CNN-based Discriminative Training for Domain Compensation in Acoustic Event Detection with Frame-wise Classifier

Tiantian Tang*, Xinyuan Zhou*, Yanhua Long*, Yijie Li† and Jiaen Liang†
* Shanghai Normal University, Shanghai, China
† Unisound AI Technology Co., Ltd., Beijing, China
E-mail: 1000479042@smail.shnu.edu.cn, yanhua@shnu.edu.cn, {liyijie,liangjiaen}@unisound.com

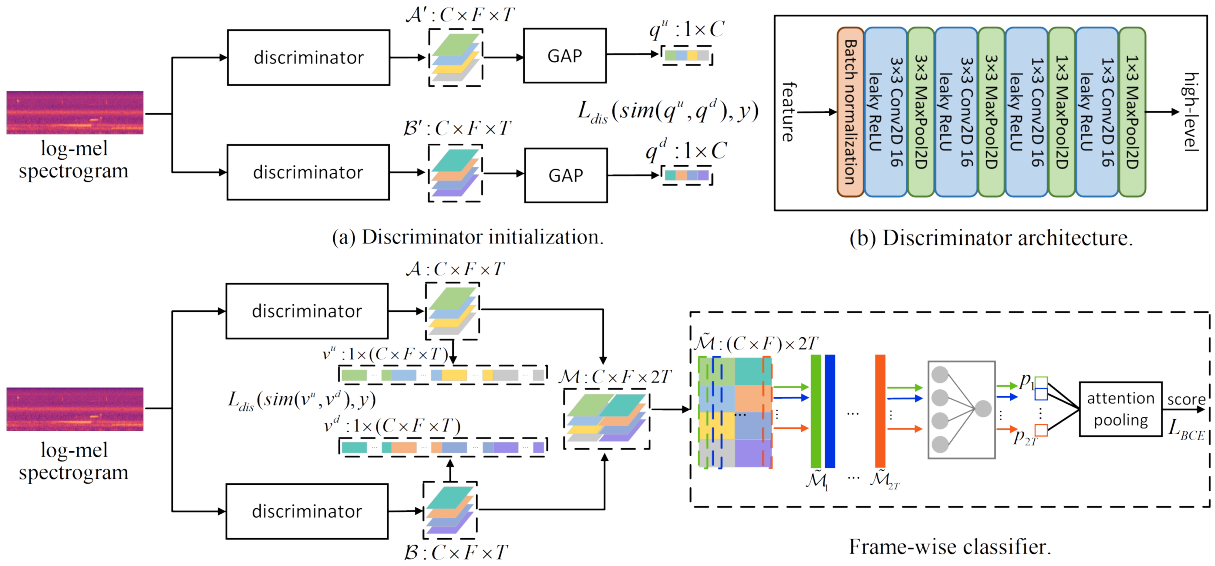
***Abstract*—Domain mismatch is a noteworthy issue in acoustic event detection tasks, as the target domain data is difficult to access in most real applications. In this study, we propose a novel CNN-based discriminative training framework as a domain compensation method to handle this issue. It uses a parallel CNN-based discriminator to learn a pair of high-level intermediate acoustic representations. Together with a binary discriminative loss, the discriminators are forced to maximally exploit the discrimination of heterogeneous acoustic information in each audio clip with target events, which results in a robust paired representations that can well discriminate the target events and background/domain variations separately. Moreover, to better learn the transient characteristics of target events, a frame-wise classifier is designed to perform the final classification. In addition, a two-stage training with the CNN-based discriminator initialization is further proposed to enhance the system training. All experiments are performed on the DCASE 2018 Task3 datasets. Results show that our proposal significantly outperforms the official baseline on cross-domain conditions in AUC by relative 1.8-12.1% without any performance degradation on in-domain evaluation conditions.**

I. Introduction

Acoustic event detection (AED) refers to the task of detecting whether interested target events occur in audios such as running water, cough, meow, etc. With the launch of the Detection and Classification of Acoustic Scenes and Events (DCASE) challenges from 2013[1], a set of AED-related tasks are provided for research and progress comparison of state-of-the-art techniques. The bird audio detection (BAD) [2] task is the DCASE 2018 Task3 that aims to detect the presence/absence of bird sound in audio clips under variety bird species, recording and background conditions. To solve this task well, the approaches are required to inherently generalize across conditions or can be self-adapted to new datasets, because there is big domain mismatch between training and evaluation sets. In this study, we also focus on the domain mismatch issue for BAD task, because the source and target domain mismatch is a common problem in most AED tasks[3], [4], and it typically results in a severe performance degradation in practical applications [5].

