Robust Acoustic Scene Classification to Multiple Devices Using Maximum Classifier Discrepancy and Knowledge Distillation

Saori Takeyama, *LINE corporation, Tokyo Institute of Technology*
Tatsuya Komatsu, *LINE corporation*
Koichi Miyazaki, *LINE corporation, Nagoya University*
Masahito Togami, *LINE corporation*
Shunsuke Ono, *Tokyo Institute of Technology*

***Abstract*—This paper proposes robust acoustic scene classification (ASC) to multiple devices using maximum classifier discrepancy (MCD) and knowledge distillation (KD). The proposed method employs domain adaptation to train multiple ASC models dedicated to each device and combines these multiple device-specific models using a KD technique into a multi-domain ASC model. For domain adaptation, the proposed method utilizes MCD to align class distributions that conventional DA for ASC methods have ignored. The multi-device robust ASC model is obtained by KD, combining the multiple device-specific ASC models by MCD that may have a lower performance for non-target devices. Our experiments show that the proposed MCD-based device-specific model improved ASC accuracy by at most 12.22% for target samples, and the proposed KD-based device-general model improved ASC accuracy by 2.13% on average for all devices.**

***Index Terms*—acoustic scene classification, domain adaptation, maximum classifier discrepancy, convolutional neural network, knowledge distillation**

I. Introduction

Acoustic scene classification (ASC) is one of the ongoing research subjects, which classifies input audio data into pre-defined scene classes, such as *office, train station, and airport* [1], and expected to be applied to various fields, e.g., robotic navigation [2], context-aware services [3], and surveillance systems [4]. For those applications that need to identify the environment in a variety of situations, the recording devices are diverse, such as surveillance cameras, smartphones, smart speakers, and embedded microphones. Since each of these devices has its own encoding method and microphone characteristics, audio data from the same sound source have different characteristics depending on the devices. For real-world ASC applications, ASCs that support all device characteristics in advance are impractical. Robustness to sounds recorded by various recording devices including unknown devices is very important.

ASC is a classification task which associate semantic labels with audio recordings (for example labeling an environment as "In a bus" or "In a meeting"). Recently, methods based on image recognition techniques that use the spectrograms as image features have become a popular approach. These methods use convolutional neural networks (CNNs) to perform the classification [5]–[7]. The CNN-based approaches have shown its effectiveness for ASC and become a strong baseline. Although the CNN-based method has shown its effectiveness, they have a problem for ASC with audio data recorded by multiple devices.

Spectral structures of recorded audio vary depending on the characteristics of the device, such as encoding and microphones. As a consequence, the sounds in the same acoustic scenes recorded by different devices have different spectral structures. This is a critical issue for CNNs that deal with spectra as images and leads to degradation of classification performance. Furthermore, collecting a huge amount of training data on multiple devices for training is unrealistic. There is a need for a method of efficiently training for utilizing audio data recorded by multiple devices with different characteristics.

To tackle this problem, domain adaptation (DA) [8] plays an important role. DA is a method for obtaining common features among multiple domains with different characteristics. In many cases, DA is employed for efficient training of data from a data-poor domain (target domain) using knowledge from a data-rich domain (source domain). Some DA methods for ASC have been proposed [9], [10] and proven that their DA-based methods are more effective than simply use all domain data as one data set. [10] employs an additional constraint for distributions of extracted features to match the distribution of both source and target device. As a consequence, the extracted features have the same distribution regardless of the difference of the recording devices.

These DA methods deal with the distribution of the dataset as a single distribution, do not consider the distribution of each class in the dataset. There is no guarantee that the distribution of each class can be matched properly. This is because most DA methods imply that each class can be properly identified in the source domain. Under this assumption, the distribution of the common feature can be appropriately obtained by matching the target domain with the source domain where each class can appropriately be classified. However, the assumption that each class can be classified with 100% accuracy in the source domain is not realistic. In reality, even if the distributions of the domains are matched, the distribution of each class may be mixed, which leads to a degradation in the classification performance of either the target or the source domain, i.e., a trade-off. With this trade-off, a multi-domain robust ASC cannot be realized. Therefore, in order to realize a multi-

