performance for the DA system is stronger than the MT based system with real data only, using stronger fixed DA system to teach weaker MT can improve the performance from 35.5% to 47.9%. Interestingly, using the weaker teacher from the MT based system to teach the stronger DA system also yield a slight performance gain from 45.5% to 47.7%, which has the same effect with teacher-free [22], where a weaker teacher can improve the performance of a stronger student by providing a label smoothing regularization. Compared against teacher-student systems with fixed teachers, the proposed MMT-DA based method further improves the SED performance to 49.4%. However, when using hard labels with BCE loss, due to the weaker performance of the MT based system, poor results are achieved on "Base to GRL" and the MMT-DA systems. This demonstrates the importance of using soft-labels for mutual mean teaching.

Table 2: *Evaluation of the proposed MMT-DA based method.*

Method	Pseudo label	CRNN+GRL	CRNN only
MT(real)	-	-	35.0
DA	-	45.5	-
GRL to Base	soft	45.5	47.9
Base to GRL	soft	47.7	35.0
MMT-DA	soft	49.4	47.0
GRL to Base	hard	45.5	46.8
Base to GRL	hard	43.4	35.0
MMT-DA	hard	40.2	40.6

5.2. Evaluation of SP_A module

Finally, we also evaluate the performance for the strip pooling based attention module, with the results presented in Table 3. "SE" ,"FSP" and "TSP" represent Squeeze-and-Excitation [15], frequency, and temporal strip pooling. Specially, we can see that, due to the channel attention module in the SE system, SED performance for our MMT-DA based system improves from 49.4% to 50.7%. When applying the TSP based attention module, an F1-score of 52.0% is achieved, demonstrating the importance of frequency and channel attention information. By contrast, the FSP based attention module provides a slight improvement on Macro-F1 score, indicating that the temporal attention is not as effective as frequency or channel attention. Finally, we combine FSP and TSP modules in parallel to obtain an "FTSP" based attention module, from which an F1-score of 51.3% is achieved. However, this does not outperform the TSP based attention module system.

Table 3: *Evaluation of the proposed strip pooling based attention module.*

Model	Macro F1, %
CRNNBase	49.4
CRNNBase+SE [15]	50.7
CRNNBase+TSP_A	**52.0**
CRNNBase+FSP_A	49.7
CRNNBase+FTSP_A	51.3

6. Conclusion

In this paper, a novel mutual mean teaching domain adaptation method is proposed to exploit synthetic data training data when combined with a real dataset. Specifically, we use two different student models to solve DA and SSL subproblems respectively, and soft-label mutual teaching is employed to guide each student model to utilize the unlabeled dataset. Furthermore, the performance of our MMT-DA based method can be improved by a strip pooling based attention module. By combining our proposed methods, an F1-score of 52.0% is achieved on the validation dataset of DCASE2020, outperforming the previous best result of 50.6% achieved by the winning system. In future, we aim to use more effective domain adaptation methods or attention modules to better exploit the synthetic data to improve SED performance further.

7. References

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

[2] R. Radhakrishnan, A. Divakaran, and A. Smaragdis, "Audio analysis for surveillance applications," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.* IEEE, 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," in *Advances in neural information processing systems*, 2017, pp. 1195–1204.

[6] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. P. Shah, "Large-scale weakly labeled semi-supervised sound event detection in domestic environments," *arXiv preprint arXiv:1807.10501*, 2018.

[7] N. Turpault, R. Serizel, A. Parag Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," June 2019, working paper or preprint. [Online]. Available: https://hal.inria.fr/hal-02160855

[8] Z. Shi, "Hodgepodge: Sound event detection based on ensemble of semi-supervised learning methods," Fujitsu Research and Development Center, Beijing, China, Tech. Rep., June 2019.

[9] L. Lin and X. Wang, "Guided learning convolution system for dcase 2019 task 4," Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, Tech. Rep., June 2019.

[10] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Convolution-augmented transformer for semi-supervised sound event detection," DCASE2020 Challenge, Tech. Rep., June 2020.

[11] L. Yang, J. Hao, Z. Hou, and W. Peng, "Two-stage domain adaptation for sound event detection," *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2020.

[12] L. Yang, Y. Wang, M. Gao, A. Shrivastava, K. Q. Weinberger, W.-L. Chao, and S.-N. Lim, "Mico: Mixup co-training for semi-supervised domain adaptation," *arXiv preprint arXiv:2007.12684*, 2020.

[13] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on Computational learning theory*, 1998, pp. 92–100.

[14] M.-F. Balcan, A. Blum, and K. Yang, "Co-training and expansion: Towards bridging theory and practice," *Advances in neural information processing systems*, vol. 17, pp. 89–96, 2005.

[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[16] X. Zheng, Y. Song, J. Yan, I. McLoughlin, L.-R. Dai, and L. Liu, "An effective perturbation based semi-supervised learning method for sound event detection," in *INTERSPEECH*, 2020.

[17] Y. Wang, J. Li, and F. Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 31–35.

[18] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

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

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

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

[22] L. Yuan, F. E. Tay, G. Li, T. Wang, and J. Feng, "Revisiting knowledge distillation via label smoothing regularization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.