Relational Teacher Student Learning with Neural Label Embedding for Device Adaptation in Acoustic Scene Classification

Hu Hu[1], Sabato Marco Siniscalchi[2], Yannan Wang[3], Chin-Hui Lee[1]

[1]Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA
[2]Computer Engineering School, University of Enna, Italy
[3]Tencent Media Lab, Tencent Corporation, Shenzhen, Guangdong, China

huhu@gatech.edu, marco.sinsalchi@unikore.it, yannanwang@tencent.com, chl@ece.gatech.edu

Abstract

In this paper, we propose a domain adaptation framework to address the device mismatch issue in acoustic scene classification leveraging upon neural label embedding (NLE) and relational teacher student learning (RTSL). Taking into account the structural relationships between acoustic scene classes, our proposed framework captures such relationships which are intrinsically device-independent. In the training stage, transferable knowledge is condensed in NLE from the source domain. Next in the adaptation stage, a novel RTSL strategy is adopted to learn adapted target models without using paired source-target data often required in conventional teacher student learning. The proposed framework is evaluated on the DCASE 2018 Task1b data set. Experimental results based on AlexNet-L deep classification models confirm the effectiveness of our proposed approach for mismatch situations. NLE-alone adaptation compares favourably with the conventional device adaptation and teacher student based adaptation techniques. NLE with RTSL further improves the classification accuracy.

Index Terms: neural label embedding, domain adaptation, acoustic scene classification, teacher-student learning

1. Introduction

In recent years, we have witnessed a great progress in the acoustic scene classification (ASC) task, as demonstrated by the high participation in the IEEE Detection and Classification of Acoustic Scenes and Events (DCASE) challenges [1, 2, 3]. Top ASC systems use deep neural networks (DNNs), and the main ingredient of their success is the application of deep convolutional neural networks (CNNs) [4, 5, 6, 7, 8, 9]. Further boost in ASC performance is obtained with the introduction of advanced deep learning techniques, such as attention mechanism [10, 11, 12], mix-up [13, 14], Generative Adversial Network (GAN) and Variational Auto Encoder (VAE) based data augmentation [15, 16], and deep feature learning [17, 18, 19, 20]. Nevertheless, those ASC systems yet do not work well when processing audios from mismatched domain, e.g., audios recorded with different devices [3]. Device mismatch is an inevitable problem in a real production, and it is therefore an important aspect to handle when deploying an ASC system. Indeed, a new sub-task, namely *Task1b*, has been added to DCASE 2018 [3] to foster research in that direction. The goal is to design a system that can attain a good performance on 10-second audios segments collected with target devices, which are either not represented at a development phase, or represented during the ASC system deployment with a scarce amount of training material compared to that available for the source device. However, Task1b attracted only a minor interest among

DCASE 2018 and 2019 participants, and even fewer teams were directly concerned with the device mismatch issue.

In the literature, there exist a few approaches that tackle the domain invariant problem in ASC. For example, multi-instance learning [21], and low-level or mid-level feature learning [22, 17, 14], which however address the robustness issue in a broader sense. Less approaches have instead been proposed to directly combat the ASC device mismatch issue, which is actually the focus of the present work. In particular, spectrum correction [23] and channel conversion [24] build a front-end module to convert speech features from the source domain to target domain before feeding them to the back-end classifier. Besides front-end features, mid-level feature based transfer systems, which uses bottleneck features [25] or hidden layer representations [26] are adopted to transfer knowledge from source to target domain. Adversarial training methods in [27, 28] leverage an extra domain discriminator to solve the device mismatch problem although the key focus is on lack of labeled target data.

Teacher-student (TS) learning, also named as knowledge distillation [29], has recently been shown to be effective in ASC and other domain adaptation speech tasks, e.g., [30, 31, 32, 33]. The key idea is to minimizes the distance measurement between teacher and student model output distributions, i.e., the information is transferred at a soft-label level. In [34], relational knowledge distillation (RKD) is demonstrated to improve the knowledge distillation process. RKD takes into account the relations of outputs rather than individual outputs themselves. Unfortunately, conventional TS learning can be applied with success if: (i) source and target data is from the same or similar domain [30, 35], or (ii) source and target data come in pair although belong to different domains [32, 33, 36]. Neural label embedding (NLE), recently proposed in [37], is an ingenues solution to distill knowledge across domains when neither of the aforementioned two requirements could be met. NLE can be viewed as the centroid of soft labels from the same class. As to extension of soft labels, it encodes the knowledge distilled from the source domain and teacher model, which can then be transferred to the target domain.