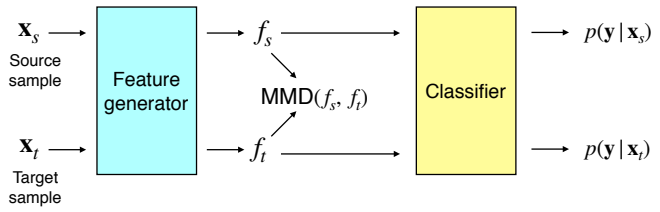

Fig. 1. Workflow of the DA method based on MMD

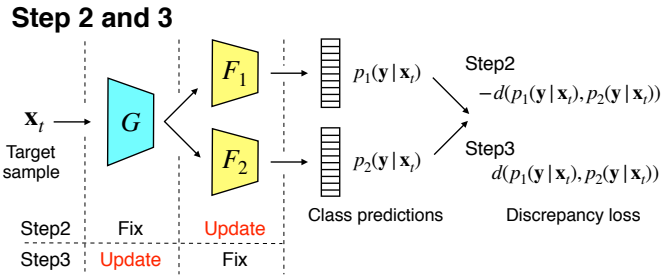

Fig. 2. Adversal training steps of MCD

domain robust ASC, there is a need for a way to align the class distributions in multiple datasets properly and to handle the knowledge of each domain properly.

We propose a new method for multi-domain robust ASC using a combination of DA and a knowledge distillation (KD) [11]. For DA in the proposed method, we utilize the state-of-the-art DA in the computer vision field, maximum classifier discrepancy (MCD) [12], which can properly consider distributions of each class within domains. By using MCD, the distribution of classes within a domain is properly aligned. However, since the trade-off problem still remains, models obtained by MCD is still device-specific. So, the proposed method newly introduces KD to utilize multi-domain knowledge. First, the proposed method uses MCD for each device to train multiple device-specific models, aligning distributions of each class. Next, a multi-device robust model is obtained by KD using the device-specific models as multiple teacher-models. The proposed method treats the class distributions within each domain by MCD, and device-specific knowledge is effectively employed by KD to obtain the multi-device robust model.

II. Domain Adaptation For ASC

The basic idea of domain adaptation methods for ASC is to extract domain-invariant scene-features by removing domain information. The DA neural network (DANN) [9] is the most basic DA method and consists of a scene feature generator and two classifiers; a scene classifier and a domain classifier that share the scene-feature as their input. The feature generator is trained to minimize the scene classifier loss and maximize the domain classifier loss. The obtained scene feature cannot classify domain of the input audio, i.e., DANN can remove the domain-wise characteristics from scene features.

Some DA method exploits an additional constraint of a distance between distributions of features from different domains to minimize the domain difference. The most popular choice of the distance measure is a maximum mean discrepancy (MMD) [10], which is a matching method of all orders of statistics between sets of samples based on a kernel method. Fig. 1 shows a network for DA using MMD, where $\mathbf{x}_s$ and $\mathbf{x}_t$ are input audio clips from source and target domains, respectively, and f_s and f_t are output of the feature generator for input $\mathbf{x}_s$ and $\mathbf{x}_t$, respectively. By minimizing the MMD of f_s and f_t, since the source and the target distribution are brought closer, the features from different domains are obtained from

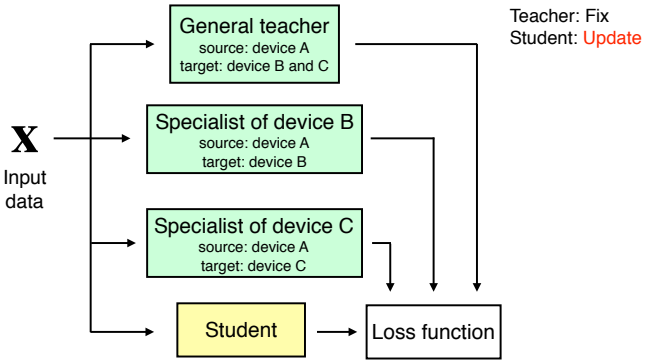

Fig. 3. The proposed KD architecture

the same distributions regardless of the difference of domains. In the Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 competition [13], ASC methods using MMD-based DA have already been proposed [14], [15].

