A Supervised Domain Adaptive Method for Multi Device Acoustic Scene Classification

Zhanqi Liu[1,2], Mei Wang[1,3], Liyan Luo[1,2(✉)], Zhenghong Liu[1,2], and Guan Wang[1,2]

[1] Ministry of Education Key Laboratory of Cognitive Radio and Information Processing, Guilin 541006, China
lly1227@guet.edu.cn

[2] School of Information and Communication, Guilin University of Electronic Technology, Guilin 541006, China

[3] School of Information Science and Engineering, Guilin University of Technology, Guilin 541006, China

Abstract. Acoustic scene classification faces performance degradation due to device mismatch when different pickup devices are used in the training and testing phases. To solve the device mismatch problem and improve the performance of the acoustic scene classification system on unseen pickup devices, we propose a joint frequency band standardization and supervised domain adaptation algorithm, which can effectively extract the domain invariant features of the acoustic scene signal to generalize the model to the unseen The algorithm can effectively extract the domain invariant features of the acoustic scene signal and generalize the model to the unseen distribution to solve the performance degradation of the model on the unseen pickups. The frequency band standardization is first used to linearly correct the extracted Log-Mel features, and then combined with supervised domain adaptation to reduce the difference between the source and target domains and correct the nonlinear differences between different kinds of pickup devices. Experimental results on the DCASE Challenge 2020 Task 1A dataset show an overall improvement of 13.9% compared to the baseline model, 6.2% on seen devices, and 20% on unseen device. The algorithm can better extract the domain invariant features of the model, has better classification performance, and enables the model to generalize to unseen device.

Keywords: Acoustic Scene Classification · Supervised Domain Adaptive · Domain generalization · Frequency band standardization

1 Introduction

In the era of Internet of Everything, acoustic scene classification can be applied in many domains, such as smart city construction [1], biodiversity monitoring [2], and urban security surveillance [3]. The goal of the acoustic scene classification task is to classify the collected sound signals to be classified according to predefined acoustic scene categories, thus providing target acoustic scene information for many domains.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2025
Published by Springer Nature Switzerland AG 2025. All Rights Reserved
J. Wang (Ed.): MobiMedia 2023, LNICST 576, pp. 15–24, 2025.
https://doi.org/10.1007/978-3-031-60347-1_2

Acoustic scene signals are complex and variable, containing a large number of different background noises, speech sounds and sudden sounds, so the difficulty of learning scene rules for acoustic scene signal classification models increases. Moreover, in practical application scenarios, the pickup devices for acoustic scene signal acquisition may use various kinds of pickup devices that have been deployed, but the problem of device diversity is raised for acoustic scene classification techniques because the distribution of data acquired by different kinds of pickup devices is not consistent. Pickup devices that are present in the training set are called seen devices, and devices that are not present in the training set but are present in the test set are called unseen device. This inconsistent data distribution phenomenon caused by device mismatch makes the trained acoustic scene classification models exhibit significant performance degradation on other devices, and thus cannot be applied to people's lives.

In the multi-device acoustic scene classification task, the most effective way is to provide more training samples, and the more types of device samples provided, the more the generalization ability of the model can be improved, however, in practical scenarios, the cost of collecting training samples is high, and we can only use the limited data available. Therefore, extracting higher-order acoustic features with invariant characteristics in the device domain and an effective classification algorithm are the keys to improve the overall classification performance of the model. It mainly includes audio sample data optimization and network model structure optimization. Audio sample data optimization mainly includes data enhancement [4, 5], frequency band standardization [6], etc. Although such methods can increase the number of samples or correct some of the sample differences brought by different pickup devices, due to the complexity of the audio sample data of acoustic scenes, the audio sample data of acoustic scenes may contain many overlapping sounds or background noise, and the methods used cannot completely characterize the differences between devices. Therefore, this type of method cannot completely compensate for the differences brought by the devices and has a large limitation. The optimization of the network model structure mainly includes the design of large-scale integrated networks [7], deep residual networks incorporating high and low frequency path separation [8] and networks using two-stage classifiers [9] for classification This type of method can improve the generalization ability of the model by better extracting key features through the optimization of the network model, but the model parameters of this type of method are more and the higher complexity is not conducive to the application of on low-cost mobile devices. The literature [10] argues that the perceptual field of the network is not as large as possible, divides the perceptual field into the maximum perceptual field and the effective perceptual field and proposes the perceptual field to actively adjust the regularization coefficient to optimize the network model.

