Two-stage Domain Adaptation for Sound Event Detection

Liping Yang, Junyong Hao, Zhenwei Hou, Wang Peng

Key Laboratory of Optoelectronic Technology and Systems, MOE
Chongqing University, Chongqing 400044, China
{yanglp, junyonghao}@cqu.edu.cn

ABSTRACT

Sound event detection under real scenarios is a challenge task. Due to the great distribution mismatch of synthetic and real audio data, the performance of sound event detection model, which is trained on strong-labeled synthetic data, degrades dramatically when it is applied in real environment. To tackle the issue and improve the robustness of sound event detection model, we propose a two-stage domain adaptation sound event detection approach in this paper. The backbone convolutional recurrent neural network (CRNN) leaned using strong-labeled synthetic data is updated by weak-label supervised adaptation and frame-level adversarial domain adaptation. As a result, the parameters of CRNN are renewed for real audio data, and the input space distribution mismatch between synthetic and real audio data is mitigated in the feature space of CRNN. Moreover, a context clip-level consistency regularization between the classification outputs of CNN and CRNN is introduced to improve the feature representation ability of convolutional layers in CRNN. Experiments on DCASE 2019 sound event detection in domestic environments task demonstrate the superiority of our proposed domain adaptation approach. Our approach achieves F1 scores of 48.3% on the validation set and 49.4% on the evaluation set, which are the-state-of-art sound event detection performances of CRNN model without data augmentation.

Index Terms— Sound event detection, Domain adaptation, Adversarial learning, Convolutional recurrent neural network

1. INTRODUCTION

Sound event detection (SED) systems, which aim to detect the onset and offset of sound events and identify events categories, have potential applications in audio events classification [1], anomalous sound detection [2], automatic monitor [3] and etc. However, performances of SED systems decrease dramatically under real scenarios due to the complexity of environment and the lack of strongly labeled real audio data.

To generate a robust sound event detection model under real scenarios, researchers usually employ two strategies: (1) weak-label supervised learning, where lots of real audio samples are weakly labeled and utilized to train a SED model; (2) semi-supervised learning, where supervised SED model trained using synthetic audio dataset is generalized using unlabeled real audio data.

Weak-label supervised sound event detection belongs to weak supervision learning, where the labels for audio samples provide event category information but miss the onset and offset of each event. Multi-instance learning, which treats each audio sample as a bag and the frames of the sample as instances, is a useful approach for SED [4]. Nevertheless, due to the property of weak supervision, weak-label supervised SED cannot precisely predict the onset and offset positions in an audio sequence.

Compared with tedious sound event accurate labeling in real data, it is much easier to collect a certain number of sound event samples and background sound. Synthesizing these sound event samples and background sound to generate high-quality labeled audio sequences for supervised SED is sensible [5]. In the Challenge of Detection and Classification of Acoustic Scenes and Events (DCASE) 2016 [6], supervised SED methods were tested both on synthetic and real-life audio datasets. The results demonstrated that sound event detection in a realistic setting was difficult. Undoubtedly, SED models trained using synthetic sequences perform robust-less under real scenarios due to the statistical distribution mismatch between synthetic and real audio data. To overcome the distribution mismatch, several semi-supervised learning approaches were proposed [7, 8]. Mean teacher method, who won the DCASE 2018 challenge, is the representative semi-supervised learning approach in sound event detection [9, 10].

However, semi-supervised SED model generalization is inadequate for fitting the gap of distribution mismatch between synthetic and real audio data. Domain adaptation, which is widely used in image classification [11] and acoustic scene classification [12], is another choice to alleviate the distribution mismatch. In domain adaptation, the space of synthetic and real audio datasets can be treated as source and target domain respectively. The objective is transferring SED models trained on source domain to target domain. In [13], the authors introduced a knowledge transfer method, which effectively transferred knowledge from a source domain convolutional neural network (CNN) model to the model in target domain, for sound event classification.