Although these DA methods have shown effectiveness, these DA methods deal with the distribution of the dataset as a single distribution, do not consider the distribution of each class in the dataset. Since the distribution for each class in each domain is not aligned, there is no guarantee that the obtained scene features are distributed for each class. Therefore, there is a possibility that an appropriate classification boundary common to the domains cannot be obtained, and the improvement of the classification accuracy in the target domain is not guaranteed. A method that takes into account the alignment of each class is required.

III. Proposed Method

We propose a domain-robust ASC method using MCD and KD. The proposed method adopts MCD to train device-specific classification models and combines the device-specific models using the KD technique into a multi-domain classification model.

A. Maximum Classifier Discrepancy

MCD is a state-of-the-art method for DA, and is proposed in [12]. The method adjusts the distribution of source and target by exploiting the decision boundary between the two classifiers, leading to domain-robust classification. Specifically, the

TABLE I
ARCHITECTURE OF BASELINE METHOD IN DCASE 2019

Layer	Output shape	Kernel size
Input	$1 \times 40 \times 500$	-
Conv2d1	$32 \times 40 \times 500$	7×7
BatchNorm2d1	$32 \times 40 \times 500$	-
ReLU1	$32 \times 40 \times 500$	-
MaxPool2d1	$32 \times 8 \times 100$	5×5
Dropout1 (30%)	$32 \times 8 \times 100$	-
Conv2d2	$64 \times 8 \times 100$	7×7
BatchNorm2d2	$64 \times 8 \times 100$	-
ReLU2	$64 \times 8 \times 100$	-
MaxPool2d2	$64 \times 2 \times 1$	4×100
Dropout2 (30%)	$64 \times 2 \times 1$	-
Flatten	128	-
Dense1	100	128×100
Dropout3 (30%)	100	-
Dense2	10	100×10

TABLE II
HYPER-PARAMETERS FOR MEL-ENERGY FEATURE EXTRACTION

window length	40 ms
window type	hamming asymmetric
hop length	20 ms
number of FFT bins	2048
number of mels	40

MCD network is composed of a feature generator G and two classifiers F_1 and F_2 and trains G, F_1, and F_2 by repeating the three steps (Step 1, 2, and 3). In Step 1, G, F_1, and F_2 are trained using only source samples as input so that the MCD network can correctly classify the source one. Fig. 2 is a learning flow diagram of Step 2 and 3. In Step 2, the generator is fixed, and only two classifiers are updated. By training the classifiers to increase the discrepancy, they can detect the target samples excluded by the support of the source. Here, $p_1(\mathbf{y}|\mathbf{x}_t)$ and $p_2(\mathbf{y}|\mathbf{x}_t)$ are predictions of F_1 and F_2 for $\mathbf{x}_t$, respectively, and the discrepancy is as follows:

$$d(p_1, p_2) := \frac{1}{K} \sum_{k=1}^{K} |p_{1k} - p_{2k}|, \tag{1}$$

where p_{1k} and p_{2k} are probability output of p_1 and p_2 for class k, respectively. The discrepancy evaluates ℓ_1 norm of difference probabilities classified class k. For this step, MCD simultaneously maximizes the discrepancy between F_1 and F_2 and minimizes the categorical cross-entropy loss for the source sample. As a result, the decision boundary of F_1 and F_2 can be made different while ensuring the classification accuracy of the source sample. In contrast, in Step 3, F_1 and F_2 are fixed, and G is learned for m times so that the discrepancy of the two classifiers for the target samples is minimized. This allows you to move the target domain in the region where the predictions of the two classifiers do not match into that where the predictions match. In [12], MCD has achieved high classification accuracy for image classification and segmentation, but a method applied to ASC has not been proposed yet. Since MCD uses task-specific information for training classifiers, we expect MCD to help to build a device-specific model.

