SEMI-SUPERVISED ACOUSTIC SCENE CLASSIFICATION UNDER DOMAIN SHIFT USING AN ATTENTION MODULE AND ANGULAR LOSS

Technical Report

Michael Neri, Marco Carli
Department of Industrial, Electronic, and Mechanical Engineering
Roma Tre University
{name}.{surname}@uniroma3.it

ABSTRACT

In this technical report, a semi-supervised solution for the classification of audio signals under domain shift of ICME 2024 Grand Challenge is described. In more detail, a low-complexity attention-based convolutional neural network is introduced for the identification of the scenes that exploit the log-Mel spectrogram and the Wavegram learning-based time-frequency representation. Experimental results on a portion of the challenge development dataset show outstanding performance. Code, model, and pre-trained weights are available at `https://github.com/michaelneri/ICME2024RM3Team`.

***Index Terms*—** Attention, Deep Learning, Wavegram, Low-complexity, Semi-supervised

1. INTRODUCTION

Acoustic scene classification (ASC) is the task that aims at identifying environments from only the sounds they produce [1]. In recent times, ASC has attracted considerable interest due to many practical applications, such as systems for detecting and classifying audio anomalies [2] and tagging of events and music in urban scenarios [3, 4]. Several studies have been conducted to identify audio patterns within this field using deep neural networks (DNNs), which are employed to generate features for classification tasks, as done in [2, 5–9].

However, domain shift is a critical issue in ASC where models trained on one set of audio conditions underperform when tested on different acoustic environments, e.g. trained on advanced recording systems and tested on commercial-off-the-shelf (COTS) devices [10] and viceversa. In [10] the authors proposed an unsupervised domain adaptation method that aligns the first- and second-order statistics of all the frequency bands of target-domain acoustic scenes to the ones of the source-domain training dataset. However, there is a lack of methods that exploit large portions of unlabelled raw data for improving supervised training of deep learning models. A recent study introduced a multi-target domain adaptation technique which focuses on reducing the domain gap by treating domain shift as a measurable distance [11].

To tackle the domain shift issue, we propose a deep learning approach that introduces an attention module and a learned time-frequency representation, namely Wavegram. Then, a multi-iteration fine-tuning (FT) process is devised to train the model on the source domain to improve its generalization ability. Finally, unlabelled data is used in a semi-supervised fashion to refine model's predictions.

The remainder of the work is organized as follows: Section 2 introduces the deep learning architecture and how it is trained. Experimental results are shown in Section 3 whereas the conclusions are drawn in Section 4.

2. PROPOSED METHOD

In this section, the proposed semi-supervised approach for ASC is detailed. The overall architecture is shown in Fig. 1.

2.1. Feature extraction

Initially, a pre-processing stage is employed to extract the complex short-time Fourier transform (STFT) $\mathrm{STFT}\{\mathbf{x}\}$ from the single-channel audio signal $\mathbf{x} \in \mathbb{R}^{1\times l}$, where l is the number of samples that is equal to the duration in seconds multiplied by the sampling frequency f_s. This transform is performed using a Hann window of length 32 ms with 50% overlap. Next, a log-Mel spectrogram $X_{\mathrm{Mel}} \in \mathbb{R}^{t\times f}$ is extracted using a Mel filterbank $\mathrm{H}_{\mathrm{Mel}}(\cdot)$ as follows:

$$X_{\mathrm{Mel}} = 20\log_{10}\mathrm{H}_{\mathrm{Mel}}(\mathrm{STFT}\{\mathbf{x}\}), \tag{1}$$

where t and f denote the number of time and frequency bins, respectively.

2.2. Wavegram

In [12] Wavegram is introduced as a new learned time-frequency representation for audio tagging. In particular, Wavegram is designed to capture time-frequency patterns that