And with the ability to increase the generalization of the model in different domains is called Domain Generalization (DG). Some of the most relevant approaches are Supervised Learning (SL), Multi-Task Learning (MLT), Transfer Learning (TL) [11], Zero-shot Learning (ZSL) and Domain Adaptation (DA) [12].

The goal of MTL is to use a single model to learn multiple related tasks simultaneously, benefiting from the regularization effect due to parameter sharing and thus applicable to DG, but MTL does not focus on unseen distributions. There are many

similarities between TL and DG, for example, the target and source distributions of TL and DG do not coincide, and TL aims to transfer knowledge learned in one or more domains to another different but related domain ZSL is related to DG in that both deal with unseen distributions, but the domain distribution shift in ZSL is mainly caused by the inconsistent labels used for training and testing. DA is the closest approach to DG, and both DA and DG aim to address the domain shift problem encountered in new testing environments (i.e., inconsistent data distribution between training and test samples has inconsistent data distributions).

The most common methods in DA are still unsupervised DA, including unsupervised adversarial domain adaptation [13] and unsupervised feature alignment domain adaptation [14]. The literature [13] combines frequency normalization and unsupervised adversarial domain adaptation methods to enable the model to obtain classification results close to the source domain on the target domain, but unsupervised DA methods require large-scale unlabeled data on the target domain to achieve better performance. In practical scenarios, collecting large-scale audio sample data requires a large investment, and in the absence of a large amount of data, the unsupervised domain adaptation method cannot extract domain-invariant features well, and the classification effect cannot be guaranteed. In the case of a small number of samples in the target domain, supervised domain adaptive training using the labels of the target domain can effectively improve the classification performance of the model.

To address the above mismatch between training and test data distribution, a multi-device supervised domain adaptive algorithm is proposed in this paper. The algorithm jointly uses frequency band standardization and supervised domain adaptive methods, firstly, in the sample data feature extraction, frequency band standardization is used to correct the extracted Log-Mel spectral map, which can correct the linear difference of seen devices, then data enhancement methods are used in the training phase to enhance the training samples, so that the model sees more training samples to improve the generalization ability of the model, and finally, supervised domain adaptive method [12], the training samples are divided into source and target domains according to the device labels in the training phase, and the nonlinear differences of the devices can be corrected by aligning the intermediate process features of the source and target domains.

2 Proposed Algorithm

2.1 Algorithm Framework

The literature [13] considers that the differences between pickup devices are mainly divided into linear and nonlinear differences. In this paper, the algorithm jointly uses the frequency band standardization method to correct the linear differences between pickup devices and the supervised domain adaptive method to correct the nonlinear differences between pickup devices. The framework of the supervised domain adaptive acoustic scene classification algorithm is shown in Fig. 1. The device domain labels in the training phase are considered as known conditions, and the main devices in the training samples are classified as source domain and other devices are classified as target domain for training according to the device labels. The source domain and target domain training samples with the same scene labels are extracted and fed into the model

separately, and the acoustic scene audio signals of the training samples are feature extracted, Log-Mel is extracted and corrected by frequency band standardization, as well as deltas, delta-deltas features are extracted, and the three features are fused in the channel dimension as the input features of the model. The source and target domain training samples will get the intermediate process features and classification loss of the three convolutional block outputs after passing through the model. The difference between the corresponding intermediate process features of the source and target domains is calculated and is called the domain difference loss, which represents the difference between the extracted features of the source and target domains. Adding this loss to the total training loss can reduce the feature difference between the source and target domains in a nonlinear way during the training process and help the model extract domain invariant features better.