B. ASC Method with MCD and KD

To build a multiple-device-robust model, we employ KD to combine device-specific models obtained with adapting for the individual devices by MCD. KD is a technique that trains a new model (student model) using both a pre-trained model (teacher model) and training data, and some conventional methods utilizes the KD techniques for the ASC tasks [17], [18]. The teacher model has a similarity between the feature domains of each class in the output of the hidden layer before the softmax processing (logit), and the student model can utilize the similarity by distillation. The loss function of KD is defined by

$$\begin{aligned}\mathcal{L}_{\mathrm{KD}} :=& (1-\lambda)\mathcal{L}_{\mathrm{soft}}(y_{\mathrm{soft}}^{\mathrm{teacher}}, y_{\mathrm{soft}}^{\mathrm{student}}) \\ &+ \lambda T^2 \mathcal{L}_{\mathrm{hard}}(y_{\mathrm{true}}, y_{\mathrm{hard}}^{\mathrm{teacher}}),\end{aligned}$$

where $\mathcal{L}_{\mathrm{soft}}$ and $\mathcal{L}_{\mathrm{hard}}$ are categorical cross-entropy loss, $y_{\mathrm{soft}}^{\mathrm{teacher}}$ and $y_{\mathrm{soft}}^{\mathrm{student}}$ are outputs of softmax with temperature T of teacher and student model, respectively, y_{true} is a ground-truth label, and $y_{\mathrm{hard}}^{\mathrm{teacher}}$ is the output of the teacher model. The parameter λ balances the weight between $\mathcal{L}_{\mathrm{soft}}$ and $\mathcal{L}_{\mathrm{hard}}$, and the student model can be effectively trained by suitable λ.

Multiple device-specific classification models are employed as teacher models: a general teacher and specialists of each device of target samples. All models are trained by using MCD. The specialists use the audio data of each device as target samples. In Step 3 of MCD, the models minimize only the discrepancy because MCD assumes the situation no having target labels. However, since we can utilize target labels in this work, models minimize both discrepancy and categorical cross-entropy loss for effective learning.

The training flow of the proposed method is shown in Fig 3. The proposed method prepares three teacher models and averages of their logit. For this setting, the student could learn features that are more robust to the device than each teacher.

IV. EXPERIMENTS

We demonstrate the performance of the proposed method on Task 1.B of DCASE 2019, which is to classify audio data collected by three devices into 10 acoustic scene classes, i.e., $k = 10$, where their devices consist of a high-performance audio recorder A and two mobile devices, B and C. We used the data of device A as source samples and that of device B and C as target samples. To evaluate the proposed method using the combination of MCD and KD, we have conducted two experiments: (i) Evaluation of the device-specific model using MCD (ii) Evaluation of the robustness of the multi-device model using MCD and KD.

A. Experimental Settings

In this section, we describe experimental settings. We employ a two layer CNNs as a base network of our proposed method, whose architecture is shown as Tab. I. The inputs

TABLE III
DATASETS OF DCASE 2019 TASK 1-B

	device A (high-performance audio recorder)	device B (mobile device)	device C (mobile device)
training data	6120	415	415
validation data	3065	125	125
test data	4185	540	540

TABLE IV
HYPER-PARAMETERS FOR TRAINING MODEL

	baseline	baseline w/ MMD	baseline w/ MCD
epoch	200		
batch size	64	83	83
optimizer	Adam [16]		
learning rate	0.001		

of the baseline method are log mel-band energies. Tab. II shows setting for the log mel-band energies. Tab. III shows the number of training, validation, and test data in the experiments, where we randomly selected 30% of the train data of Task1.B as validation data.

For all ASC method based on DA, the feature generator consists of two CNN layers without batch normalization layers in Tab. I, i.e., from the Conv2d1 layer to Dropout2 layer, and the other layers are the classifier. In the paper, we adopt MMD as compared DA method, and the parameters of the compared and proposed method were set like Tab. IV. In MCD method, we set the number of the generator update $m = 2$.

In the second experiment, we set the parameters $\lambda = 0.5$ and $T = 0.01$ in the loss function of KD, and the training data of the general teacher is the audio data of all devices.