In the literature, only few previous works have been proposed to improve the domain robustness of BAD systems. Such as in [6], authors applied a per-channel energy normalization to alleviate the outdoor acoustic environment distortions. In [7], authors used the wasserstein distance guided representation learning [8] to incorporate the domain knowledge during model training. And works in [9] applied the CORrelation ALignment [10] to minimize the domain shift by aligning the second-order statistics of source and target distributions. There are also some domain compensation or adaptation methods are proposed for other acoustic processing tasks [11], [12], [13], [14], [15], [16]. For example, in acoustic scene classification task, a spectrum correction [17] was proposed to corrected the mismatched front-end by adjusting the varying frequency response of different recording devices. [18] proposed a neural label embedding together with a relational teacher-student learning to perform the device adaptation. And in [19], the unsupervised adversarial learning was used to leverage an extra domain discriminator for device adaptation and it was further generalized for AED tasks in [3].

Unlike previous domain adaptation methods, in this study, we deal with the domain mismatch in BAD tasks by proposing a novel discriminative training framework with two CNN-based discriminators, where each input audio clip is transformed into a pair of discriminative high-level acoustic representations before feeding them to the back-end binary classifier. It is motivated by the intuition that extracting high-level representation using standard network as CNN or LSTM optimized only by the final task-dependent loss might not be the best choice, as it may tend to be trapped in local optima and fail to extract the fine-grained heterogeneous acoustic information between targets and interferences we need. Therefore, we wonder if it is possible to learn two discriminative representations from each audio clip instead of one to enhance the domain robustness of AED systems. In the BAD task, we design a two-stage training strategy with a binary discriminative loss to force the CNN-based discriminators to learn the acoustic discrimination between bird calls and background interferences separately. The resulted paired representations are then feed into a specially designed frame-wise classifier to further capture the transient characteristics. All experiments are performed on the DCASE 2018 Task3 datasets. Compared with the official baseline system [20], results show that our proposed framework can achieve significant performance improvements on cross-domain test conditions without degrading performance of in-domain test conditions.



(a) Discriminator initialization.

(b) Discriminator architecture.

(c) The BAD system with the parallel CNN-based discriminator.

Fig. 1. Framework of the proposed CNN-based discriminative training with frame-wise classifier.

II. Proposed Method

A. Architecture

Fig. 1 gives an overview of our proposed method. As in block (c), the BAD system consists of two parts: a parallel CNN-based discriminator and followed by a frame-wise binary classifier. The discriminators are with the same structure as shown in block (b). Given a labeled dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ with N samples, where x_i is the input audio clip and $y_i \in \{0, 1\}$ is a label to indicate the absence/presence of any bird calls within that clip. Each input x is first transformed to a log-mel spectrogram $X \in \mathbb{R}^{1\times F\times T}$, where F and T are the number of mel-frequency bins and frames respectively. By taking X as input feature, the parallel discriminator $d_{cnn}(\cdot, \theta)$ with C-channel output layers are employed to obtain a pair of intermediate discriminative representations $\mathcal{A} \in \mathbb{R}^{C\times F\times T}$ and $\mathcal{B} \in \mathbb{R}^{C\times F\times T}$,

$$\mathcal{A} = d_{\text{cnn}}(X, \theta_u) \tag{1}$$

$$\mathcal{B} = d_{\text{cnn}}(X, \theta_d) \tag{2}$$

where θ_u and θ_d denote the CNN parameters of the upper and lower discriminators in Fig. 1 (c) respectively. Then we concatenate $\mathcal{A}$ and $\mathcal{B}$ into a tensor $\mathcal{M} \in \mathbb{R}^{C\times F\times 2T}$ as

$$\mathcal{M} = [\mathcal{A}, \mathcal{B}] \tag{3}$$

Next, we reshape $\mathcal{M}$ into a $(C \times F) \times 2T$ matrix $\tilde{\mathcal{M}}$ as input of the frame-wise classifier $\mathcal{G}(\cdot)$ to perform the binary classification. Details of the parallel discriminator, the frame-wise binary classifier and system training strategy are presented in the following subsections.

