TOWARDS AUDIO DOMAIN ADAPTATION FOR ACOUSTIC SCENE CLASSIFICATION USING DISENTANGLEMENT LEARNING

Jakob Abeßer[1] *Meinard Müller*[2]

[1] Semantic Music Technologies Group, Fraunhofer IDMT, Ilmenau, Germany
[2] International Audio Laboratories Erlangen, Germany

ABSTRACT

The deployment of machine listening algorithms in real-life applications is often impeded by a domain shift caused for instance by different microphone characteristics. In this paper, we propose a novel domain adaptation strategy based on disentanglement learning. The goal is to disentangle task-specific and domain-specific characteristics in the analyzed audio recordings. In particular, we combine two strategies: First, we apply different binary masks to internal embedding representations and, second, we suggest a novel combination of categorical cross-entropy and variance-based losses. Our results confirm the disentanglement of both tasks on an embedding level but show only minor improvement in the acoustic scene classification performance, when training data from both domains can be used. As a second finding, we can confirm the effectiveness of a state-of-the-art unsupervised domain adaptation strategy, which performs across-domain adaptation on a feature-level instead.

Index Terms— acoustic scene classification, domain adaptation, disentanglement learning

1. INTRODUCTION

Acoustic scene classification (ASC) is an essential task of auditory scene analysis. Its goal is to classify the audio recording's environment according to a given set of pre-defined classes. Examples of such classes are indoor locations (e. g., restaurants and metro stations) and outdoor locations (e. g., pedestrian streets and parks). State-of-the-art ASC algorithms are mostly driven by deep learning (DL) techniques and, to this day, face several challenges in real-world application scenarios, see [1] for a summary.

A key challenge is that the data used to train ASC algorithms is typically recorded under different acoustic conditions (microphone characteristics, reverberations, background noises) than the audio data recorded by the potential application device. Such a microphone mismatch can cause a "domain shift," which is a distribution mismatch between the source domain (training) data and the target domain (test) data to be expected in the application scenario. It has been shown in related disciplines like computer vision that such domain shifts can cause significant performance degradation of (deep) neural networks [2, 3]. Several domain adaptation (DA) strategies have been proposed in the past as countermeasures. Amongst others, these strategies include a fine-tuning of pre-trained classification models onto target domain data (transfer learning) or an adaptation/normalization of the input features without changing the initial model. From a practical viewpoint, both strategies require an additional adaptation effort every time an ASC system faces audio data from a novel target domain.

As the main contribution of this paper, we propose to use disentanglement representation learning during the model training process to learn task-specific yet domain-agnostic feature representations. Our assumption is that there exist two main factors which influence the acoustic variability of multi-device acoustic scene recordings. First, there are domain-specific properties such as microphone characteristics and room acoustics, which do not carry meaningful information but rather confuse ASC algorithms. The second factor are the relevant characteristics of different acoustic scenes. By learning domain-agnostic feature representations, our objective is to develop more robust and context-sensitive ASC approaches, which can be used for devices such as hearing aids and cellphones to adapt to changes in the surrounding acoustic environment automatically. In order to foster scientific reproducibility, we publish Python code to reproduce our experiments.[1]

2. RELATED WORK

Modern ASC approaches almost exclusively rely on use convolutional neural networks (CNN) or convolutional recurrent neural networks (CRNN) to learn discriminative features in time-frequency audio representations such as Mel spectrograms. We refer the reader to [1] for a survey on state-of-the-art ASC approaches. Several publications proposed DA methods for audio signals to compensate for domain shift caused, for instance, by microphone mismatch conditions.

DA methods can be categorized into unsupervised and supervised methods depending on whether labels of the target domain data exist or not. In supervised DA scenarios, ASC algorithms can be trained jointly for ASC and domain classification (DC) in a multitask learning fashion [4, 5]. A similar multitask learning approach can also be used for semi-supervised learning, as shown in [6].

For unsupervised DA, adversarial training strategies were proposed to align the distributions of intermediate feature representations in ASC models obtained from source and target domain data [7, 8]. or by mapping data from both domains to another (universal) domain [9,10]. One approach to achieve such an alignment is to standardize data per frequency band independently per domain [11] or by matching the band-wise statistics between domains [12]. In general, such an adaptation can not only be performed on the feature-level but also using internal hidden layer activations [13].