B. Comparison between MMD and MCD

We demonstrate the performance of the specialists of devices B and C using MCD. In this experiment, we compared the specialists using MCD with that using MMD. Fig. 4 shows the ASC accuracy by each method for the test samples. In Fig. 4, one can see that MCD outperforms the other methods for target samples, and in the case of the specialist of device C, the ASC accuracy of unknown device data is also improved. On the other hand, in the case of the specialist of B, MCD reduces the ASC accuracy for unknown device compared with MMD, because device A and C would have similar characteristics. The results show that MCD achieves high DA performance. However, MCD cannot successfully classify source samples compared with baseline, so one can see that there is a trade-off between source and target domains. Therefore, only using MCD is incomplete to achieve device-robust classification.

C. Knowledge Distillation

We inspect the performance of the proposed device-specific ASC method using MCD and KD compared with ASC methods using MMD and KD. We trained the general teacher and

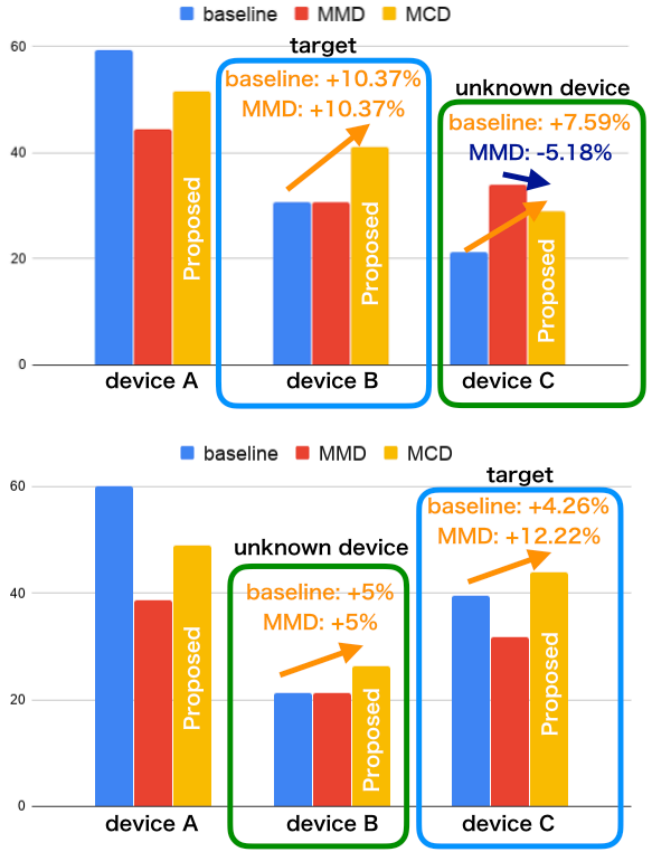

Fig. 4. ASC accuracy on the specialist of device B (top) and the specialist of device C (bottom). This figure shows that MCD outperforms the other methods regarding the target samples but reduces the ASC accuracy of the source samples.

the specialist of B and C using the audio data collected by all devices, devices A and B, and devices A and C, respectively.

Fig. 5 shows the ASC accuracy of the test samples by the ASC methods with DA (top) and the ASC methods with DA and KD (bottom). In the top of Tab. 5, one can see that the method only using MCD improves ASC accuracy for the target samples, but that of the source samples drops because of the trade-off between source and target domains. The bottom of Tab. 5 shows that the proposed KD-MCD method can achieve higher ASC accuracy for the source and target samples. Therefore, the proposed KD-MCD can successfully extract robust features to the device without trade-off between source and a target domain and achieve to device-robust classification.