In this study, we extend the NLE adaptation scheme [37] by taking into account relationships among different acoustic scenes during adaptation. We achieve this goal by proposing a relational teacher student learning (RTSL) approach based on NLE for ASC device mismatching problem. First, NLE is learned from a relatively large-size source data set, i.e., collected with the source devices. Next, ASC system is adapted to the target device leveraging upon target domain data only, i.e., teacher-student learning with unpaired data, and the set of NLE, one each per acoustic scene class. The proposed solution is assessed against the DCASE 2018 Task1b data. Experimental re-
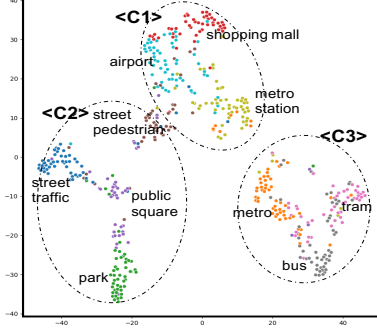
Figure 1: *SKLD based t-SNE plot of an AlexNet-L model outputs, i.e., posterior probabilities. $C1 - C3$ indicate public indoor area, public out-door area and transportation clusters, respectively.*

sults confirm our intuitions and demonstrate that our adaptation technique generates a significant classification improvement on target domain data. Indeed, NLE-based TS adaptation outperforms both (i) multi-device training strategies, and (ii) conventional TS adaptation schemes. Furthermore, an additional boost is obtained when TS adaptation is carried out leveraging structural information.

2. Neural Label Embedding for Adaptation

We focus on transfer learning of structural relationships among ASC classes. Conventionally one-hot label is used to train the ASC system, assuming the similarities between each class pair to be the same at the label level. To visualize the distance between two probability densities, we use SKLD based t-SNE [38], an extension of Stochastic neighbor embedding (SNE) [39], to transform the observations in high-dimensional space into a low-dimensional space which preserves neighbour identities by minimizing the symmetrical Kullback-Leibler divergence (SKLD) of the pairwise distributions between the two spaces where Student t distribution is assumed. In Figure 1, the points are 2-dim representations of audio samples, recorded with Device A, scattering all over the figure. On the other hand, three clusters, labeled C1, C2 and C3, representing public indoor, public out-door and transportation areas, respectively, can be observed. This confirms our conjecture and intuition that the intrinsic relationships between different acoustic scenes depend more on the recording environments.

2.1. Source Model Training

We now describe NLE's generation and usage as suggested in [37]. NLE starts with building a stand-alone deep model for the ASC task, as shown in the <Step 1> of Figure 2. Here, the source data, $X^S = \{x_1^S, x_2^S, ..., x_{N_S}^S\}$, and the related one-hot class label vectors, $Y^S = \{y_1^S, y_2^S, ..., y_{N_S}^S\}$, are employed to estimate the model parameters with a Cross Entropy (CE) loss:

$$L_{CE}(X^S, Y^S) = -\frac{1}{N_S} \sum_{x_i^S, y_i^S \in D_S} \sum_{j=1}^{K} y_{i,j}^S \log F_S(x_i^S)_j, \tag{1}$$

where D_S indicates source domain data, N_S is the number of input samples, K is the number of output classes which is also the dimension of the one-hot label vector, and F_S represents the
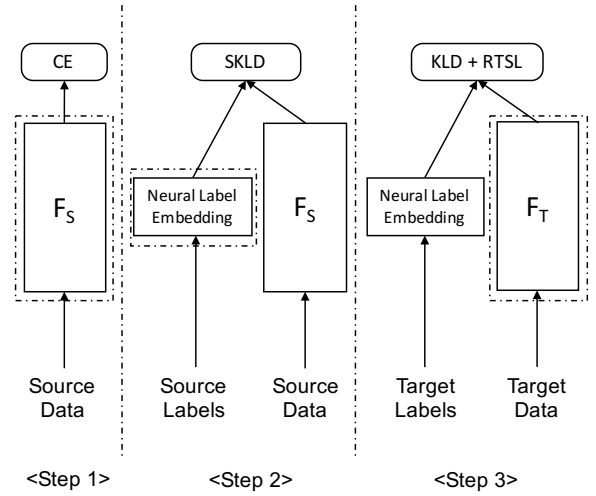
source domain model, performing input-output mapping.