This research was partially supported by H2020 EU project AI4Media—A European Excellence Centre for Media, Society and Democracy—under Grand Agreement 95191 and by the Fraunhofer Innovation Program "SEC Learn FLY". The authors would like to express gratitude to Alessandro Ilic Mezza, Sebastian Ribecky, and Sascha Grollmisch for valuable discussions. The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and the Fraunhofer Institute for Integrated Circuits IIS.

[1] https://github.com/jakobabesser/ICASSP_2022_DisentanglementASC

In order to compensate for differences in the microphone frequency responses, Kosmider [14] propose a "spectral correction" step. Based on time-aligned recordings from different recording devices (each one is considered as one domain), frequency-dependent magnitude coefficients are estimated from the source data and used to calibrate the target domain data. Mezza et al. [10] proposed to project source and target domain features to a lower-dimensional subspace spanned by the eigenvectors of the source domain feature covariance matrix. Hu et al. [15] propose to use neural label embedding (NLE) to encode structural relationships between different acoustic scene classes from source domain data. In order to mitigate microphone mismatch, this knowledge is then transferred to target domain data using relational teacher–student learning.

A promising approach to train more robust classification models is disentanglement representation learning. Mun and Shon [9] proposed to use a factorized hierarchical variational autoencoder (FHVAE) in order to first disentangle the audio features' component related to the recording device, shift it according to a universal domain mean value, and finally decode and reconstruct the modified audio features. In the field of music information retrieval, Lee et al. [16] used a Conditional Similarity Network (CSN) [17] to disentangle multiple semantic concepts such as genre, mood, or instrumentation in music recordings in order to learn configurable similarity metrics. The authors applied a binary masking procedure on internal embedding representations in order to enforce a specific distribution of semantic concepts within this embedding. We adapt this approach to disentangle the ASC and DC components as will be detailed in the next section. As the main conceptual difference to previous DA approaches, we aim to learn domain-agnostic feature representations already at the training stage, a strategy that does not require any further adaptation when novel target domains are faced.

3. PROPOSED METHOD

In this section, we first introduce in Section 3.1 the ASC dataset used for experimental validation. Then, we describe in Section 3.2 the applied neural network architecture, which implements a multitask learning approach of ASC and DC as well as an embedding masking procedure in order to disentangle both concepts. Finally, Section 3.3 summarizes different training configurations, covering regular ASC training as well as unsupervised and supervised DA scenarios.

3.1. Dataset & Feature Extraction

In this paper, we focus on the DCASE 2018 Task 1B entitled "Acoustic Scene Classification with mismatched recording devices", which addresses ASC under microphone mismatch conditions. The corresponding dataset includes ten second long excerpts of acoustic scene recordings (denoted as "clips" in the following) recorded with three devices: a source domain device (A) as well as two target domain devices (B & C). For the sake of reproducibility, we use in our experiments a set of pre-computed Mel spectrogram features extracted from the task's development set, which were published alongside with [7][2]. These features cover $F = 64$ Mel bands and were computed using a window size of 2048 samples with a 50 % overlap. This feature set was used in follow-up publications on unsupervised DA [8, 12]. Based on the provided dataset split, we use the training set with 5510/612/612 clips for devices A/B/C for model training and the test set with 2518/180/180 clips for evaluation. Throughout this paper, the subscript "A" indicates the acoustic scene and "D" the domain classification tasks. Furthermore, the superscripts "S" and "T" refer to the source and target domain, respectively. We consider $A = 10$ acoustic scene classes and $D = 3$ domain classes.

[2] https://zenodo.org/record/1401995

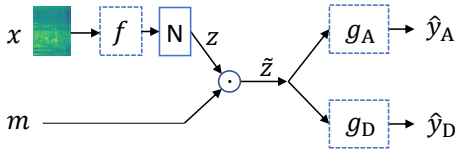

Fig. 1. General neural network architecture: A core model f transforms input features x to a normalized (N) embedding representation z, which is multiplied (element-wise) with a binary mask m. The masked embedding $\tilde{z}$ is processed by two output branches to obtain acoustic scene and domain predictions $\hat{y}_\mathrm{A}$ and $\hat{y}_\mathrm{D}$, respectively. Trainable and fixed parts of the network are shown with dashed and solid blue rectangles, respectively.