In this paper, we propose an end-to-end two-stage domain adaptation method for robust SED under real scenarios. We employ a convolutional recurrent neural network (CRNN) as the backbone network for sound event detection. We first train the CRNN model using strong-labeled synthetic audio samples in a supervised learning manner. Then, two-stage domain adaptation is conducted to align the distributions of synthetic and real audio samples in feature space: (1) Parameters of the CRNN trained using synthetic audio data are updated with weakly supervised learning on weak-labeled real audio samples. (2) CRNN features of both source domain synthetic audio samples and target domain real audio samples (including weak-labeled and unlabeled) are fed into a domain discriminator for frame-level adversarial learning. A zero-sum game between CRNN and domain discriminator is performed to renew the parameters of CRNN for feature alignment between source and target domain samples. In addition, to improve the source and target domain feature representation ability of convolutional layers in the CRNN, we design a context clip-level consistency constraint between the classification outputs of CNN and RNN. Experimental results on DCASE 2019 sound event

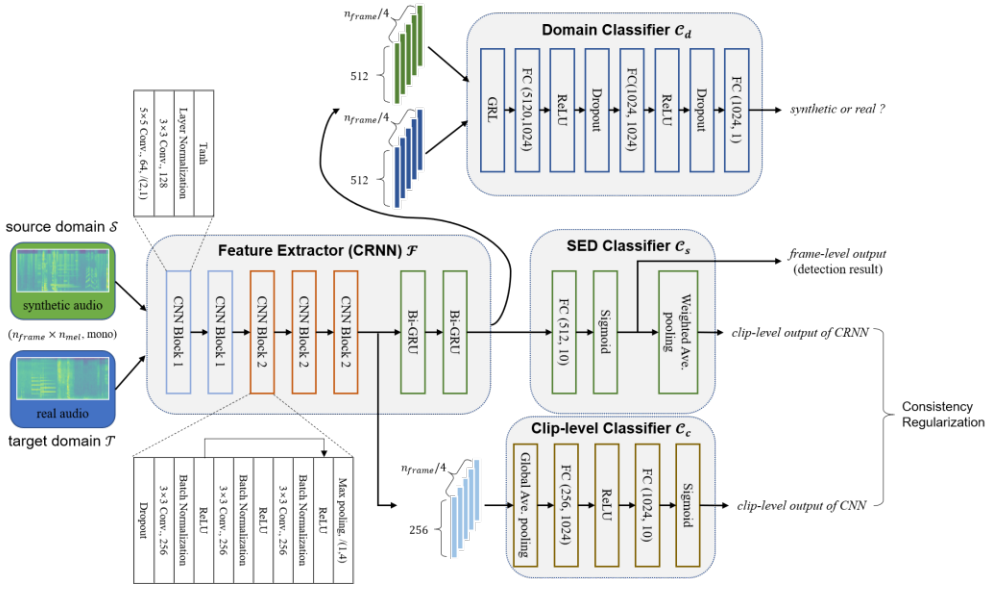

Figure 1. Framework of domain adaptation sound event detection model.

detection in domestic environments (task 4) demonstrate that our proposed two-stage domain adaptation method can efficiently transfer SED models from source domain to target domain and significantly promote the robustness of sound event detection in real life.

2. TWO-STAGE DOMAIN ADAPTATION

Convolutional recurrent neural network, which consists of several cascaded convolutional layers and gated recurrent units, is the representative model for sound event detection [7, 9]. In this paper, the overall domain adaptation sound event detection framework is shown in Figure 1. We employ a CRNN with 13 convolutional layers and 2 bidirectional gated recurrent units (Bi-GRU) as backbone feature extraction network $\mathcal{F}$ of domain adaptation sound event detection. Different from the baseline system of DCASE 2019, we use two types CNN blocks. In the first block (CNN Block 1), we use a layer normalization (LN) operation [14] to replace commonly used batch normalization (BN) operation in the early stage of CRNN. In the second block (CNN Block 2), a shortcut is added between the first and last ReLU. Audio sample features extracted using $\mathcal{F}$ are then fed to SED classifier $\mathcal{C}_s$ to produce frame-level and clip-level outputs. To overcome the distribution mismatch between synthetic and real audio samples, we design a domain classifier $\mathcal{C}_d$, which consists of several fully connected layers, to discriminate synthetic and real audio samples for domain adversarial learning. Finally, a context clip-level consistency constraint between the classification outputs of CNN and CRNN is used to improve feature representation ability of convolutional layers.

In sound event detection task, we have three types of audio datasets: strong-labeled synthetic dataset $\mathcal{D}_S = \{(x_s^i, y_s^i)\}_{i=1}^{Ns}$, weak-labeled real dataset $\mathcal{D}_w = \{(x_w^i, y_w^i)\}_{i=1}^{Nw}$ and unlabeled real dataset $\mathcal{D}_u = \{x_u^i\}_{i=1}^{Nu}$, where y_s^i denotes the frame-level label for a synthetic sample x_s^i, y_w^i denotes the weak label for a weak-labeled real sample x_w^i and x_u^i denotes a unlabeled real sample. From the perspective of domain adaptation, all the samples from synthetic dataset belong to source domain (domain label is $d = 0$), all the samples from real datasets belong to target domain (domain label is $d = 1$). Therefore, the source domain dataset is $\mathcal{S} = \mathcal{D}_S$, and the target domain dataset is $\mathcal{T} = \mathcal{D}_w \cup \mathcal{D}_u$.