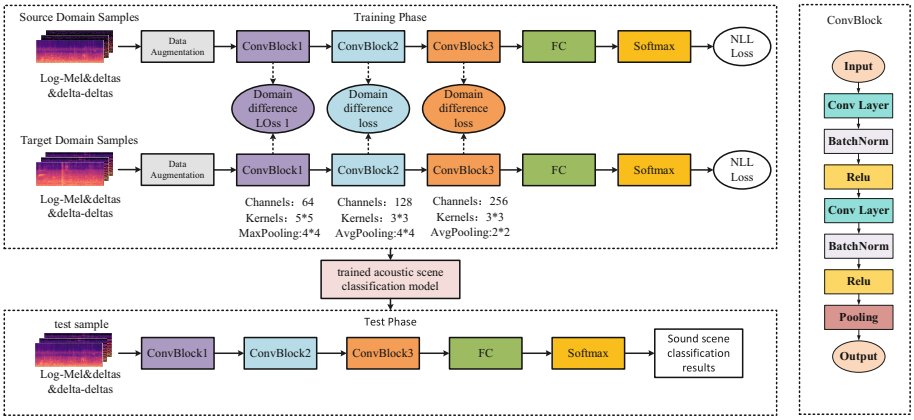

Fig. 1. Framework of supervised domain adaptive acoustic scene classification algorithm

2.2 Frequency Band Standardization

The frequency band standardization is first divided on the sample space of the training set according to the type of equipment, and then the mean and standard deviation are calculated for each band of the acoustic characteristic spectrogram of the equipment, respectively, with the following formula:

$$\mu_{dk} = \frac{1}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} x_{dnmk} \tag{1}$$

$$\sigma_{dk} = \sqrt{\frac{1}{NM-1} \sum_{n=1}^{N} \sum_{m=1}^{M} (x_{dnmk} - \mu_{dk})^2} \tag{2}$$

where d is the device class of the training samples, N is the number of training samples, M is the number of time frames of the training samples, and k is the Mel frequency band of the training samples;

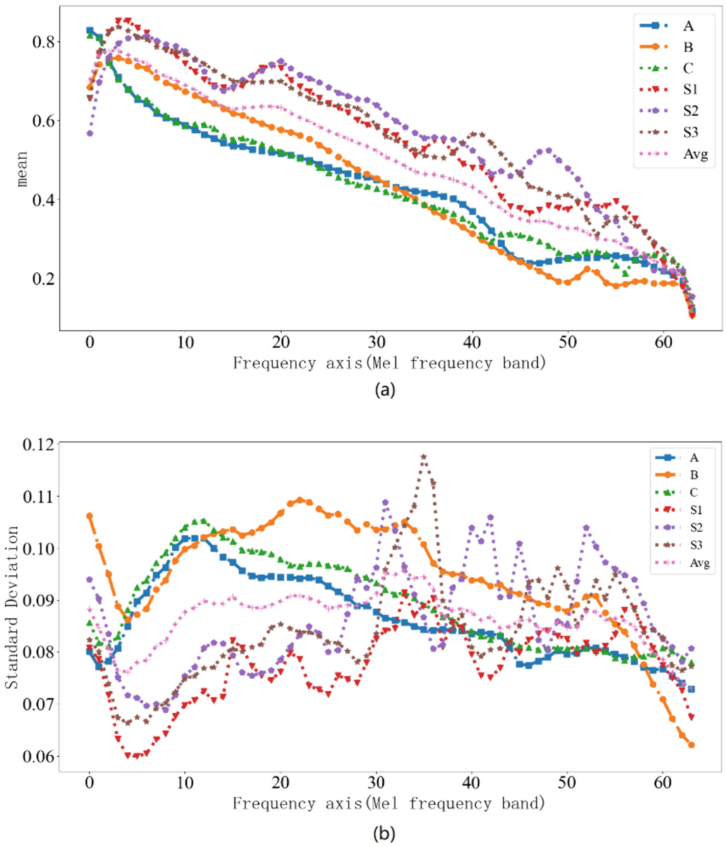

Fig. 2. Frequency band standardization mean (a) and standard deviation (b) curves

The 64-dimensional mean and standard deviation profiles computed by this algorithm on the DCASE2020Task1A [15] multi-device acoustic scene classification dataset are shown in Fig. 2.