Figure 2: *Framework of NLE with RTSL. The parameters of the models in the dashed boxes in Steps 1-3 are learnable.*

2.2. NLE Label Generation

NLE can function as labels in place of the one-hot class labels for the source data. The goal is to distill the knowledge of F_S into a dictionary of NLE vectors, one for each acoustic scene class, which is predicted at the output layer in <Step 2> of Figure 2. Each K-dim NLE vector for class c, NLE_c, encodes the statistical relationship between class c and all other classes via the output distributions of the source-domain F_S given all features aligned with class c at the input. Thus, NLE is generated by SKLD (γ_2), which is to measure the distance between two distributions, P and Q, defined in the following:

$$\gamma_1(P, Q) = \sum_{p_i, q_i \in P, Q} p_i \log \frac{p_i}{q_i}, \tag{2}$$

$$\gamma_2(P, Q) = \frac{1}{2}(\gamma_1(P, Q) + \gamma_1(Q, P)). \tag{3}$$

A $K \times K$ NLE matrix, $[NLE_1, NLE_2, ..., NLE_K]$, denoted as NLE_M, can thus be trained using SKLD. The associated loss function L_{LE} is defined as follows:

$$L_{LE} = \frac{1}{N_S} \sum_{x_i^S, y_i^S \in D_S} \gamma_2(F_S(x_i^S), NLE_M \cdot y_i^S) \tag{4}$$

where $NLE_M \cdot y_i^S$ represents the label embedding vector of input x_i^S with the one-hot label vector y_i^S. The matrix multiplication "·" is an indexing operation on NLE_M. It should be noted that all elements in the vector NLE_c should sum to 1.

2.3. NLE for Device Domain Adaptation

The final step to accomplish device adaptation, as indicated in <Step 3> in Figure 2, is to obtain the adapted target model, F_T, starting from the seed source model, F_S, using only the trained NLE_M and target domain data D_T with available inputs $X^T = \{x_1^T, x_2^T, ..., x_{N_T}^T\}$ and labels $Y^T = \{y_1^T, y_2^T, ..., y_{N_T}^T\}$. This is different from conventional TS, which requires paired source-target data, including soft labels obtained with F_S. In our case, Y^T, the one-hot labels available

for the target data, are used to fetch and be replaced by the corresponding NLE in NLE_M. Finally, KLD can be used to train F_T with the following criterion:

$$L_{NLE}(X^T, Y^T) = \frac{1}{N_T} \sum_{x_i^T, y_i^T \in D_T} \gamma_1(NLE_M \cdot y_i^T, F_T(x_i^T)). \tag{5}$$

It should be noted that with a relatively large-scale target domain data, the simplest way to adapt F_S to the target condition is to use the target data and one-hot labels Y^T to fine-tune the original model using the CE loss. However, that easily leads to over-fitting as the amount of target data decreases. The model to be used for TS learning in <Step 3> will be presented next.

3. Relational Teacher Student Learning with Neural Label Embedding

As aforementioned, for conventional TS learning for domain adaptation, soft labels generated from teacher are employed to train the student with the KLD. With well-trained source model F_S, the loss L_{TS} for TS learning is

$$L_{TS}(X^T, Y^T) = \frac{1}{N_T} \sum_{X_i^{S'} \in D_S, X_i^T \in D_T} \gamma_1(F_S(x_i^{S'}), F_T(x_i^T)), \tag{6}$$

where $X^{S'}$, in the case of acoustic scene classification, is often a subset of X^S owing paired target data, X^T. In this work, paired data represent two acoustic scenes collected in the same location and time but with two different devices. Traditional TS learning encodes the knowledge from source domain data and source model into soft labels. Benefiting from this, knowledge can be transferred from the source domain to target domain.