To generate a robust CRNN sound event detection model under real scenarios, we transfer a backbone CRNN learned in source domain to target domain by the proposed two-stage domain adaptation method.

2.1. Learning CRNN on strong-labeled synthetic audio

Given a strong-labeled synthetic audio dataset $\mathcal{D}_S = \{(x_s^i, y_s^i)\}_{i=1}^{Ns}$, parameters of the backbone CRNN can be learned in a supervised manner by minimizing the binary cross entropy (BCE) loss of all samples in $\mathcal{D}_S$:

$$L_s = -\frac{1}{N_s}\sum_{i=1}^{N_s} y_s^i \log\left(\bar{y}_s^i\right) + \left(1 - y_s^i\right)\log\left(1 - \bar{y}_s^i\right) \qquad (1)$$

where $\bar{y}_s^i$ is the frame-level prediction probabilities of $\mathcal{C}_s$.

Since that there is a great gap between the distribution of synthetic and real audio samples, the CRNN model learned on strong-labeled synthetic audio cannot be directly used under real scenarios. On the other hand, training the CRNN by minimizing the BCE loss is equivalent to minimizing the empirical risk of given samples. Therefore, the CRNN model trained by synthetic audio is biased and need to do generalization.

2.2. Domain Adaptation

Since that there are two types of real audio samples, weak-labeled and unlabeled samples, we design a two-stage domain adaptation method to transfer the backbone CRNN learned in source domain to target domain. At the first stage, we use weak-labeled real audio samples to update the parameters of CRNN by performing weak-label supervised learning. At the second stage, frame-level adversarial learning is conducted between synthetic audio samples and all real audio samples to align source and target domain feature distributions.

2.2.1. Weak-label supervised adaptation

Given a weak-labeled dataset $\mathcal{D}_w = \{(x_w^i, y_w^i)\}_{i=1}^{Nw}$ in target domain, the parameters of CRNN are updated by performing weak-label supervised learning. The loss function is:

$$L_w = -\frac{1}{N_w}\sum_{i=1}^{N_w} y_w^i \log\left(\bar{y}_w^i\right) + \left(1 - y_w^i\right)\log(1 - \bar{y}_w^i) \qquad (2)$$

where $\bar{y}_w^i$ is the clip-level prediction output of CRNN [15].

At this stage, we use weak-labeled real audio samples of the target domain to update CRNN model. The parameters of CRNN is optimized by minimizing weak-label BCE loss shown in (2). Therefore, such a weak-label supervising learning procedure can be regarded as a model transformation procedure from source domain to target domain. Moreover, due to that weak-label supervised learning provides a constraint on clip-level, it cannot guarantee that the transformed model is suitable for sound event detection task, which is a frame-level classification task.

2.2.2. Frame-level adversarial learning domain adaptation

To investigate a frame-level domain adaptation solution, we introduce the idea of adversarial learning at this stage. Suppose that the source domain synthetic audio dataset is $\mathcal{S} = \{x_s^i, d = 0\}_{i=1}^{N_s}$, the target domain real audio dataset is $\mathcal{T} = \{x_T^i, d = 1\}_{i=1}^{N_T}$, and $N = N_s + N_T$. For a sample $x_s^i \in \mathcal{S}$ and a sample $x_T^i \in \mathcal{T}$, we can get their feature maps $\mathcal{F}(x_s^i)$ and $\mathcal{F}(x_T^i)$ with the previously trained CRNN, i.e., the feature extractor $\mathcal{F}$. The feature maps for each frame of all the source and target domain samples are fed into the domain classifier $\mathcal{C}_d$ to make frame-level decisions whether they are synthetic or real. The feature extractor $\mathcal{F}$ and the domain classifier $\mathcal{C}_d$ are trained in an adversarial manner. Inspired by [16], the input of $\mathcal{C}_d$ is a joint variable of each frame feature and its SED prediction.