Finally, the frequency band standardization process is carried out according to the corresponding equipment type of the input acoustic characteristics, which is calculated as follows:

$$\hat{x}_{dnmk} = \frac{x_{dnmk} - \mu_{dk}}{\mu_{dk}} \tag{3}$$

where x_{dnmk} is the input acoustic feature spectrogram and $\hat{x}_{dnmk}$ is the acoustic feature spectrogram after Frequency band standardization;

2.3 Data Augmentation

The data enhancement module used in this paper algorithm is based on the joint composition of Mixup [4] and SpecAugment [5] methods to obtain the acoustic features

after data enhancement by applying time warping, time masking, frequency masking and hybrid data enhancement methods to the input acoustic features.

The data enhancement methods consist of Mixup and SpecAugment jointly, which are composed of the following methods:

The processing methods of SpecAugment include:

Time warping: overlaying any length of the time spectrogram of the input acoustic feature spectrogram onto any time spectrogram of the input acoustic feature spectrogram;
Temporal Masking: masking the temporal spectrogram of any length in the input acoustic feature spectrogram;
frequency masking: masking the frequency spectra of any length in the input acoustic feature spectrogram;

The composition methods of Mixup are:

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j \tag{4}$$

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j \tag{5}$$

where i, j are positive integers, $\lambda \in (0, 1)$ and conform to the beta distribution, x_i denotes the i-th sample of the input acoustic feature, x_j denotes the j-th sample of the input acoustic feature, $\tilde{x}$ denotes the acoustic feature obtained by hybrid enhancement, y_i denotes the label corresponding to the i-th sample of the input acoustic feature, y_j denotes the label corresponding to the j-th sample of the input acoustic feature, and $\tilde{y}$ denotes the label corresponding to the acoustic feature obtained by hybrid enhancement;

2.4 Domain Difference Loss

In order to reduce the feature differences between source and target domains, in the training phase the main device is divided into source domain and other devices are divided into target domains according to device categories, and each round of training uses source domain samples and target domain samples in the supervised domain adaptive acoustic scene classification model using the mean square error MSE loss function to measure the differences between source and target domains, and performs domain difference calculation to obtain the domain difference loss, and the total loss is obtained by adding the domain difference loss to the source and target domain classification loss, and the model is updated in the reverse direction according to the total loss so as to train the model;

The MSE loss function is calculated as:

$$Loss_{mse} = \frac{1}{m} \sum_{m}^{1} \left(y_i^s - y_i^t\right)^2; \tag{6}$$

where m is the number of elements of the input feature, y_i^s is the value of the i-th element of the source domain, and y_i^t is the value of the i-th element of the target domain;

The supervised domain-adaptive acoustic scene classification model has a total of three feature-aligned convolutional blocks with domain difference losses of $Loss_{mse1}$, $Loss_{mse2}$, $Loss_{mse3}$, Therefore, the total domain difference loss:

$$LOSS_{ddl} = Loss_{mse1} + Loss_{mse2} + Loss_{mse3} \tag{7}$$

NLL losses are used for source and target domain classification losses. The total loss of the supervised domain adaptive acoustic scene classification model is a weighted sum of the domain difference loss, the source domain loss and the target domain loss.

Since the domain difference loss is added to the total loss function, it may be difficult to fit the classification loss by over-optimizing the domain difference during the model training process. In this paper, the algorithm also adopts a pre-training approach, in which the classification loss of the model is trained first, and then the domain classification loss is added to continue the training with a smaller learning rate, so as to better extract domain invariant features.

3 Simulation Experiment and Result Analysis

3.1 A Datasets and Experimental Parameters

To verify the performance of the algorithm proposed in this paper, an experimental environment with Window10 system, RTX3060 graphics card, R7-5800H CPU and 16G RAM was used. Using pytorch as the deep learning framework, the DCASE2020Task1A multi-device sound scene classification dataset from the DCASE competition was used, The dataset contains a total of ten predefined acoustic scene categories. According to the official dataset division, the number of samples in the training set is 13926, where the number of main device A is 10215, and the number of devices B, C, S1, S2, and S3 are all about 750, which are called seen devices, and the number of samples in the test set is 2968, where the number of devices A, B, C, S1, S2, S3, S4, S5, and S6 are all about 330, and devices S4, S5, and S6 do not appear in the training set, which are called unseen device.