Taking into account the structural relationships among output classes has shown to be beneficial in [34]. We therefore propose to use relational teacher student learning (RTSL) to carry out device adaptation and generate the target model leveraging upon NLE, and their structural relationships. Using ASC task as an example, as visualized in Figure 1, the measured distance between sounds from metro and park should be larger than the sounds from metro and train. This property is from the feature of acoustic scene itself, and it should not be modified by the nature of the device used during the recording phase. Thus, we can minimize the distance of each class between source and target domain data. In our experiments, the SKLD is used as the measurement of distance among probability distributions.

The total mutual distance value, V, of the target model output probabilities, $F_T(X^T)$, is obtained considering all possible N^2 pairs in the training set. V is thus givn in Eq. (7).

$$V(F_T(X^T)) = \frac{1}{N_T^2} \sum_{x_i^T \in D_T} \sum_{x_j^T \in D_T} \gamma_2(x_i^T, X_j^T). \tag{7}$$

During RTSL, we minimize the divergence between total mutual distance of output and NLE. In our experiments, we use smoothed L1 loss, a.k.a. Huber loss [40], which is given by

$$SL1(x, y) = \begin{cases} \frac{1}{2}(x-y)^2, & |x-y| \leq 1, \\ |x-y| - \frac{1}{2}, & |x-y| > 1. \end{cases} \tag{8}$$

Therefore, the loss for RTSL is

$$L_{RTSL} = SL1(V(NLE_M), V(F_T(X^T))). \tag{9}$$

During target model training of relational teacher student learning with NLE, we use the linear combination of L_{NLE} in Eq. (5) and L_{RTSL} in Eq. (9), which is given by

$$L_{NLE-RTSL} = L_{NLE} + \lambda * L_{RTSL} \tag{10}$$

where λ is a tunable hyper-parameter to balance the loss terms. According to our experiments, this parameter is very robust to the model training. We set $\lambda = 10$ to make the two contributors of our loss have the same scale range.

4. Experiments and Analysis

4.1. Experimental Setup

The proposed adaptation framework is evaluated on the DCASE 2018 Task1b development data set [3]. Task1b is provided with 28 hours of acoustic scene audio recordings acquired with three real devices, namely device A (24 hours), device B (2 hours) and device C (2 hours). Audio segments are recorded in 10 different acoustic scenes, and each audio segment lasts 10 seconds. For each 10-second single-channel audio segment, short-time Fourier transofrm (STFT) with 2048 FFT points is applied, using a window analysis of 25ms and a shift of 10ms. Mel filter-banks with 128 bins are used to extract the 128-dimension log-Mel filter bank energy (128-D LMFB) features. In our experiments, we follow the officially recommended train-test partition of the data. Training data from device A is regarded as the source domain data, and device B and device C are regarded as target domain. It should be noticed that our goal is slightly different from that for the Task1b. We regard data from devices B and C as two separate target domains. We aims to solve the device mismatch issue for one specific target device at a time, which is a more common scenario in real applications.

All ASC deep models used in our experiments use an AlexNet [41] based CNN structure. Resource constraints impose us to reduced the amount of neural parameters, and we refer to this model as AlexNet-L, and it's implemented in PyTorch. AlexNet-L has five convolutional layers with the kernel size of 4×4 and two fully connected layers with the hidden dimension of 1024. Each convolutional layer consists of a convolution operation, a batch normalization block, non-linear processing block with ReLU activation function, and a max pooling block. Each input audio recording is chopped into segments of 20 frames (0.2 seconds per segment). Therefore, final scene classification is obtained through majority voting among all segments. AlexNet-L is trained with stochastic gradient descent (SGD) algorithm with cosine based learning rate scheduler. The initial learning rate is set to 0.01 for baseline training, and 0.002 for adaptations on target device. The temperature parameter [31] is used and set to 2.0 in all soft labels and NLE-based experiments. The linear combination parameter λ in Eq. (10) is set to 10 to balance the loss terms.

4.2. Experimental Results

Table 1 shows our experimental results on DCASE 2018 Task1. From the top three rows in Table 1, we can observed that device mismatch is indeed an critical aspect in acoustic scene classification. The official baseline system [3], and the AlexNet-L (All Devs), listed in the two first rows, are trained using data from all the three devices, namely A, B, and C. The amount of Device A training material is much larger than that available for Devices B & C (24 hours against 2 hours); therefore, when the baseline systems are skewed toward Device A, and the degradation

in the classification accuracy when moving from Device A test data to device B & C test data is significant. In the third row in Table 1, we report acoustic scene classification accuracy when AlexNet-L is trained only on device A data (this is, Device B & C data is missing). This experiment is conducted to clarify any doubt about the severity of the device mismatch issue. In fact, the accuracy drops from 69.6% on device A down to 13.4% and 16.2% on Devices B and C, respectively.