For the implementation of adversarial learning, a gradient reverse layer (GRL) [17] is used to identity transform the feature maps before they are fed to and forward propagate in $\mathcal{C}_d$. During back-propagation, the gradient is reversed by changing its sign when it passes through the GRL. As a result, the feature extractor $\mathcal{F}$ and the domain classifier $\mathcal{C}_d$ are optimized in a reverse direction. The objective function of adversarial learning is:

$$L_d = -\frac{1}{N}\sum_{i=1}^{N}\left(d\ log(\overline{d}) + (1-d)log\left(1-\overline{d}\right)\right) \qquad (3)$$

where, $\overline{d}$ is the output of domain classifier $\mathcal{C}_d$.

By minimizing (3), the domain classifier $\mathcal{C}_d$ gets the ability of discriminating the distribution mismatch between source and target domain. Meanwhile, the feature extractor $\mathcal{F}$ can extract domain invariant features. In this way, source and target domain sample distributions are aligned in the feature space.

2.3. Clip-level consistency regularization

As stated earlier, the backbone CRNN and SED classifier $\mathcal{C}_s$ learned using synthetic audio samples is biased. To improve feature representation power and classification generalization ability of transferred CRNN, a clip-level consistency regularization strategy is introduced.

Given all the N audio samples in source and target domain, as shown in Figure 1, the CRNN clip-level prediction $\bar{y}^i$ of each sample is obtained by weighted average pooling of its frame-level predictions [9]. On the other hand, we feed the CNN feature map of the sample to a clip-level classifier $\mathcal{C}_c$ and get another clip-level prediction $\bar{y}_{cnn}^i$. The clip-level consistency regularization is achieved by minimizing the ℓ_1 norm:

$$L_c = \frac{1}{N}\sum_{i=1}^{N}\left\|\bar{y}_{cnn}^i - \bar{y}^i\right\|_1 \qquad (4)$$

The objective of this clip-level consistency regularization is keeping the clip-level predictions of CNN and CRNN the same. Based on this, the convolutional layers of CRNN grab more feature representation of the whole audio sample, so that the bidirectional recurrent layers of CRNN are able to describe the frame relationship of the audio sample more effectively.

To implement the proposed two-stage domain adaptation, we minimize the loss functions (1) ~ (4) sequentially on each mini-batch sample in each epoch. After hundreds of iterations, a robust CRNN model for sound event detection under real scenarios is obtained.

3. EXPERIMENTS

Experiments on DCASE 2019 sound event detection in domestic environments (task 4) task are conducted to demonstrate performances of the proposed two-stage domain adaptation approach.

3.1. Dataset

The dataset of DCASE 2019 task 4 has 10 sound event classes, including *alarm*/*bell*/*ringing*, *blender*, *cat*, *dishes*, *dog*, *electric shaver*/*toothbrush*, *frying*, *running water*, *speech* and *vacuum cleaner*. The dataset includes 2045 synthetic audio clips and 14850 real audio chips in total. The synthetic audio clips are generated with Scaper [5]. The real audio clips are extracted from Audioset [18], which contains 2 million human-labeled sound clips drawn from 2 million Youtube videos.

To compare with the existing methods, we use the dataset according to the partition rules given by official DCASE challenge. The dataset is divided into a development set and an evaluation set. The development set contains two partitions: training set and validation set. In training set, there are three types of audio clips, including 2045 synthetic strongly labeled clips, 1578 weal-labeled real audio clips and 11412 unlabeled real audio clips. The validation set contains 1168 strongly labeled real audio clips. The evaluation set contains 692 strongly labeled real audio clips. Both validation set and evaluation set are used to evaluate the performances of SED systems.

3.2. Experimental setup

In our experiments, mel-spectrogram features are used as the basic features for sound event detection. Each audio clip in the dataset is down-sampled at 22050Hz and transformed using fast Fourier transform with a 2048 points Hanning window and 1617 points overlap. Then, a mel filter-bank with 64 bandpass filters is applied to obtain the mel-spectrogram feature of the clip. Finally, the values of mel-spectrogram are normalized to 0-1. As a result,