In the experiments of this paper, all acoustic scene signals are down sampled to 32000 Hz, Log-Mel spectrograms are extracted in 64 frequency bands, window size is 2048, optimizer is SGD, momentum = 0.9, learning rate is 0.0015, cosine annealing strategy is used for learning rate reduction, and 500 epochs are set for training.

3.2 Comparison of Algorithms

The literature [16] argues that when larger neural networks are used for training, specially designed domain generalization methods usually perform similarly to common models (called empirical risk minimization) and that domain generalization algorithms should be compared on the same data set and the same model to be meaningful, so the experiments in this paper are all compared on the same network model.

The experimental results are shown in Table 1 and Fig. 3. When using only the samples from device A in the training set for training, it can achieve 76.7% accuracy on device A in the test set, but lower accuracy on other untrained devices, indicating that

Table 1. Performance comparison of different methods

Method	Device A	Seen-device (A,B,C,S1-S3)	Unseen-device (S4-S6)	Avg
Only A	0.767	0.32	0.163	0.265
All train	0.68	0.608	0.561	0.593
Base-Line	0.728	0.631	0.372	0.514
FBS	**0.77**	0.677	0.504	0.62
SDA	0.727	0.663	**0.594**	0.643
FBS-SDA	0.726	**0.693**	0.572	**0.653**

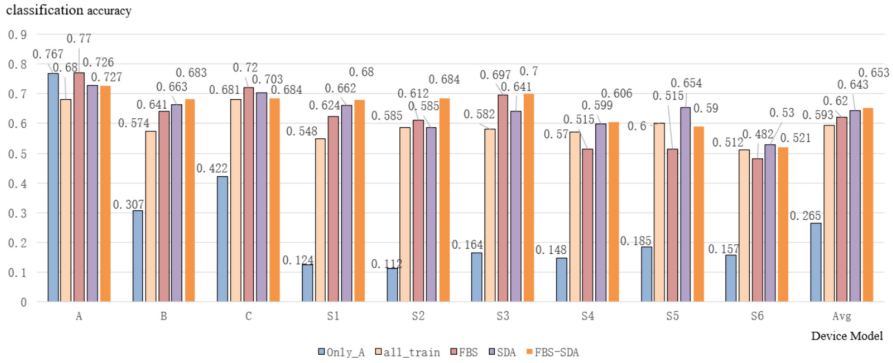

Fig. 3. Detailed performance comparison of different methods

the model trained on a single device does not generalize to other devices for application, while when using all seen devices in the officially divided training set for training, its recognition performance on the main The recognition on device A is reduced by 8% when trained on all seen devices in the officially divided training set, but the recognition performance of the model on other devices is improved due to the participation of more types of devices in the model training, and the model is even generalized to unseen device, which indicates that the participation of more types of devices in the model training can help the model to better extract common features or domain invariant features among different devices.

And after adding the algorithm proposed in this paper, the model showed different performance on seen and unseen device, respectively, and the frequency band standardization(FBS) [6] reached 77% accuracy on the main device A, showing 57 comparable to the training device A alone, and improved 7% accuracy in the seen device A-S3, but suffered to some extent in the unseen device S4–S6, indicating that the frequency band standardization is mainly a linear correction of the frequency axis of the seen device, which can effectively improve the recognition accuracy of the seen device and improve the average classification accuracy by 2% on the test set.

Meanwhile, the supervised domain adaptive method(SDA) [12] improved the accuracy by 4% on device A and achieved some improvement on all other devices, especially

on the unseen device S4–S6, which improved the accuracy by 3%, and the improvement effect is obvious, which indicates that the supervised domain adaptive method actively aligns the features of the source and target domains during the training process and corrects the nonlinear distortion, so that the model can better extract the domain invariant features and enhance the classification accuracy and generalization ability of the model, which improved the average classification accuracy by 4% on the test set.

While combining the frequency band standardization and supervised domain adaptive methods (FBS-SDA), the frequency band standardization method is used to correct linear distortion and the supervised domain adaptive method is used to correct nonlinear distortion. Combining the advantages of these two methods, the resulting trained model improves by 13.9% overall compared to the baseline model, 6.2% on seen devices, and 20% on unseen device.