3.2. Embedding Normalization & Masking

Figure 1 gives an overview our network architecture. Each clip is represented by a feature tensor $x \in \mathbb{R}^{T\times F\times C}$ covering $T = 431$ time frames, $F = 64$ Mel bands, and one depth channel ($C = 1$). The output of the core model $f : \mathbb{R}^{T\times F\times C} \to \mathbb{R}^E$ is an embedding vector $z = f(x)$. Zhai et al. [18] discussed the importance of normalizing embedding vectors when being used for classification tasks. Therefore, we apply layer normalization [19] to normalize each embedding vector independently to zero mean and unit variance. After the normalization step, each embedding z is multiplied element-wise with a binary mask $m \in \{0, 1\}^E$ yielding $\tilde{z} = z \odot m$. Since we aim to disentangle both semantic concepts (acoustic scene and domain), we organize m such that its first half contains zeros and the second half contains ones or vice versa. The main intuition is that only the unmasked parts of the embedding vector can provide information for each of the two classification tasks. As a consequence, this task specialization should support the disentanglement [16].

The masked embedding vectors $\tilde{z}$ are propagated to two output branches each with one dense layer with softmax activation in order to predict the acoustic scene and the audio domain. The output branch operations are denoted as g_A and g_D and the acoustic scene and domain predictions $\hat{y}_\mathrm{A} \in \mathbb{R}^A$ and $\hat{y}_\mathrm{D} \in \mathbb{R}^D$ are computed as $\hat{y}_\mathrm{A} = g_\mathrm{A}(\tilde{z})$ and $\hat{y}_\mathrm{D} = g_\mathrm{D}(\tilde{z})$, respectively. Therefore, the model training involves learning f, g_A, and g_D from batches of data instances, which consist of feature-mask pairs (x, m) and the corresponding targets $(y_\mathrm{A}, y_\mathrm{D})$.

For the sake of comparability with prior research, we use the "Kaggle" CNN model, which is detailed in Table II in [12]. The network includes five convolutional layers followed by two dense layers each with 256 units. We use this model as core model f and consider the output of the second dense layer as embedding z.

3.3. Training Strategies & Loss Configurations

In our work, we compare different approaches to train the neural network for ASC alone or jointly for ASC and domain classification (DC). In particular, we investigate six training strategies as shown in Figure 2. First, `C0` involves a conventional ASC model training solely on the source domain data without considering the domain classification task (comparable to [12]). In `C1`, while following the same training objective, we use both source and target domain data in

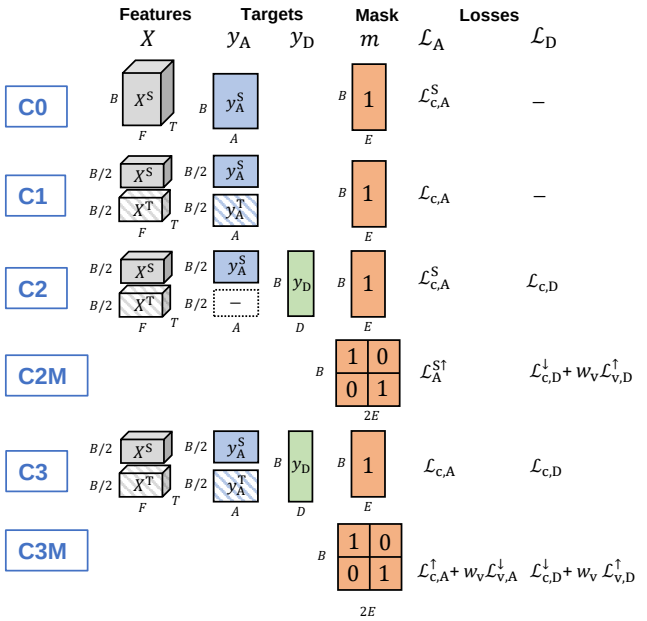

Fig. 2. Overview of different loss configurations. For each configuration, the composition of features, targets, and masks per training batch as well as the composition of the two losses $\mathcal{L}_A$ and $\mathcal{L}_D$ is shown (see Section 3.3 for more details). For the masks (orange), "1" and "0" indicate all-ones and all-zeroes blocks, respectively.

equal share. In `C2`, we consider a hybrid form between unsupervised and supervised DA, where ASC annotations are used only for source but not for target domain data (denoted "U*" in Table 4). Therefore, we can only evaluate the ASC task on the source domain data but at the same time can combine both source and target domain data for the additional DC task. In `C3`, a supervised DA scenario is simulated with ASC labels available for both source and target domain data.