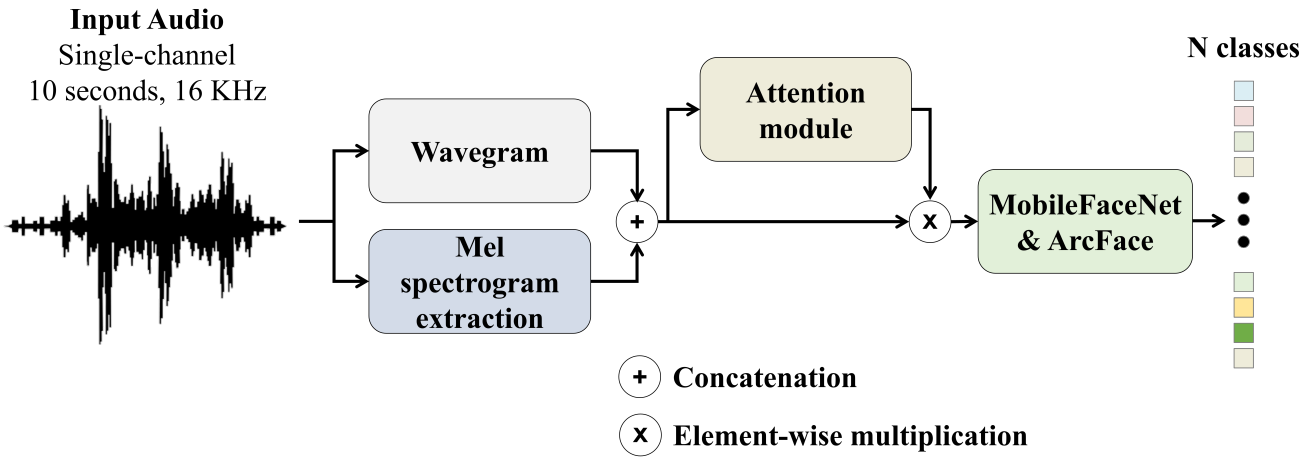

Fig. 1. Proposed approach for semi-supervised classification of ASC.

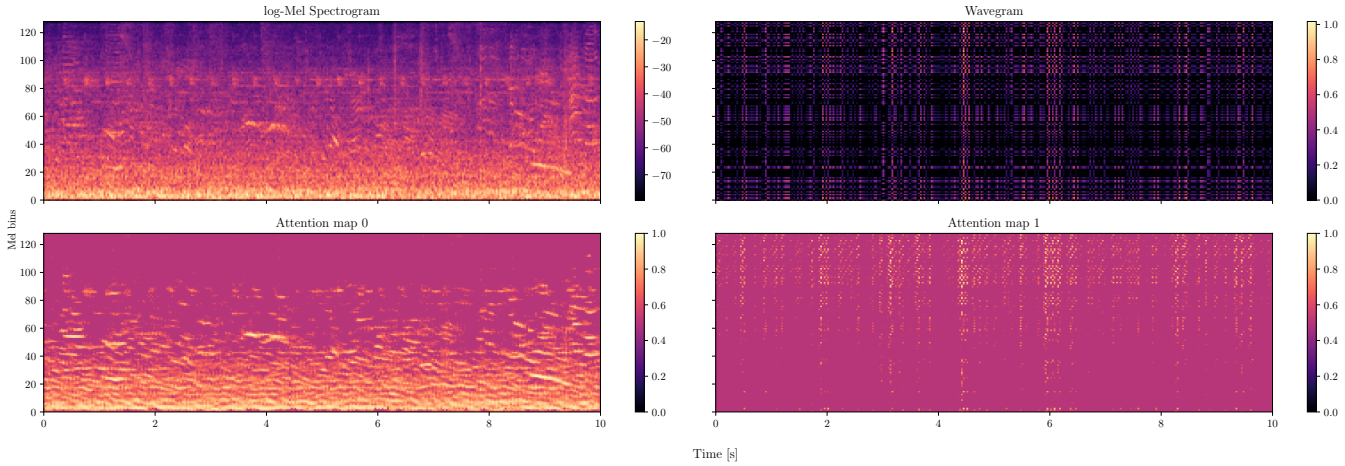

Fig. 2. Example of acoustic features X and corresponding attention maps $H = f_{\mathrm{ATT}}(X)$ for a *Construction site* audio recording. The first row depicts the log-Mel spectrogram and the Wavegram, respectively. The second row shows the attention maps that are element-wise multiplied with the acoustic features to obtain $\tilde{X}$.

are generally lost during the extraction of hand-crafted filterbanks, e.g., Mel spectrograms [12]. Several methods have been based on Wavegram by applying a 1D convolution that acts as a learnable STFT [13–15]. Next, the features have been further processed by layer normalization and 1D convolutions with small kernel sizes [14, 15]. To reduce the computational complexity, in this work, Wavegram consists only of a separable 1D convolutional layer with $f = 128$ filters with 1024 neurons each. To mimic the windows' overlap in the STFT computation, stride and padding of 512 samples are applied. The output of the Wavegram is denoted as $X_{\mathrm{Wave}} \in \mathbb{R}^{t \times f}$.