4 Conclusions

In this paper, a multi-device sound scene algorithm combining frequency band standardization and supervised domain adaptive methods is proposed to solve the device mismatch problem of sound scene classification in multi-device conditions. Since the differences between different kinds of pickup devices are mainly divided into linear and nonlinear distortions, the frequency band standardization method is used to make linear corrections to Log-Mel features by device type, and then the supervised domain adaptive method is combined to reduce the differences between the source and target domains for nonlinear corrections. The experimental results show that the proposed algorithm achieves some improvement on both seen pickup devices and unseen pickup devices compared with the baseline model and other domain generalization algorithms, which can be used to better extract domain invariant features, improve the classification performance of the model, and better generalize the model to unseen device.

Acknowledgements. This research was funded by the National Natural Science Foundation of China (No. 62071135), the Project of Guangxi Natural Science Foundation (No. 2020GXNSFAA159004), the Project of Guangxi Technology Base and Talent Special Project (No. GuiKe AD20159018).

References

1. AbeBer, J., Gotze, M., Kuhnlenz, S., et al.: A distributed sensor network for monitoring noise level and noise sources in urban environments. In: 2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud), pp. 318–324. IEEE (2018)
2. Hao, Z., Zhan, H., Zhang, C., et al.: Assessing the effect of human activities on biophony in urban forests using an automated acoustic scene classification model. Ecol. Ind. **144**, 109437 (2022)
3. Bear, H.L., Heittola, T., Mesaros, A., et al.: City classification from multiple real-world sound scenes. In: 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 11–15. IEEE (2019)

4. Zhang, H., Cisse, M., Dauphin, Y.N., et al.: mixup: Beyond Empirical Risk Minimization (2017). arXiv preprint arXiv:1710.09412
5. Park, D.S., Chan, W., Zhang, Y., et al.: Specaugment: a Simple Data Augmentation Method for Automatic Speech Recognition (2019). arXiv preprint arXiv:1904.08779
6. Kosmider: Calibrating neural networks for secondary recording devices. In: Detection and Classification of Acoustic Scenes and Events, Technical Report (2019)
7. Yang, L., Tao, L., Chen, X., et al.: Multi-scale semantic feature fusion and data augmentation for acoustic scene classification. Appl. Acoust. **163**, 107238 (2020)
8. Mcdonnell, M.D., Gao, W.: Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths. In: Detection and Classification of Acoustic Scenes and Events, Technical Report (2019)
9. Hu, H., Yang, C.H., Xia, X.J., et al.: A two-stage approach to device-robust acoustic scene classification. In: Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), pp. 845–849 (2021)
10. Koutini, K., Eghbal-zadeh, H., Widmer, G.: CP-JKU submissions to DCASE19: acoustic scene classification and audio tagging with receptive field regularized CNNs. In: Detection and Classification of Acoustic Scenes and Events, Technical Report (2019)
11. Ye, M., Zhong, H., Song, X., et al.: Acoustic scene classification using deep convolutional neural network via transfer learning. In: 2019 International Conference on Asian Language Processing (IALP), pp. 19–22. IEEE (2019)
12. Zhao, J., Kong, Q., Song, X., et al.: Feature alignment for robust acoustic scene classification across devices. IEEE Signal Process. Lett. **29**, 578–582 (2022)
13. Olvera, M., Vincent, E., Gasso, G.: On the impact of normalization strategies in unsupervised adversarial domain adaptation for acoustic scene classification. In: ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 631–635. IEEE (2022)
14. Rozantsev, A., Salzmann, M., Fua, P.: Beyond sharing weights for deep domain adaptation. IEEE Trans. Pattern Anal. Mach. Intell. **41**(4), 801–814 (2018)
15. Heittola, T., Mesaros, A., Virtanen, T.: Acoustic Scene Classification in DCASE 2020 Challenge: Generalization Across Devices and Low Complexity Solutions (2020). arXiv:2005.14623
16. Gulrajani, I., Lopez-Paz, D.: In Search of Lost Domain Generalization (2020). arXiv preprint arXiv:2007.01434