Table 1: *Evaluation results on DCASE 2018 Task1b.*

Model	Dev A acc.(%)	Dev B acc.(%)	Dev C acc.(%)
Official Baseline [3]	58.9	45.6	52.3
AlexNet-L (All Devs)	67.0	52.5	54.7
AlexNet-L (Dev A)	69.6	13.4	16.2
AlexNet-L (Dev A)			
+ one-hot adaptation	-	57.0	60.8
+ soft labels adaptation	-	56.7	60.6
+ NLE adaptation	-	58.7	62.0
+ NLE-RTSL adaptation	-	**59.2**	**64.2**

In Section 2, we pointed out that a simple yet viable domain adaptation approach with labeled adaptation data is to fine-tune the source deep model using adaptation data and one-hot labels. We put forth this supervised adaptation scheme using AlexNet-L (Dev A), trained on Device A data only, as a seed model, and and Device B & C data (separately) along with one-hot class labels, and the CE loss in Equation 1. From Table 1, we can observe that one-hot adaptation improves performance on target devices (see 4th row), and classification accuracy goes from 13.4% to 57.0% for Device B and from 16.2% to 60.8% for Device C. By comparing with multi-device training (2nd row), we can conclude that one-hot adaptation outperforms multi-device training.

To better appreciate our contribution, the conventional TS learning, dubbed soft labels adaptation in Table 1, is also implemented. This adaptation approach can be put forth because Device B & C data are paired with some of Device A data. The teacher is AlexNet-L (Dev A). The student is obtained by copying the teacher architecture and then adapting it using target device data and soft labels generated by the teacher model fed with the Device A paired data. In soft labels adaptation, the teacher parameters remain unchanged. From the results in the 4rd, and 5th rows in Table 1, we can observe that slightly worse results are obtained using TS with soft labels adaptation than using a simpler one-hot adaptation solution. That may be due to the the mismatch between source and target domain is large.

Finally, we turn our attention to the key architecture proposed for device adaptation. In the last two rows in Table 1, experimental results of TS adaptation with NLE (LE adaptation) and NLE with RTSL framework (NLE-RTSL adaptation) are reported. First, we can observe that both NLE adaptation approaches attain superior performance compared to one-hot adaptation. For Device B, classification accuracy goes from 57.0% to 58.7%, and from 60.8% to 62.0% for Device C. A significant boost in the classification results can further be obtained by imposing structural constrains during knowledge distillation. Indeed, NLE-RTSL adaptation allows us to attain top performance of 59.2% for Device B, and 64.2% on Device C, respectively.
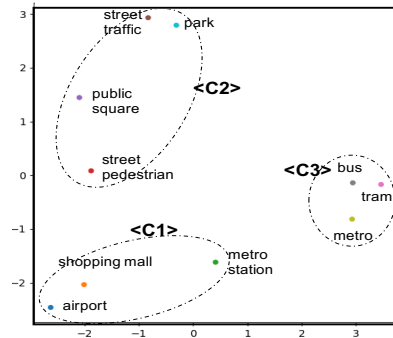

Figure 3: *Scatter plot on two-dimensional space of NLE vectors. $C1 - C3$ indicates the three clusters of public in-door area, public out-door area and transportation, respectively.*

4.3. Visual Analysis

A visual analysis can shed light on the nature of our proposed NLE for acoustic scenes. Recall in Figure 1 we show audio samples belonging to Device A obtained with SKLD based t-SNE method. It is easy to see that three main clusters arise. They are circled with a dashed ellipses labeled as C1, C2 and C3, for public in-door, public out-doors, and transport-related areas, respectively. A close visual inspection immediately reveals some key features of those acoustic scenes. For instance, metro samples are close to the metro station samples, which are both semantically related to the metro concept. In the same way, street pedestrian samples are close to street traffic samples. In addition, street pedestrian is also close to the public in-door area, since both scenes can capture walking people. Otherwise, street pedestrian scene is far away from bus-, metro-, and tram-related acoustic scenes. However, within classes we can observe some overlap, and that because not all soft labels are equally good. In Figure 3, we plot the NLE vectors, one centriod per acoustic scene. NLE are generated from the source Device A training data and Principal component analysis (PCA) is applied for dimension reduction. Comparing Figures 1 and Figure 3, we can see that the three main clusters are present in both plots, yet we have a single NLE per acoustic scene which makes our proposed adaptation approach less sensitive to the quality of the soft labels. In summary, the two plots reveal the correctness of our conjecture for an existence of structural relationships between acoustic scenes. Moreover, NLE encodes these structural relationships in a compact way in low-dimensional spaces.