Finally, the log-Mel spectrogram and the output of Wavegram are concatenated along the channel dimension:

$$X = [X_{\mathrm{Mel}}, X_{\mathrm{Wave}}] \in \mathbb{R}^{t \times f \times 2}. \tag{2}$$

2.3. Attention module

The attention module is devised to construct an attention map $H \in \mathbb{R}^{+t \times f \times 2}$ utilizing both the log-Mel spectrogram and the Wavegram. Its objective is to highlight the most significant regions of features for the task of classification. This module is represented as: $\mathrm{f}_{\mathrm{ATT}} : \mathbb{R}^{t \times f \times 2} \rightarrow \mathbb{R}^{+t \times f \times 2}$. It encompasses two separable convolutional blocks with 16 and 64 filters of size 3×3, sequentially. Following these blocks, a convolutional layer of 1×1, i.e., a projection layer, is applied, and a sigmoid activation function maps each pixel to a probability, thus producing a $t \times f \times 2$ attention map. The enhanced acoustic features $\tilde{X} \in \mathbb{R}^{t \times f \times 2}$ result from the element-wise product ($\otimes$) of the time-frequency representations and the attention map, defined as

$$\tilde{X} = f_{\mathrm{ATT}}(X) \otimes X. \tag{3}$$

An example of log-Mel spectrogram, Wavegram, and their attention maps is depicted in Fig. 2.

2.4. Loss function

The classification layer we employ is ArcFace [16]

$$\mathcal{L}_{\mathrm{AF}}(\boldsymbol{\theta}, \mathbf{y}) = -\mathbf{y}^T \frac{e^{s\cos(\boldsymbol{\theta}+m\mathbf{y})}}{\sum_{i=1}^{c} e^{s\cos(\theta_i+m\hat{y}_i)}}, \tag{4}$$

where the vector of angles $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_c]$ corresponds to each class and is derived by computing $\theta_i = \arccos(\mathbf{w}_i^T \mathbf{h})$, reflecting the correlation between the features classified as $\mathbf{h} \in \mathbb{R}^{h\times 1}$ and the ArcFace weights learned as $\mathbf{w}_i \in \mathbb{R}^{h\times 1}$ for the i-th class. The constants $s \in \mathbb{R}^+$ and $m \in \mathbb{R}^+$ denote the scaling and margin coefficients for the ArcFace loss, respectively.

2.5. Semi-supervised pipeline

Initially, the model is pre-trained on the TUT Mobile dataset [17], similarly to the baseline provided by the organizers [18]. This stage involves adjusting the model's weights to recognize sound patterns and features of the urban scenario.

Then, multiple FT iterations are performed on the labelled development dataset of the Grand Challenge. In this work, a FT iteration consists in removing the last classification layer, i.e., ArcFace [16], keeping untouched the model's weights. This iterative process helps the model adapt to specific tasks, handle class imbalances, and enhance its ability to generalize to new data. It also offers insights for further model refinement, making the final model more suited to real-world applications [19].

Next, the trained model is used to assign *soft labels* to the unlabelled development dataset. Fig. 3 and Fig. 4 depict the histograms of *soft labels* occurrences and confidence scores during inference, respectively.

The latter is then utilized during the final FT with the whole development dataset in the training loss

$$\mathcal{L}(\boldsymbol{\theta}, \mathbf{y}, \alpha) = \alpha \mathcal{L}_{\mathrm{AF}(\boldsymbol{\theta},\mathbf{y})}, \tag{5}$$

where $\alpha \in \mathbb{R}^+$ is the confidence score of the prediction $\boldsymbol{\theta}$ from the single-channel recording $\mathbf{x}$ with one-hot ground truth $\mathbf{y}$. By doing so, the model during FT tends to disregard errors on samples of the unlabelled dataset that have low confidence scores. Finally, the trained model is used to perform inference on the evaluation set.

3. EXPERIMENTAL RESULTS

3.1. Datasets

The 2023 Chinese Acoustic Scene (CAS) [18] dataset is an extensive resource foundational to studies on environmental acoustic scenes, containing 10 scenes with a collective length of more than 130 hours. Each of the dataset's 10-second sound clips is accompanied by metadata detailing its recording location and time. Derived from the CAS 2023, the dataset for the ICME 2024 challenge includes development and evaluation parts. The evaluation part comprises $1,100$ recordings chosen from 12 cities, incorporating 5 cities not previously included to enrich the evaluation process for domain shift scenarios. Due to the nature of the challenge, we randomly split the development dataset into training, validation, and testing sets using a percentage ratio of 80%-10%-10%, respectively.