B. CNN-based Discriminator

Motivated by the fine-grained structure extraction [21] and domain-invariant representation learning [22] for image classification, here we investigate to use two same structure CNN networks as a parallel discriminator that shown in Fig.1 (c) to extract a pair of intermediate discriminative representations to enhance the bird calls detection system. We aim to mine the intermediate feature discrimination of heterogeneous acoustic information that embedded in each input audio clip, such as the target events (bird calls) and variety background interferences in a given clip of BAD task.

To force the designed parallel discriminator to well learn the heterogeneous acoustic characteristic separately, we introduce a novel binary discriminative loss as a training criteria to trained the parallel discriminator simultaneously as below:

$$L_{\text{dis}}(s, y) = -[y \log(1 - \mathcal{H}(s)) + \lambda(1 - y) \log(\mathcal{H}(s))] \tag{4}$$

where $y \in \{0, 1\}$ denotes the ground-truth label of input clip, $\lambda \in [0, 1]$ is a tuning parameter to balance the loss contribution of positive and negative training data. $\mathcal{H}(\cdot)$ is the rectified linear unit (ReLU) to ensure that the s is non-negative. s is a cosine similarity measure that defined as :

$$s = \text{sim}(\mathcal{T}(\mathcal{A}), \mathcal{T}(\mathcal{B})) = \frac{\mathcal{T}(\mathcal{A}) \cdot \mathcal{T}(\mathcal{B})}{\|\mathcal{T}(\mathcal{A})\| \, \|\mathcal{T}(\mathcal{B})\|} \tag{5}$$

where $\mathcal{T}(\cdot)$ is the flatten operation that transforms a tensor into a vector, as show in Fig. 1 (c), the $\mathcal{T}(\mathcal{A}), \mathcal{T}(\mathcal{B})$ transforms the $\mathcal{A}, \mathcal{B}$ into vector v^u, v^d respectively.

Based on (4), we minimize the $L_{\text{dis}}(s, y)$ to achieve our goal. It means that when the input clip is a positive sample ($y = 1$) with heterogeneous acoustic information, i.e., the clip is a mixture signal that contains both bird calls and background

noises, then only the 1st part $y\log(1-\mathcal{H}(s))$ contributes to $L_{\text{dis}}(s,y)$ and maximum $\mathcal{H}(s)=0$ ($\mathcal{A}$ and $\mathcal{B}$ are totally different, one learns background characteristics, while the other emphasizes the bird calls). When the input is a negative sample ($y=0$, only background sounds), only the 2nd part of (4) contributes to the loss, then the discriminators do not differentiate their outputs, because minimizing $L_{\text{dis}}(s,y)$ leads to a maximum value of $\mathcal{H}(s)=1$, it results similar paired $\mathcal{A}$ and $\mathcal{B}$. That's to say, $L_{\text{dis}}(s,y)$ only discriminates the heterogeneous information in positive samples to highlight the acoustic characteristics of target events. These target representations should be more robust to domain variation, they can be taken as domain-invariant intermediate features because they capture the acoustic properties of bird calls more explicitly.

In most real applications as our BAD task, the positive samples are always with background sounds. Therefore, we think that if the BAD system is trained on a source domain, the robust target representations learned from the parallel CNN-based discriminator can leverage a better model generalization to the target domain bird call detection. That's to say, the discriminators play a domain compensation role in the whole BAD system.

C. Frame-wise (F-W) Binary Classifier

Different from other bioacoustics signals, the bird calls are normally short, their spectrograms have strong transient characteristics. Instead of using the conventional classifier of official baseline [23] that accepts the whole flattened CNN feature maps as one input, here we propose to use a frame-wise classifier to learn the transient bird chirping characteristics. As shown in the last block of Fig.1, each column $\tilde{\mathcal{M}}_j \in \mathbb{R}^{C\times F}$, $j=1,2,\ldots,2T$, is taken as j-th frame, then each $\tilde{\mathcal{M}}_j$ is learned independently by a two-layers feed-forward neural network (FFN) followed with a sigmoid activation to achieve a prediction score p_j. All $\tilde{\mathcal{M}}_j$ share the same FFN parameters. Finally, all the $2T$ prediction scores are further attention-weighted by the attention pooling [24], [25] to automatically control their contribution for decision making. The final score p is computed as:

$$p=(\sum_j p_j w_j)/\sum_j w_j \tag{6}$$

where w_j is the learnable weight for each p_j. Details of the attention pooling can be found in [26].