5. Conclusion

In this paper, a relational teacher student learning framework with neural label embedding is proposed to resolve the device mismatch issue in acoustic scene classification. We explore the similarities or dissimilarities between pairs of classes. This structural relationship is learned and encoded into NLE and then transferred from the source device domain to the target device domain via the relational teacher-student approach. Our proposed framework is assessed against the DCASE 2018 Task1b development data set, and experimental results demonstrate not only the viability of our approach, but also that a significant improvement of the classification accuracy on the target device data can be obtained. Furthermore, a visual analysis is provided to shed light on the key characteristics of the proposed neural label embedding concept.

6. References

[1] A. Mesaros, T. Heittola, 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 T-ASLP*, vol. 26, no. 2, pp. 379–393, 2018.

[2] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proc. DCASE 2017 Workshop*, November 2017, pp. 85–92.

[3] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proc. DCASE 2018 Workshop*, November 2018, pp. 9–13.

[4] A. Mesaros, T. Heittola, 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 T-ASLP*, vol. 26, no. 2, pp. 379–393, 2018.

[5] Y. Han and K. Lee, "Acoustic scene classification using convolutional neural network and multiple-width frequency-delta data augmentation," *arXiv preprint arXiv:1607.02383*, 2016.

[6] D. Battaglino, L. Lepauloux, and N. Evans, "Acoustic scene classification using convolutional neural networks," DCASE2016 Challenge, Tech. Rep., September 2016.

[7] H. Hu, C.-H. H. Yang, X. Xia, X. Bai, X. Tang, Y. Wang, S. Niu, L. Chai, J. Li, H. Zhu, F. Bao, Y. Zhao, S. M. Siniscalchi, Y. Wang, J. Du, and C.-H. Lee, "Device-robust acoustic scene classification based on two-stage categorization and data augmentation," DCASE2020 Challenge, Tech. Rep., June 2020.

[8] K. Koutini, H. Eghbal-zadeh, M. Dorfer, and G. Widmer, "The receptive field as a regularizer in deep convolutional neural networks for acoustic scene classification," *CoRR*, vol. abs/1907.01803, 2019.

[9] H. Hu, S. M. Siniscalchi, Y. Wang, X. Bai, J. Du, and C.-H. Lee, "An acoustic segment model based segment unit selection approach to acoustic scene classification with partial utterances," in *Interspeech*, 2020.

[10] J. Wang and S. Li, "Self-attention mechanism based system for dcase2018 challenge task1 and task4," in *Proc. DCASE Challenge*, 2018, pp. 1–5.

[11] Z. Ren, Q. Kong, K. Qian, M. D. Plumbley, B. Schuller *et al.*, "Attention-based convolutional neural networks for acoustic scene classification," in *DCASE 2018 Workshop Proceedings*, 2018.

[12] X. Bai, J. Du, J. Pan, Z. Heng-shun, Y.-H. Tu, and C.-H. Lee, "High-resolution attention network with acoustic segment model for acoustic scene classification," in *Proc. ICASSP*. IEEE, 2020.

[13] K. Xu, D. Feng, H. Mi, B. Zhu, D. Wang, L. Zhang, H. Cai, and S. Liu, "Mixup-based acoustic scene classification using multichannel convolutional neural network," in *Pacific Rim Conference on Multimedia*. Springer, 2018, pp. 14–23.

[14] T. Nguyen and F. Pernkopf, "Acoustic scene classification with mismatched devices using cliquenets and mixup data augmentation," in *Interspeech*, 2019.