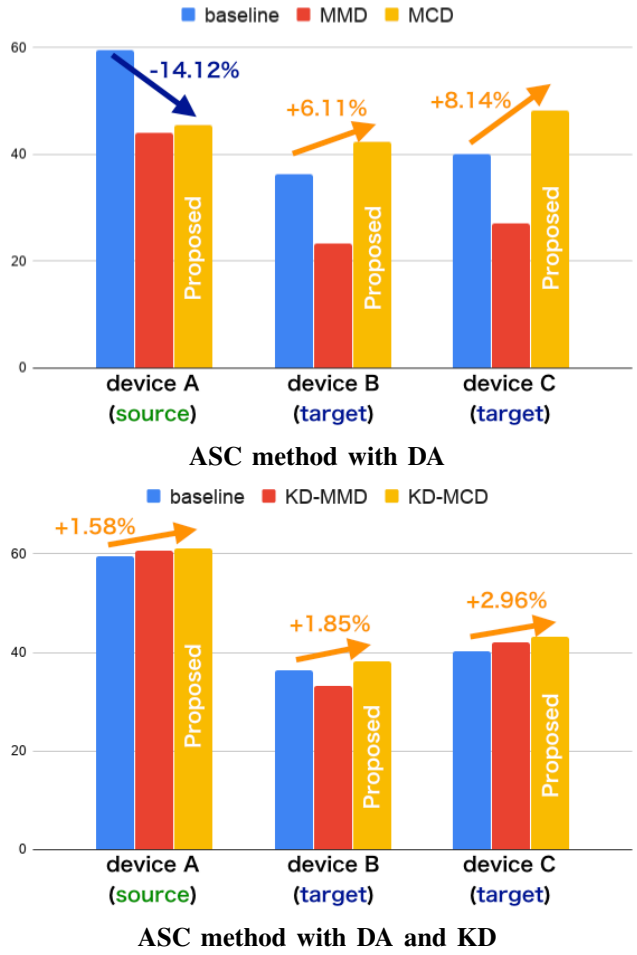

Fig. 5. ASC accuracy on DA methods (top) and KD methods (bottom). One can see that the DA methods reduces the accuracy of the source samples, but the KD methods using MCD improve it for all devices.

V. Conclusion

We proposed a new robust ASC method to multiple devices, which trains device-specific classification models using MCD and combines them by KD. MCD is one of the state-of-the-art DA methods, which can move the target domain in the source domain using a discrepancy between the output of two classifiers. However, since the source sample cannot be classified correctly, MCD faces a trade-off between the classification accuracy of the source and target samples. In order to resolve the problem, our proposed method adopts KD with one general classification model and two device-specific classification models. Therefore, the proposed method can effectively classify the audio data of all devices than the method only using DA. In the experiments, the proposed method successfully classifies for all devices. In the future, we will try to use other powerful networks as the base.

References

[1] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, 2018, pp. 9–13.

[2] S. Chu, S. Narayanan, C-C. J. Kuo, and M. J. Mataric, "Where am i? scene recognition for mobile robots using audio features," in *IEEE International Conference on Multimedia and Expo(ICME)*, 2006, pp. 885–888.

[3] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 14, no. 1, pp. 321–329, 2005.

[4] R. Radhakrishnan, A. Divakaran, and A. Smaragdis, "Audio analysis for surveillance applications," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2005, pp. 158–161.

[5] Y. Han, J. Park, and K. Lee, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," *the Detection and Classification of Acoustic Scenes and Events (DCASE)*, pp. 1–5, 2017.

[6] S. Mun, S. Shon, W. Kim, D. K. Han, and H. Ko, "Deep neural network based learning and transferring mid-level audio features for acoustic scene classification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2017, pp. 796–800.

[7] H. Zeinali, L. Burget, and H. Cernocky, "Convolutional neural networks and x-vector embedding for dcase2018 acoustic scene classification challenge," Tech. Rep., DCASE2018 Challenge, 2018.

[8] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.

[9] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," *arXiv preprint arXiv:1409.7495*, 2014.

[10] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. Mar, pp. 723–773, 2012.

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

[12] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 3723–3732.

[13] "DCASE2019," http://dcase.community/challenge2019/.

[14] M. Wang and R. Wang, "Ciaic-ASC system for DCASE 2019 challenge task1," Tech. Rep., DCASE2019 Challenge, June 2019.

[15] K. Koutini, H. Eghbal-zadeh, and G. Widmer, "Cp-jku submissions to dcase'19: Acoustic scene classification and audio tagging with receptive-field-regularized cnns," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 2019.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations (ICLR)*, 2015.

[17] J. Jung, HS. Heo, H. Shim, and HJ. Yu, "Distilling the knowledge of specialist deep neural networks in acoustic scene classification," 2019.

[18] L. Gao, H. Mi, B. Zhu, D. Feng, Y. Li, and Y. Peng, "An adversarial feature distillation method for audio classification," *IEEE Access*, vol. 7, pp. 105319–105330, 2019.