D. Two-stage Training Strategy

In intuition, the proposed BAD system in Fig.1(c) should be trained in one-stage using a combination loss that defined as:

$$L_{\text{total}}=L_{\text{dis}}(\text{sim}(v^u,v^d),y)+L_{\text{BCE}}(\hat{y},y) \tag{7}$$

where $\hat{y}$ is the final prediction score p, $L_{\text{dis}}(\cdot)$ is the binary discriminative loss defined in (4), and $L_{\text{BCE}}(\cdot)$ is the tradition binary cross entropy (BCE) loss as in [27], [28]. v^u, v^d are the flattened representations used in (5).

However, from our extensive tryout experiments, we find that it's better to use a two-stage training strategy with the parallel CNN-based discriminator initialization. In `stage 1`, as shown in Fig.1(a), we only train the parallel discriminator using a binary discriminative loss defined as,

$$L_{\text{pre}}=L_{\text{dis}}(\text{sim}(q^u,q^d),y) \tag{8}$$

where $q^u=\text{GAP}(\mathcal{A}')$, $q^d=\text{GAP}(\mathcal{B}')$, the $\mathcal{A}',\mathcal{B}'$ and the size of q^u,q^d are illustrated in Fig.1(a). The GAP denotes using the global average pooling [29] to map each channel representation into a average one. It is different from the flatten that used in one-stage training loss.

Based on the well pre-trained discriminators, in `stage 2`, the whole system is then trained using the above combination loss L_{total}, but with the discriminators are initialized by the pre-trained CNN parameters in `stage 1`. We speculate that an effective initialization may avoid local optima and provide a good guidance to enhance the whole model training, because the pre-trained discriminators can provide a stable and discriminative perception to the frame-wise classifier.

III. Experimental Setup

A. Dataset

The DCASE 2018 Task 3 (bird audio detection) provides 3 separate labeled development and 3 evaluation datasets, each recorded under different conditions. As the ground-truth of evaluation set is not released publicly. Only the development sets are used in our work. The datasets have different balances of positive/negative cases, different bird species and a wide-domain coverage of background sounds and recording equipments. Each audio clip is 10s-length and sampled at 44.1kHz.

Specifically, three development sets are the "freefield1010" (ff1010bird), the "warblrb10k" and the "BirdVox-DCASE-20k" (BirdVox-20k). The ff1010bird contains 7,690 excerpts from field recordings around the world with a diverse location and environments. The warblrb10k contains 8,000 smartphone audio recordings from around the UK, the audio covers a wide distribution of UK locations and environments, and it includes weather/traffic noise, human speech and even human bird imitations. The BirdVox-20k consists of 20,000 audio clips that collected from remote monitoring units placed near Ithaca, NY, USA during the autumn of 2015. Compared with ff1010bird and BirdVox-20k, the warblrb10k contains much more diverse background acoustics. Instead of using the experimental procedure recommended by DCASE challenge to achieve one general model, our goal is to examine the model generalization ability for cross-domain evaluation tasks, so we construct our own BAD tasks using the provided development sets, for each of the above mentioned dataset, we select 60%, 20%, 20% audio clips for training, validation and test respectively.

B. Features and Models

Each clip is down-sampled to 22.05 kHz and then divided into 46 ms frames using hanning window with a hop size of 14 ms. 80-dimensional log-mel filter banks extracted across a frequency range from 50 to 11 kHz are used as input features for both the baseline and our method. The official "Area Under

the Curve (AUC) of Receiver Operating Characteristic curve (ROC)" [30] is used to evaluate the system performances.

The official baseline of DCASE 2018 BAD challenge [23] is taken as our baseline. Its also a CNN-based encoder-classifier structure. The CNN-based encoder is the same as our discriminator as shown in block (b) of Fig.1. This encoder is then followed by three dense (fully-connected) layers with 256, 32 and 1 unit(s) as a binary classifier. Each convolution and dense layer use the leaky rectifier nonlinearity as their activation function except for the sigmoid output layer.