The embedding masking procedure detailed in Section 3.2 is only applied in configurations `C2M` and `C3M`. Here, different parts of the embeddings in a training batch contribute to different loss terms. Since in `C2M` and `C3M` only the upper half of the embedding vector contributes to the ASC classification, we double in these cases the embedding size (i. e., using an embedding size of $2E$ instead of E) to remain comparable with the other configurations. The embedding masks m and the loss functions $\mathcal{L}_\mathrm{A}$ and $\mathcal{L}_\mathrm{D}$ (defined later) are designed in such way that the first half of the embedding vector supports the ASC task but remains agnostic for DC task and vice versa for the second half.

For each configuration, Figure 2 illustrates in a column-wise fashion the composition of the feature tensor[3] x, the acoustic scene target y_A, the domain target y_D, the embedding mask m, as well as the calculation of the loss function $\mathcal{L}$. The tensor dimensions are given in terms of the batch size B, the number of Mel bands F, the number of time frames T, and the embedding size E. For each batch, data instances and corresponding targets are randomly sampled from the relevant domains.

The total loss $\mathcal{L}$ is computed as weighted sum of the individual output branch losses $\mathcal{L}_\mathrm{A}$ and $\mathcal{L}_\mathrm{D}$ for both classification tasks:

$$\mathcal{L} = \mathcal{L}_\mathrm{A} + w_\mathrm{D}\mathcal{L}_\mathrm{D}. \tag{1}$$

As shown in the final two columns of Figure 2, the losses $\mathcal{L}_\mathrm{A}$ and $\mathcal{L}_\mathrm{D}$ are composed of different variations of the categorical cross-entropy loss $\mathcal{L}_\mathrm{c}$ and a variance-based loss $\mathcal{L}_\mathrm{v}$ defined in (2) and (3). Empirically, we found the two weighting factors $w_\mathrm{D} = 10$ and $w_\mathrm{v} = 500$ (compare `C2M` and `C3M`, Figure 2) to balance out the contribution of the individual loss terms in the weighted sums.

To define these losses, we consider a multi-class classification with N classes, a batch size of B, batch-wise network predictions $\hat{y} \in \mathbb{R}^{B\times N}$, and the corresponding one-hot encoded targets $y \in \mathbb{R}^{B\times N}$. The categorical cross-entropy loss can be expressed as

$$\mathcal{L}_\mathrm{c}(\hat{y}, y) = -\frac{1}{B}\sum_{b=1}^{B}\sum_{n=1}^{N} y_{b,n}\log\hat{y}_{b,n}. \tag{2}$$

This loss function guides the optimization of a neural network to improve its performance for a given classification task. Opposed to that, the variance-based loss

$$\mathcal{L}_\mathrm{v}(\hat{y}) = -\frac{1}{B}\sum_{b=1}^{B}\frac{1}{N}\sum_{n=1}^{N}(\hat{y}_{b,n} - \overline{\hat{\mu}_b})^2. \tag{3}$$

when applied on the networks softmax predictions maximizes the uncertainty for a given classification task. The average of all class predictions is denoted as $\overline{\hat{\mu}_b}$. Losses which are only computed using the first and second half of each mini batch are indicated by the superscripts $\uparrow$ and $\downarrow$, respectively.

4. EVALUATION

In our experiments, we use an embedding size of $E = 256$ and a batch size of $B = 512$. The Adam optimizer is used for training with an initial learning rate of 10^-3 for 400 epochs. The learning rate is halved every 50 epochs. In order to better evaluate the effectiveness of the proposed disentanglement approaches, we avoid any data augmentation and model regularization during training. We evaluate six models trained with the loss configurations introduced in Section 3.3. Furthermore, for the configuration `C0`, which does not involve any adaptation to the target domain, we investigate the effectiveness of the unsupervised band-wise adaptation proposed by Mezza et al. [12] (entitled `C0`[12]). Here, the target domain data in the test set is adapted to the source domain data in the training set.

In Table 4, the experimental results are reported as classification accuracy values for the ASC (A) and DC (D) tasks. These values are shown individually for the source domain (S) and target domain (T) parts of the test set. The ASC accuracy values of the related work [7, 8, 12] are provided as baseline systems.