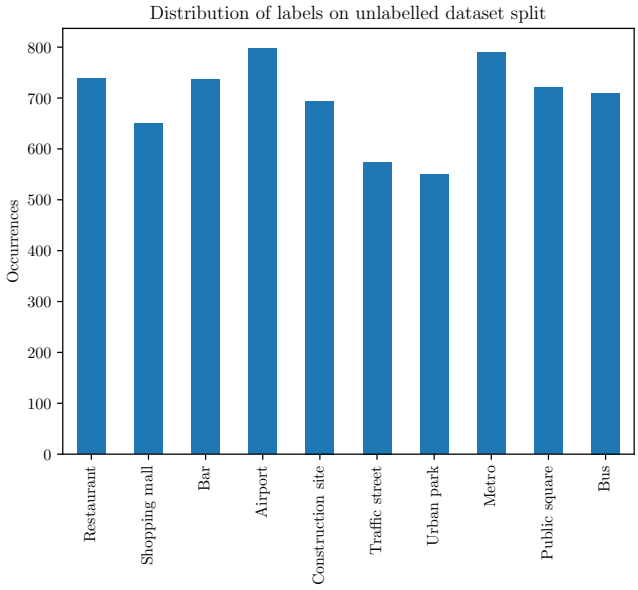

Fig. 3. Distribution of predicted labels in the unlabelled dataset.

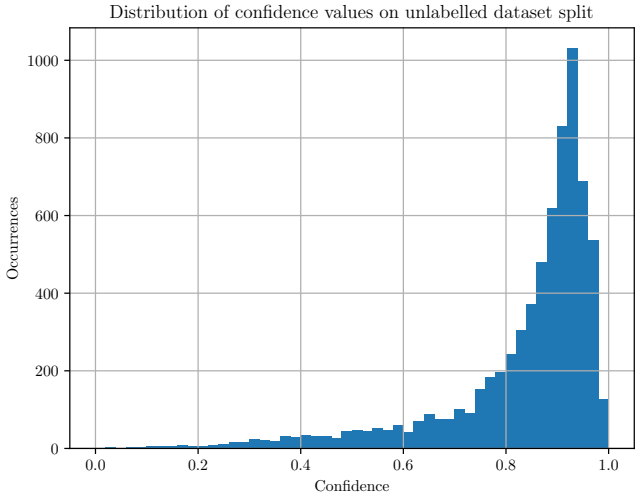

Fig. 4. Distribution of confidence values in the unlabelled dataset.

TAU Urban Acoustic Scenes 2020 Mobile development dataset [17] is used to pre-train the proposed approach. The dataset encompasses recordings from 12 European cities across 10 distinct acoustic scenes, captured with 4 different devices. Moreover, synthetic data was generated for 11 mobile devices, drawing on the original recordings. Among the 12 cities, two are exclusively included in the evaluation set. The overall length of the dataset is of 64 hours. Training, validation, and testing have been carried out following the official splits provided by the organizers of the challenge.

For both CAS 2023 [18] and TAU Urban Acoustic Scene (UAS) 2020 [17], we follow the authors where accuracy is employed to assess the performance of models

$$\text{Acc} = \frac{TP}{TP + FP}, \tag{6}$$

where TP denotes the number of true positives, e.g., correct classifications, whereas FP represents the false positives, e.g., misclassifications.

3.2. Implementation details

In this work, $f_s = 16$ kHz to reduce the computational complexity of the approach. The classifier at the end of the DNN is MobileFaceNet [20]. The number of trainable parameters is 874k, highlighting the low-complexity characteristic of the approach. Regarding the training and FT procedure, the model is trained for 100 epochs with batches of size 32. A cosine annealing learning rate is employed with initial learning rate $\eta_{\max} = 0.001$ with a maximum number of steps $T_{\max} = 100$. Pytorch-Lightning and Weights&Biases are utilized for training and logging, respectively. ArcFace's scale and margin coefficients are set to $s = 8$ and $m = 0.2$, respectively, following [16]. Number of FT iterations on the labelled ICME 2024 development dataset is set to 3, since no improvement on the validation loss has been observed. More implementation details are available at `https://github.com/michaelneri/ICME2024RM3Team`.