Different from baseline, our model Fig.1 (c) uses two same structure CNN-based encoders as a parallel discriminator, but the followed frame-wise classifier only has two dense layers with 32 and 1 unit(s). Besides using the frame-wise classifier, as the baseline model, we also investigate to use the conventional dense layers as the classifier (F-C) to learn the directly flattened discriminator outputs $\tilde{\mathcal{M}}$. As the flattened vector dimension is too large (2816) than that in the baseline, a four dense layers (512, 256, 32 and 1 unit(s)) instead of three is used in the F-C to achieve a better results. The Adam Optimizer [31] with a learning rate of 10^{-4} is used for both the baseline and one-stage training. The two-stage training uses 10^{-3} as initial learning rate, and then gradually decaying to 10^{-5} in `stage 1`, then fixed to 10^{-4} in `stage 2`. 200 epochs are used in each stage.

IV. Results

A. Results with one-stage training

Table I shows the performance comparison using one-stage training. Two training-test tasks are constructed to evaluate the effectiveness of the proposed methods. One is using "BirdVox-20k" as the training set while the other is using "warblrb10k" to train the model. Both of them are tested on the same three subset of "BirdVox-20k, warblrb10k and ff1010bird", there is no clip overlap between training and test data.

From the baseline results of Table I, it's clear that there are big performance gaps between in-domain and cross-domain tasks. Results on the in-domain test sets are much better than those on cross-domain test sets. By comparing the F-C and baseline results, we see significant AUC improvements on the cross-domain test tasks, such as when we train the model on "BirdVox-20k", there are relative 11.8% and 5.7% improvements on the "warblrb10k" and "ff1010bird" respectively. In the 2nd block of Table I, we only achieve limited gains (relative 2.2% on "BirdVox-20k") or even a little bit worse (relative 2.6% on "warblrb10k") results when the model is trained on a very wide-domain acoustic coverage dataset "warblrb10k". These improvements indicate that the proposed CNN-based discriminator is very effective to enhance the cross-domain performances when the model is trained on "BirdVox-20k" that with no richness background acoustics. Because: 1) both the baseline and F-C system are with the same type of classifiers; 2) as shown in section III-A, the "warblrb10k" is very diverse that contains a rich acoustic environment while "BirdVox-20k" is recorded from a fixed place with remote monitoring units.

Interestingly, by comparing the results in last two columns of Table I, we see that under one-stage training strategy, almost no improvements can be found when the model is trained on "BirdVox-20k" , however, the proposed frame-wise classifier achieves around absolute 2.6-3.5% AUC improvements over the F-C on the cross-domain test sets when the model is trained on "warblrb10k". In addition, it's clear to see that there is almost no performance change on the in-domain test set results, either for the "BirdVox-20k" or "warblrb10k" in-domain tasks, it indicates that both of the proposed CNN-based discriminative training and the frame-wise classifier are effective to improve the cross-domain BAD performances without worsening any in-domain performances.

B. Results with two-stage training

Table II shows the results of the proposed method with two-stage training using different CNN-based discriminator initialization. Comparing the results of TS-GAP with TS-fla, we see that pre-training the discriminators using L_{pre} with GAP achieves much better results than using flatten operation. Performances of systems using L_{pre} with flatten as initialization are even worse than the ones from one-stage training strategy. This may due to the fact that global average pooling as a structural regularizer sums out the spatial information, which is less prone to overfitting than traditional flatten operation [29]. Furthermore, when comparing the TS-GAP with F-W, it's clear that the performances from two-stage training is slightly better than the ones from one-stage training on all in-domain and cross-domain tasks. However, the gains shown in the 1st block of Table II are much larger than the ones shown in the 2nd block. This phenomenon

TABLE I
Results (AUC%) of the proposed model with F-C or frame-wise (F-W) classifier using one-stage training strategy. $\lambda = 0.1$ and 1.0 in (4) achieve the best results for both F-C and F-W that shown in the 1st and 2nd blocks respectively.