We make the following observations. First, the effectiveness of the domain adaptation proposed by Mezza et al. [12] has been verified, as it increases the ASC target domain accuracy a_A^T from 0.19 (without any domain adaptation) to 0.46, which is only slightly below the accuracy value of 0.51 reported in [12]. Second, the results from configuration `C1` show that by simply combining training data from both domains, one cannot outperform this unsupervised domain adaptation approach, which does not require any target domain training data. Third, throughout all configurations `C2`, `C2M`, `C3`, and `C3M`, the model makes perfect DC predictions for the source domain test data while failing to generalize to the target domain.

[3]The singleton channel dimension is omitted for better readability.

Config	DA	ASC				DC			
		$a_{\mathrm{A}}^{\mathrm{S}}$		$a_{\mathrm{A}}^{\mathrm{T}}$		$a_{\mathrm{D}}^{\mathrm{S}}$		$a_{\mathrm{D}}^{\mathrm{T}}$	
		$(a_{\mathrm{A}}^{\mathrm{S}+})$	$(a_{\mathrm{A}}^{\mathrm{S}-})$	$(a_{\mathrm{A}}^{\mathrm{T}+})$	$(a_{\mathrm{A}}^{\mathrm{T}-})$	$(a_{\mathrm{D}}^{\mathrm{S}+})$	$(a_{\mathrm{D}}^{\mathrm{S}-})$	$(a_{\mathrm{D}}^{\mathrm{T}+})$	$(a_{\mathrm{D}}^{\mathrm{T}-})$
Related Work									
[7]	U*		0.65		0.32		-		-
[8]	U*		0.64		0.45		-		-
[12]	U		0.66		**0.51**		-		-
Proposed Method									
C0	-		0.66		0.19		-		-
C0[12]	U		0.66		**0.46**		-		-
C1	S		0.64		0.45		-		-
C2	U*		0.61		0.40		1.00		0.02
C2M	U*	0.63	0.08	0.38	0.09	0.29	1	0.41	0.03
C3	S		0.62		0.37		1.00		0.02
C3M	S	0.62	0.12	0.39	0.12	0.71	1	0.16	0.01

Table 1. Experimental results for the ASC (A) and DC (D) tasks for the loss configurations described in Section 3.3. Accuracy values are provided for both source (S) and target (T) domain data taken from the test set. Results of prior work on unsupervised DA on the same dataset from [7, 8, 12] are added as reference. The second column indicates whether the DA is unsupervised (U), supervised (S), or a hybrid form of both (U*) (see Section 3.3) For both configurations C2M and C3M, which include the embedding masking step (see Section 3.2), accuracy values are provided both embedding mask settings (additional superscript "+" indicates that m is filled with ones in its first half and with zeros in its second half vice versa for "-") to illustrate the task disentanglement.

In order to get a better insight into the effectiveness of the proposed embedding masking (see Section 3.2) for task disentanglement, we report individual accuracy values for all task-domain pairs for both embedding mask settings in C2M and C3M. An additional superscript "+" indicates that the embedding vector m is filled with ones in its first half and zeros in the second half vice versa for "-". We observe that both the ASC and DC tasks were disentangled to a certain degree since the upper embedding half shows clearly higher accuracy values for the ASC task (compare $a_{\mathrm{A}}^{\mathrm{T}+}$ against $a_{\mathrm{A}}^{\mathrm{T}-}$ and $a_{\mathrm{A}}^{\mathrm{S}+}$ against $a_{\mathrm{A}}^{\mathrm{S}-}$) while the lower embedding half shows a better performance on the DC task (compare $a_{\mathrm{D}}^{\mathrm{S}+}$ against $a_{\mathrm{D}}^{\mathrm{S}-}$). However, despite the task disentanglement, both models do not generalize at all to the target domain for the DC task as it can be seen for $a_{\mathrm{D}}^{\mathrm{T}+}$. Finally, adding the embedding masking shows a minor performance improvement on the target domain test set for the supervised DA scenario (compare $a_{\mathrm{A}}^{\mathrm{T}+}$ in C3M against $a_{\mathrm{A}}^{\mathrm{T}}$ in C3), while the opposite can be observed for the unsupervised DA scenario (C2M against C2).