3.3. Results on TAU Urban Acoustic Scenes 2020 Mobile

Fig. 5 reports the confusion matrix of the proposed approach on the TAU Urban Acoustic Scenes 2020 Mobile dataset. Overall, the model achieves an average accuracy of 45%, which is consistent with the performance of architectures that are not ensembles of models [17], following the rules of the ICME 2024 Grand Challenge.

3.4. Results on ICME 2024 development dataset

Table 3.4 shows the performance of the proposed approach with several training setups. Training from scratch yields the worst performance with an average accuracy of 63.79%. Instead, pretraining on the TAU Urban Acoustic Scenes 2020 improves the generalization ability of the model.

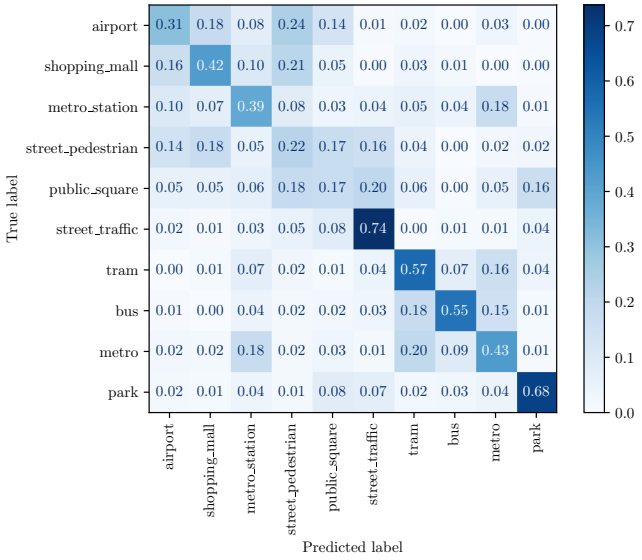

Fig. 5. Confusion matrix of proposed approach on TAU Urban Acoustic Scenes 2020 development dataset.

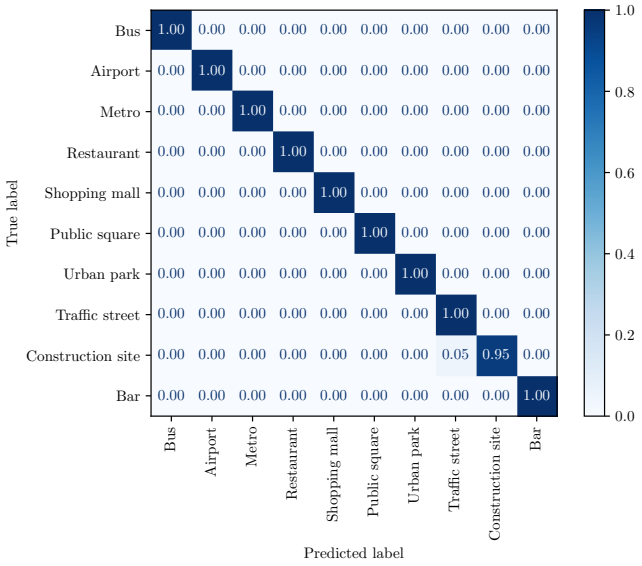

Fig. 6. Confusion matrix of proposed approach on ICME 2024 Grand Challenge test set dataset before exploiting the unlabelled dataset.

Moreover, the multi-iteration FT process further enhances the performance, achieving a remarkable 99.43% of accuracy on the test set, as it can be inspected from the confusion matrix in Fig. 6. With the addition of the unlabelled dataset in the FT, the proposed approach achieves optimal classification performance, showing a diagonal confusion matrix in Fig. 7.

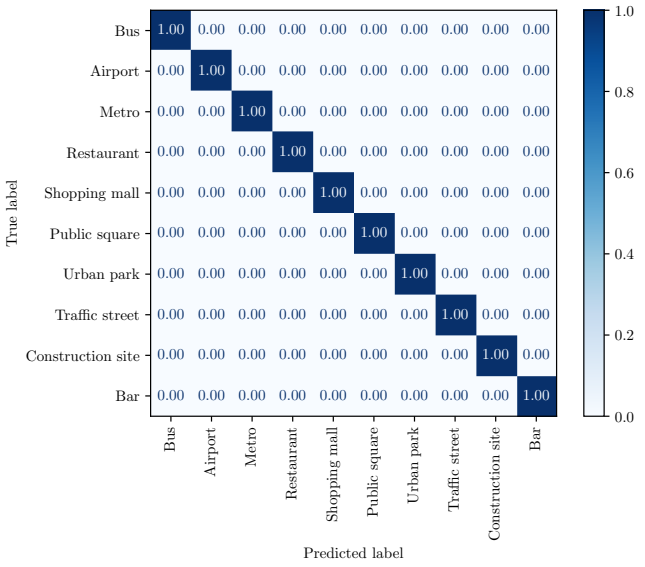

Fig. 7. Confusion matrix of proposed approach on ICME 2024 Grand Challenge test set dataset after exploiting the unlabelled dataset.