Train set	Test set	Baseline	F-C	F-W
BirdVox-20k	BirdVox-20k	94.62	94.57	93.63
	warblrb10k	62.57	69.98	68.96
	ff1010bird	75.11	79.42	79.61
warblrb10k	warblrb10k	94.29	94.39	94.66
	BirdVox-20k	64.33	65.74	68.37
	ff1010bird	85.22	82.98	86.47

TABLE II
Comparison (in AUC%) of training strategy with and without discriminator initialization. F-W is the one-stage training, TS-fla and TS-GAP represent two-stage training strategy using the L_{pre} with flatten and GAP operation respectively. $\lambda = 0.3$ and 0.1 in (4) achieve the best results for both TS-fla and TS-GAP that shown in the 1st and 2nd blocks respectively.

Train set	Test set	F-W	TS-fla	TS-GAP
BirdVox-20k	BirdVox-20k	93.63	93.54	94.29
	warblrb10k	68.96	66.28	70.13
	ff1010bird	79.61	79.13	82.52
warblrb10k	warblrb10k	94.66	94.23	94.86
	BirdVox-20k	68.37	63.64	68.61
	ff1010bird	86.47	86.48	86.73

is consistent with the observation from Table I. Finally, by comparing the AUCs of the TS-GAP and the baseline, our proposed method can bring relative 12.1%, 9.9% and 6.7%, 1.8% AUC improvements over baseline on the "BirdVox-20k" and "warblrb10k" based cross-domain tasks, respectively. These gains also indicate that the proposed discriminative training is more effective when there is large background domain mismatch between training and test data.

C. Visualization

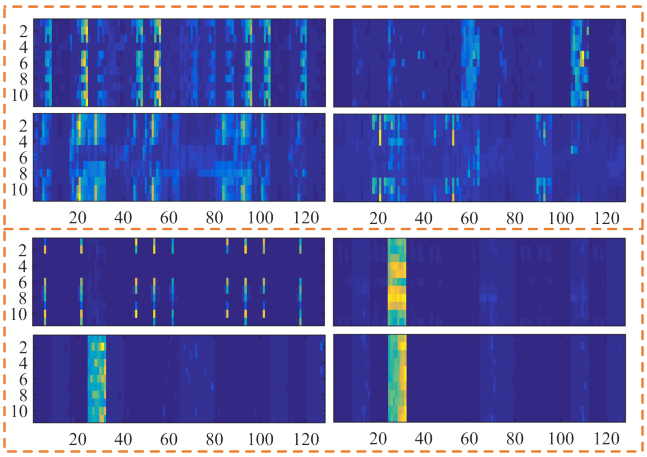


Fig. 2. Visualization of four pairs of discriminative acoustic representations. Each row is corresponding to one audio clip. The top part of the box shows the examples when the model trained and test both on warblrb10k. Bottom part is the examples when model trained on BirdVox-20k and test on ff1010bird.

In Fig.2, we visualize four audio samples' acoustic representations of the final layer of each discriminator. The vertical axis represents the frame index, and the horizontal axis represents the frequency index of all stacked channels. The upper and lower parts within the dashed box respectively show the representations on in-domain and cross-domain testing. The 1st and 3rd rows indicate the audio representations with bird calls while the 2nd and 4th rows refer to the ones without bird calls. It can be observed that each pair in the 1st and 3rd rows are very different. Each pair in the 2nd and 4th have something in common which represent the background sounds. According to this visualization, we can conclude that the parallel discriminator is able to produce the discriminative representations as we expect.

V. Conclusion

This paper investigates a new CNN-based architecture for acoustic event detection task to alleviate the domain mismatch problem, which features two CNN discriminators and an additional discriminative loss. Moreover, we design two kinds of training strategy and two alternative binary classifiers to further improve the performances. Results on DCASE2018 task3 dataset have shown that our two-stage training strategy with frame-wise classifier significantly outperforms the baseline system in most cross-domain evaluation cases.

Acknowledgment

Thanks to DCASE 2018 Challenge for providing the datasets. Yanhua Long is the corresponding author and the work is supported by the National Natural Science Foundation of China (No.62071302).

References

[1] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, October 2015.

[2] D. Stowell, Y. Stylianou, M. Wood, H. Pamuła, and H. Glotin, "Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge," *Methods in Ecology and Evolution*, vol. 10, pp. 2672–2680, March 2019.

[3] W. Wei, H. Zhu, E. Benetos, and Y. Wang, "A-crnn: A domain adaptation model for sound event detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 276–280.