5. CONCLUSION

In this paper, we adapt and investigate a disentanglement learning approach based on embedding masking for the task of acoustic scene classification. We propose to combine this method with a combination of cross-entropy and variance based losses in order to better disentangle the microphone characteristics (audio domain) and the acoustic scene. To get a better insight into the effectiveness of the disentanglement learning approach, we conduct a systematic study on six different training configurations, which model different unsupervised and supervised DA scenarios. Our results show that while the two tasks were disentangled in the internal embedding representations, the generalization towards target domain data could only slightly be improved in a supervised DA setting. Furthermore, we could confirm the effectiveness of a state-of-the-art unsupervised domain adaptation approach [12], which—in comparison to the proposed method—performs an across-domain adaptation of the data in the feature space. A possible research question for future work is how to best combine such adaptation strategies both at the feature level as well as the level of internal data representations within the ASC model.

6. REFERENCES

[1] Jakob Abeßer, "A Review of Deep Learning Based Methods for Acoustic Scene Classification," *applied sciences*, vol. 10, no. 6, 2020.

[2] Wouter M. Kouw and Marco Loog, "An introduction to domain adaptation and transfer learning," *arXiv preprint arXiv:1812.11806*, 2018.

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

[4] Pavel Denisov, Ngoc Thang Vu, and Marc Ferras Font, "Unsupervised domain adaptation by adversarial learning for robust speech recognition," *arXiv preprint arXiv:1807.11284*, 2020.

[5] Tiantian Tang, Xinyuan Zhou, Yanhua Long, Yijie Li, and Jiaen Liang, "CNN-based Discriminative Training for Domain Compensation in Acoustic Event Detection with Frame-wise Classifier," *arXiv preprint arXiv:2103.14297*, 2021.

[6] Donmoon Lee and Kyogu Lee, "Cross-Domain Semi-Supervised Audio Event Classification using Contrastive Regularization," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2021.

[7] Shayan Gharib, Konstantinos Drossos, Cakir Emre, Dmitriy Serdyuk, and Tuomas Virtanen, "Unsupervised Adversarial Domain Adaptation for Acoustic Scene Classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE)*, Surrey, UK, 2018.

[8] Konstantinos Drossos, Paul Magron, and Tuomas Virtanen, "Unsupervised Adversarial Domain Adaptation based on the Wasserstein Distance for Acoustic Scene Classification," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2019, pp. 259–263.

[9] Seongkyu Mun and Suwon Shon, "Domain Mismatch Robust Acoustic Scene Classification Using Channel Information Conversion," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 845–849.

[10] Alessandro Ilic Mezza, Emanuël A. P. Habets, Meinard Müller, and Augusto Sarti, "Feature Projection-Based Unsupervised Domain Adaptation for Acoustic Scene Classification," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Espoo, Finland, 2020, pp. 1–6.

[11] David S. Johnson and Sascha Grollmisch, "Techniques improving the robustness of deep learning models for industrial sound analysis," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Amsterdam, The Netherlands, 2020, pp. 81–85.

[12] Alessandro Ilic Mezza, Emanuël A. P. Habets, Meinard Müller, and Augusto Sarti, "Unsupervised domain adaptation for acoustic scene classification using band-wise statistics matching," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Dublin, Ireland, 2021, pp. 11–15.

[13] Paul Primus, Hamid Eghbal-zadeh, David Eitelsebner, Khaled Koutini, Andreas Arzt, and Gerhard Widmer, "Exploiting Parallel Audio Recordings to Enforce Device Invariance in CNN-based Acoustic Scene Classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE)*, New York, NY, USA, 2019, pp. 204–208.

[14] Michał Kośmider, "Calibrating Neural Networks for Secondary Recording Devices," Tech. Rep., DCASE2019 Challenge, 2019.

[15] Hu Hu, Sabato Marco Siniscalchi, Yannan Wang, and Chin Hui Lee, "Relational Teacher Student Learning with Neural Label Embedding for Device Adaptation in Acoustic Scene Classification," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020, pp. 1201–1205.

[16] Jongpil Lee, Nicholas J. Bryan, Justin Salamon, Zeyu Jin, and Juhan Nam, "Metric learning vs classification for disentangled music representation learning," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Montréal, Canada, 2020, pp. 439–445.

[17] Andreas Veit, Serge Belongie, and Theofanis Karaletsos, "Conditional similarity networks," in *arXiv preprint arXiv:1603.07810*, 2017.

[18] Andrew Zhai and Hao-Yu Wu, "Classification is a strong baseline for deep metric learning," in *Proceedings of the British Machine Vision Conference 2019 (BMVC)*, Cardiff, UK, 2020, pp. 1–12.

[19] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton, "Layer Normalization," *arXiv preprint arXiv:1607.06450*, 2016.