Table 1. Sound event detection performances of backbone CRNN and two-stage domain adaptation approach

	backbone CRNN						two-stage domain adaptation					
	validation set			evaluation set			validation set			evaluation set		
	F1(%)	DR	IR	F1(%)	DR	IR	F1(%)	DR	IR	F1(%)	DR	IR
alarm	32.6	0.74	0.33	23.0	0.84	0.22	46.1	0.61	0.29	47.9	0.62	0.20
blender	29.1	0.69	0.83	20.2	0.80	0.80	59.5	0.40	0.43	48.8	0.50	0.55
cat	7.3	0.94	0.56	10.5	0.93	0.35	48.9	0.59	0.27	63.5	0.42	0.25
dishes	20.2	0.83	0.54	21.3	0.85	0.27	28.5	0.76	0.43	35.5	0.75	0.17
dog	9.3	0.93	0.35	7.8	0.95	0.23	29.8	0.74	0.47	40.5	0.68	0.27
electric	48.8	0.54	0.43	33.5	0.73	0.33	62.9	0.40	0.31	55.2	0.56	0.17
frying	35.5	0.64	0.68	44.0	0.63	0.30	40.6	0.55	0.77	50.3	0.54	0.36
running water	27.7	0.71	0.78	22.9	0.75	0.92	44.9	0.60	0.37	35.7	0.70	0.39
speech	27.5	0.78	0.35	27.4	0.79	0.34	51.6	0.54	0.33	59.8	0.48	0.23
vacuum	54.2	0.48	0.40	51.6	0.51	0.41	69.7	0.34	0.24	57.3	0.49	0.27
overall	29.2	0.73	0.53	26.2	0.78	0.42	48.3	0.55	0.39	49.4	0.57	0.29

we obtain a 512×64 two-dimensional normalized mel-spectrogram feature for an audio clip in the dataset. During training process, we use Adam to optimize all the loss functions of our two-stage domain adaptation approach.

Median filtering and tagging embedding [19] are employed as post processing strategies of sound event detection. The threshold of tagging embedding is set to 0.5 for filtering out the unstable frame-level predictions.

3.3. Experimental results

In this section, we conduct experiments to verify the effectiveness of our two-stage domain adaptation method. First of all, an ablation experiment is implemented to verify the feasibility of using the proposed two-stage domain adaptation approach on transferring the backbone CRNN. Then, the performances of the proposed two-stage domain adaptation approach are compared with some representative sound event detection systems in DCASE 2019. We use F1-measure (F1), deletion rate (DR) and insertion rate (IR) to evaluate the performances of these systems.

To verify that our proposed two-stage domain adaptation approach can effectively transfer the CRNN SED model trained on synthetic audio to real audio, we compare the sound event detection performances of the backbone CRNN with that of the two-stage domain adaptation approach. Table 1 lists the F1 scores, deletion rates and insertion rates of backbone CRNN and two-stage domain adaptation approach. The overall F1 scores of two-stage domain adaptation approach are 19.0% and 23.2% higher than that of the backbone CRNN on validation set and evaluation set. It is obvious that the performances of two-stage domain adaptation approach significantly outperform those of backbone CRNN, which demonstrates the superiority of the two-stage domain adaptation approach. From the results of the two comparative experiments on validation set and evaluation set, we can see that *cat* and *dog* have the most significant performances improvement when the two-stage domain adaptation approach is used. Biological sound events are usually rich diversity. When training data is limited, it is difficult for neural networks to learn the diversity of events. Nevertheless, adversarial learning in the proposed two-stage domain adaptation approach can effectively force CRNN feature extractor to extract domain-invariant features. As a result, the diversity of biological sounds may decrease in our proposed two-stage domain adaptation approach.

To further illustrate the performance of the two-stage domain adaptation method, we compare the sound event detection results of our proposed approach with several representative SED systems of DCASE 2019 in Table 2. Compared to the baseline system of DCASE 2019, which is the top-1 solution of DCASE 2018, our backbone CRNN achieved similar performance both on validation set and evaluation set. Combining our backbone CRNN with semi-supervised mean teacher algorithm [9] achieves top-5 on validation dataset and top-2 on evaluation dataset. Compared with the semi-supervised supervised learning of [7,20,22,23], Our two-stage domain adaptation method has achieved best results both on validation set and evaluation set. Moreover, data augmentation and model fusion methods are used in [7,20] to improve SED performance. However, our backbone CRNN and the proposed two-stage domain adaptation use a single CRNN model without data augmentation. Therefore, the method we proposed is very simple and effective.