[4] E. Fonseca, M. Plakal, F. Font, D. P. Ellis, and X. Serra, "Audio tagging with noisy labels and minimal supervision," in *Acoustic Scenes and Events 2019 Workshop (DCASE2019),New York (USA)*, October 2019, pp. 69–73.

[5] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1-2, pp. 151–175, 2010.

[6] V. Lostanlen, J. Salamon, A. Farnsworth, S. Kelling, and J. P. Bello, "Robust sound event detection in bioacoustic sensor networks," *PloS one*, vol. 14, no. 10, p. e0214168, 2019.

[7] F. Berger, W. Freillinger, P. Primus, and W. Reisinger, "Bird audio detection-dcase 2018," DCASE2018 Challenge, Tech. Rep., June 2018.

[8] J. Shen, Y. Qu, W. Zhang, and Y. Yu, "Wasserstein distance guided representation learning for domain adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 32, no. 1, 2018.

[9] S. Liaqat, N. Bozorg, N. Jose, P. Conrey, A. Tamasi, and M. T. Johnson, "Domain tuning methods for bird audio detection," in *Acoustic Scenes and Events 2018 Workshop (DCASE2018), Surrey (UK)*, November 2018, pp. 163–167.

[10] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 30, no. 1, 2016.

[11] T. Asami, R. Masumura, Y. Yamaguchi, H. Masataki, and Y. Aono, "Domain adaptation of dnn acoustic models using knowledge distillation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5185–5189.

[12] S. Mun and S. Shon, "Domain mismatch robust acoustic scene classification using channel information conversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 845–849.

[13] V. Hubeika, L. Burget, P. Matějka, and P. Schwarz, "Discriminative training and channel compensation for acoustic language recognition," in *Ninth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2008, pp. 301–304.

[14] J. Rohdin, T. Stafylakis, A. Silnova, H. Zeinali, L. Burget, and O. Plchot, "Speaker verification using end-to-end adversarial language adaptation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6006–6010.

[15] R. Duroselle, D. Jouvet, and I. Illina, "Metric learning loss functions to reduce domain mismatch in the x-vector space for language recognition," in *Twenty-first Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020, pp. 447–451.

[16] S. Mirsamadi and J. H. Hansen, "On multi-domain training and adaptation of end-to-end rnn acoustic models for distant speech recognition." in *Eighteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 404–408.

[17] M. Kośmider, "Calibrating neural networks for secondary recording devices," DCASE2019 Challenge, Tech. Rep., June 2019.

[18] H. Hu, S. M. Siniscalchi, Y. Wang, and C.-H. Lee, "Relational teacher student learning with neural label embedding for device adaptation in acoustic scene classification," in *Twenty-first Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020, pp. 1201–1205.

[19] S. Gharib, K. Drossos, E. Cakir, D. Serdyuk, and T. Virtanen, "Unsupervised adversarial domain adaptation for acoustic scene classification," in *Acoustic Scenes and Events 2018 Workshop (DCASE2018), Surrey (UK)*, November 2018, pp. 138–142.

[20] [Online]. Available: http://github.com/DCASE-REPO/bulbul_bird_detection_dcase2018

[21] D. Chang, Y. Ding, J. Xie, A. K. Bhunia, X. Li, Z. Ma, M. Wu, J. Guo, and Y.-Z. Song, "The devil is in the channels: Mutual-channel loss for fine-grained image classification," *IEEE Transactions on Image Processing*, vol. 29, pp. 4683–4695, 2020.

[22] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS)*, 2016, pp. 343–351.

[23] T. Grill and J. Schlüter, "Two convolutional neural networks for bird detection in audio signals," in *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 1764–1768.

[24] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 121–125.

[25] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "Audio set classification with attention model: A probabilistic perspective," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 316–320.

[26] 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.

[27] M. Lasseck, "Acoustic bird detection with deep convolutional neural networks," in *Acoustic Scenes and Events 2018 Workshop (DCASE2018), Surrey (UK)*, November 2018, pp. 143–147.

[28] I. Himawan, M. Towsey, and P. Roe, "3d convolution recurrent neural networks for bird sound detection," in *Acoustic Scenes and Events 2018 Workshop (DCASE2018), Surrey (UK)*, November 2018, pp. 1–4.

[29] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[30] [Online]. Available: http://dcase.community/challenge2018/task-bird-audio-detection

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