Approach	Acc (%)
from scratch	63.79%
1 FT iteration	97.70%
2 FT iterations	98.28%
3 FT iterations	99.43%
3 FT iterations + unlabelled dataset	100%

Table 1. Comparison of several training setup with respect to test accuracy.

3.5. Results on ICME 2024 evaluation dataset

Blank subsection until the announcement of the results.

4. CONCLUSIONS

In this work, a semi-supervised learning approach for ASC that addresses domain shift is proposed for the ICME 2024 Grand Challenge. Thanks to an attention-based convolutional neural network (CNN), a learning-based time-frequency representation, namely Wavegram, and an iterative FT process, our model demonstrated optimal performance on the development dataset of the challenge. *Conclusion adding the results on domain shift will be completed after the announcements.*

5. REFERENCES

[1] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic Scene Classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.

[2] K. Qiuqiang, C. Yin, I. Turab, W. Yuxuan, W. Wang, and M. D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[3] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015.

[4] J. Salamon, C. Jacoby, and J. P. Bello, "A Dataset and Taxonomy for Urban Sound Research," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014.

[5] J. Salamon and J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[6] B. Bahmei, E. Birmingham, and S. Arzanpour, "CNN-RNN and Data Augmentation Using Deep Convolutional Generative Adversarial Network for Environmental Sound Classification," *IEEE Signal Processing Letters*, vol. 29, pp. 682–686, 2022.

[7] H. Park and C. D. Yoo, "CNN-Based Learnable Gammatone Filterbank and Equal-Loudness Normalization for Environmental Sound Classification," *IEEE Signal Processing Letters*, vol. 27, pp. 411–415, 2020.

[8] H. Song, S. Deng, and J. Han, "Exploring Inter-Node Relations in CNNs for Environmental Sound Classification," *IEEE Signal Processing Letters*, vol. 29, pp. 154–158, 2022.

[9] M. Neri, F. Battisti, A. Neri, and M. Carli, "Sound event detection for human safety and security in noisy environments," *IEEE Access*, vol. 10, pp. 134230–134240, 2022.

[10] A. I. Mezza, E. A. P. Habets, M. Müller, and A. Sarti, "Unsupervised Domain Adaptation for Acoustic Scene Classification Using Band-Wise Statistics Matching," in *EUSIPCO*, 2021.

[11] D. Yang, H. Wang, and Y. Zou, "Unsupervised multi-target domain adaptation for acoustic scene classification," *arXiv preprint arXiv:2105.10340*, 2021.

[12] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[13] Y. Liu, J. Guan, Q. Zhu, and W. Wang, "Anomalous Sound Detection Using Spectral-Temporal Information Fusion," in *IEEE ICASSP*, 2022.

[14] H. Chen, L. Ran, X. Sun, and C. Cai, "SW-WAVENET: Learning Representation from Spectrogram and Wavegram Using Wavenet for Anomalous Sound Detection," in *IEEE ICASSP*, 2023.

[15] S. Choi and J Choi, "Noisy-ArcMix: Additive Noisy Angular Margin Loss Combined With Mixup Anomalous Sound Detection," *arXiv preprint arXiv:2310.06364*, 2023.

[16] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in *IEEE/CVF CVPR*, 2019.

[17] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *DCASE*, 2018.

[18] J. Bai, M. Wang, H. Liu, H. Yin, Y. Jia, S. Huang, Y. Du, D. Zhang, D. Shi, W. Gan, M. D. Plumbley, S. Rahardja, B. Xiang, and J. Chen, "Description on IEEE ICME 2024 Grand Challenge: Semi-supervised Acoustic Scene Classification under Domain Shift," *arXiv:2402.02694*, 2024.

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

[20] S. Chen, Y. Liu, X. Gao, and Z. Han, "MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices," in *CCBR*, 2018.