Table 2 F1 scores (%) of the proposed approach and the representative SED systems for DCASE 2019 task 4.

methods	validation	evaluation
Lin [19]	45.3	47.7
Delphin [7]	43.6	45.8
Shi [20]	42.5	46.1
Pellegrini [21]	39.9	43.0
Yan [22]	42.6	38.8
Lim [23]	40.9	38.6
Baseline (DCASE2019)	23.7	29.0
backbone CRNN	29.2	26.2
backbone CRNN+Mean teacher	42.2	46.7
two-stage domain adaptation	**48.3**	**49.4**

4. CONCLUSIONS

In this paper, we present a two-stage domain adaptation approach for sound event detection. The approach transfers a backbone convolutional recurrent neural network trained on synthetic audio data to real audio data using weak-label supervised adaptation and adversarial domain adaptation. Through domain adaptation, the distribution mismatch between synthetic data and real data is mitigated in feature space. The performance of the transferred CRNN model significantly outperforms that of the CRNN model trained on synthetic data under real environment. In addition, the clip-level consistency regularization between the classification outputs of CNN and CRNN is able to promote the feature representation power and generalization ability of the transferred CRNN model.

5. REFERENCES

[1] I. Martin-Morato, M. Cobos, and F. Ferri, "Adaptive midterm representations for robust audio event classification," IEEE/ACM Trans. Audio Speech Lang., vol. 26, no. 12, pp.2381–2392, 2018.

[2] A. I. Humayun, S. Ghaffarzadegan, Z. Feng, and T. Hasan, "Learning front-end filter-bank parameters using convolutional neural networks for abnormal heart sound detection," in Proc. IEEE EMBC, 2018, pp. 1408–1411.

[3] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," IEEE Trans. Audio Speech Lang., vol. 10, no. 5, pp. 293–302, 2002.

[4] A. Kumar and B. Raj, "Audio event detection using weakly labeled data," in Proc. ACM MM, 2016, pp. 1038–1047.

[5] N. Turpault, R. Serizel, J. Salamon, and A. P. Shah, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in Proc. IEEE DCASE, 2019,pp. 253–257.

[6] A. Mesaros, H. Toni, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," IEEE/ACM Trans. Audio Speech Lang., vol. 26, no. 2, pp. 379–393, 2018.

[7] L. Delphin-Poulat and C. Plapous, "Mean teacher with data augmentation for DCASE 2019 task 4," Orange Labs Lannion, France, Tech. Rep., 2019.

[8] T. K. Chan, C. S. Chin, and Y. Li, "Non-negative matrix factorization-convolutional neural network (NMF-CNN) for sound event detection," in Proc. IEEE DCASE, 2019, pp. 40–44.

[9] J. K. Lu, "Mean teacher convolution system for DCASE 2018 task 4," PFU Shanghai Co., LTD, China, Tech. Rep., 2018.

[10] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in Proc. NIPS, 2017, pp. 1195–1204.

[11] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in Proc. ICML, 2015, pp. 97–105.

[12] P. Primus, H. Eghbal-zadeh, D. Eitelsebner, K. Koutini, A. Arzt, and G. Widmer, "Exploiting parallel audio recordings to enforce device invariance in CNN-based acoustic scene classification," in Proc. IEEE DCASE, 2019, pp. 204–208.

[13] A. Kumar, M. Khadkevich, and C. Fugen, "Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes," in Proc. IEEE ICASSP, 2018, pp. 326–330.

[14] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, arXiv preprint arXiv:1607.06450.

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

[16] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in Proc. NIPS, 2018, pp. 1640–1650.

[17] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in Proc. ICML, 2015, pp. 1180–1189.

[18] 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 Proc. IEEE ICASSP, 2017, pp. 776–780.

[19] L. Lin, X. Wang, H. Liu, and Y. Qian, "Specialized decision surface and disentangled feature for weakly-supervised polyphonic sound event detection," IEEE/ACM Trans. Audio Speech Lang., vol. 28, no. 4, pp. 1466–1478, 2020.

[20] Z. Shi, L. Liu, H. Lin, and R. Liu, "HODGEPODGE: Sound event detection based on ensemble of semi-supervised learning methods," in Proc. IEEE DCASE, 2019, pp. 224–228.

[21] L. Cances, T. Pellegrini, and P. Guyot, "Multi task learning and post processing optimization for sound event detection," IRIT, Universite de Toulouse, France, Tech. Rep., 2019.

[22] J. Yan and Y. Song, "Weakly labeled sound event detection with residual CRNN using semi-supervised method," University of Science and Technology, China, Tech. Rep., 2019.

[23] W. Lim, S. Sun, S. Park, and Y. Jeong, "Sound event detection in domestic environments using ensemble of convolutional recurrent neural networks," Electronics and Telecommunications Research Institute, Korea, Tech. Rep., 2019.