[15] H. Chen, Z. Liu, Z. Liu, P. Zhang, and Y. Yan, "Integrating the data augmentation scheme with various classifiers for acoustic scene modeling," DCASE2019 Challenge, Tech. Rep., June 2019.

[16] S. Mun, S. Park, D. K. Han, and H. Ko, "Generative adversarial network based acoustic scene training set augmentation and selection using svm hyper-plane," *Proc. DCASE*, pp. 93–97, 2017.

[17] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *Proc. ICASSP*. IEEE, 2019, pp. 3852–3856.

[18] H. Zeinali, L. Burget, and J. Cernocky, "Convolutional neural networks and x-vector embedding for dcase2018 acoustic scene classification challenge," *arXiv preprint arXiv:1810.04273*, 2018.

[19] S. Mun, S. Shon, W. Kim, and H. Ko, "Deep neural network bottleneck features for acoustic event recognition." in *Interspeech*, 2016, pp. 2954–2957.

[20] P. Sharma, V. Abrol, and A. Thakur, "Ase: Acoustic scene embedding using deep archetypal analysis and gmm." in *Interspeech*, 2018, pp. 3299–3303.

[21] H. Song, J. Han, S. Deng, and Z. Du, "Acoustic scene classification by implicitly identifying distinct sound events," *arXiv preprint arXiv:1904.05204*, 2019.

[22] H. Eghbal-Zadeh, B. Lehner, M. Dorfer, and G. Widmer, "Cp-jku submissions for dcase-2016: A hybrid approach using binaural i-vectors and deep convolutional neural networks," *Proc. IEEE AASP Challenge Workshop on DCASE*, 2016.

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

[24] S. Mun and S. Shon, "Domain mismatch robust acoustic scene classification using channel information conversion," in *Proc. ICASSP*. IEEE, 2019, pp. 845–849.

[25] S. Mun, S. Shon, W. Kim, and H. Ko, "Deep neural network bottleneck features for acoustic event recognition." in *Interspeech*, 2016, pp. 2954–2957.

[26] 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," *arXiv preprint arXiv:1909.02869*, 2019.

[27] S. Gharib, K. Drossos, E. Cakir, D. Serdyuk, and T. Virtanen, "Unsupervised adversarial domain adaptation for acoustic scene classification," *arXiv preprint arXiv:1808.05777*, 2018.

[28] K. Drossos, P. Magron, and T. Virtanen, "Unsupervised adversarial domain adaptation based on the wasserstein distance for acoustic scene classification," in *Proc. WASPAA*. IEEE, 2019, pp. 259–263.

[29] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[30] H.-S. Heo, J.-w. Jung, H.-j. Shim, and H.-J. Yu, "Acoustic scene classification using teacher-student learning with soft-labels," *arXiv preprint arXiv:1904.10135*, 2019.

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

[32] J. Li, M. L. Seltzer, X. Wang, R. Zhao, and Y. Gong, "Large-scale domain adaptation via teacher-student learning," *arXiv preprint arXiv:1708.05466*, 2017.

[33] Z. Meng, J. Li, Y. Zhao, and Y. Gong, "Conditional teacher-student learning," in *Proc. ICASSP*. IEEE, 2019, pp. 6445–6449.

[34] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3967–3976.

[35] H. Hu, T. Tan, and Y. Qian, "Generative adversarial networks based data augmentation for noise robust speech recognition," in *Proc. ICASSP*. IEEE, 2018, pp. 5044–5048.

[36] W. Li, S. Wang, M. Lei, S. M. Siniscalchi, and C.-H. Lee, "Improving audio-visual speech recognition performance with cross-modal student-teacher training," in *Proc. ICASSP*. IEEE, 2019, pp. 6560–6564.

[37] Z. Meng, H. Hu, J. Li, C. Liu, Y. Huang, Y. Gong, and C.-H. Lee, "L-vector: Neural label embedding for domain adaptation," in *Proc. ICASSP*. IEEE, 2020.

[38] L. van der Maaten and G. E. Hinton, "Visualizing data using t-sne," in *Journal of Machine Learning Research, vol. 9, no. Nov, pp. 2579–2605*, 2008.

[39] G. E. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *Advances in Neural Information Processing Systems 15, S. Becker, S. Thrun, and K. Obermayer, Eds.*, 2003.

[40] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, no. 1, pp. 73–101, 03 1964.

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.