2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW) | 979-8-3503-7981-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICMEW63481.2024.10645362

SEMI-SUPERVISED ACOUSTIC SCENE CLASSIFICATION WITH TEST-TIME ADAPTATION

Wen Huang[1], Anbai Jiang[2], Bing Han[1], Xinhu Zheng[2], Yihong Qiu[3], Wenxi Chen[1], Yuzhe Liang[1],
Pingyi Fan[2], Wei-Qiang Zhang[2], Cheng Lu[3], Xie Chen[1], Jia Liu[2,4], Yanmin Qian[1†]

[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[2]Department of Electronic Engineering, Tsinghua University, Beijing, China
[3]School of Economics and Management, North China Electric Power University, Beijing, China
[4]Huakong AI Plus, Beijing, China

ABSTRACT

Acoustic Scene Classification (ASC) plays a crucial role in audio signal processing, with applications ranging from urban soundscapes to smart homes. However, challenges like domain shift and scarce labeled data hinder its development, highlighting the need for semi-supervised learning strategies. In the context of ICME 2024 Grand Challenge, aimed at the semi-supervised acoustic scenes classification under domain shift, our endeavor has been to devise a system that navigates these challenges. Our submission outlines a semi-supervised ASC system that employs pretraining on available datasets, followed by finetuning through FixMatch and pseudo-labeling, and concludes with test-time adaptation. This approach seeks to effectively utilize unlabeled data and mitigate domain shift, ultimately enhancing the ASC system's performance. Our final entry achieved a third-place position with a macro accuracy rate of 70.0% on the evaluation set.

***Index Terms*—** acoustic scene classification, semi-supervised learning, domain shift, test-time adaptation

1. INTRODUCTION

Acoustic Scene Classification (ASC) is a pivotal task in the realm of audio signal processing, aiming to categorize audio recordings into predefined scenes based on their acoustic characteristics. This technology underpins numerous applications, from enhancing urban soundscapes to advancing smart home devices, making its development a focal point for researchers and technologists alike.

However, as ASC research advances, it confronts significant obstacles. Challenges such as domain shift significantly influence ASC, where discrepancies in acoustic properties between training and testing scenarios can degrade model performance [1]. The issue of domain shift encompasses a variety of factors such as recording devices, environmental conditions, and language or culture differences, making it particularly challenging to address. Numerous studies [2, 3, 4] have explored various domain adaptation strategies, such as domain alignment and feature disentanglement, to counteract discrepancies arising from device or city mismatches within the TAU urban acoustic scene dataset [1]. Nevertheless, these techniques often depend on prior knowledge of domain characteristics, posing challenges when the domain information is unknown or cannot be precisely defined.

Furthermore, due to the high cost of annotating acoustic scenes, the limited availability of labeled data poses a significant challenge for supervised learning approaches. This scarcity underscores the importance of semi-supervised methods, which leverage the abundant unlabeled audio data, offering a practical solution to this constraint. In computer vision, semi-supervised learning serves as a powerful intermediary between supervised and unsupervised learning. It uti-

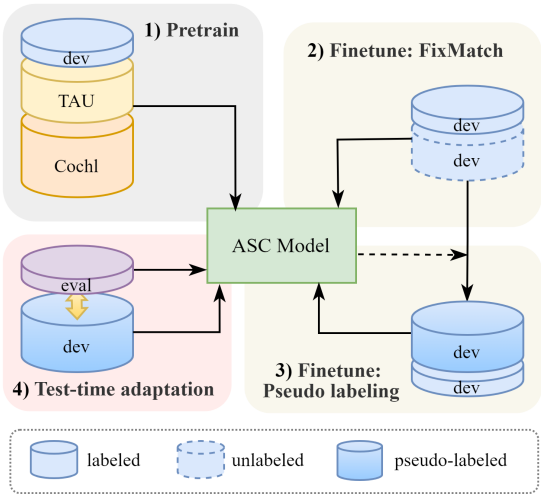

Fig. 1. Overview of our system's pipeline. "dev" and "eval" refer to the development and evaluation sets of the competition, respectively, while "TAU" and "Cochl" denote the TAU Urban Acoustic Scenes and CochlScene datasets.

† Yanmin Qian is the corresponding author

979-8-3503-7981-5/24/$31.00 ©2024 IEEE

lizes techniques like pseudo labeling, Mean Teacher [5], MixMatch [6], and FixMatch [7], which meld consistency regularization, data augmentation, and the strategic incorporation of unlabeled data to enhance model robustness and performance. These approaches are not only effective in visual contexts but also demonstrate significant potential for adaptation in audio processing.

The aforementioned challenges also form the core of the "Semi-supervised Acoustic Scene Classification under Domain Shift" challenge [8]. Within this context, the challenge provides a development dataset from the CAS 2023 collection, featuring 4.8 hours of labeled and 19.3 hours of unlabeled data. A significant domain shift characterizes the evaluation dataset, which includes recordings from cities not present in the development dataset. Importantly, the dataset only provides basic information such as the scene category and filename for each audio clip, presenting additional hurdles in managing and understanding domain-specific information.

To navigate these obstacles, we propose a semi-supervised ASC system. Figure 1 illustrates our method, which unfolds in four steps. Initially, we pretrain the model using a variety of available datasets. Next, we finetune it on the challenge development dataset employing the FixMatch [7] strategy. As the model acquires knowledge and becomes accustomed to the development set, we generate pseudo labels for the remaining unlabeled data and further finetune the model using these labels. Lastly, in the testing phase, due to the inability to use the evaluation set for training and the unavailability of domain information, we opt for a test-time adaptation [9] method to mitigate the domain shift between the development and evaluation sets. This strategic approach enables our system to effectively mitigate the challenges of domain shift and label scarcity, thereby enhancing the predictive accuracy for acoustic scene classification.

2. DATASETS

In compliance with the challenge rules, apart from the ASC challenge development dataset [8], we utilize the TAU Urban Acoustic Scenes (UAS) 2020 Mobile development dataset [1] and the CochlScene dataset [10] for model pretraining. These are the only two additional datasets permitted for use in this challenge.

Challenge dataset. The ICME ASC challenge development set comprises approximately 24 hours of audio, encompassing a total of 8,700 recordings sourced from eight different cities. These recordings span across 10 distinct acoustic scene classes. Notably, only a fraction of the data, amounting to 20%, comes with associated scene labels.

TAU UAS. The TAU Urban Acoustic Scenes 2020 Mobile dataset [1] features 64 hours of recordings from various European cities across ten acoustic scenes, captured simultaneously using four devices (A, B, C, and D). Additionally, it includes synthetic recordings from devices S1-S11, created by simulating audio from device A, a high-quality binaural recorder, to enhance the dataset's diversity.

CochlScene. The Cochl Acoustic Scene Dataset [10], also known as CochlScene, is an acoustic scene dataset with recordings entirely sourced from crowdsourcing participants in Korea. By selecting a subset pertinent to Acoustic Scene Classification (ASC) from the full collection, it has 76,115 ten-second audio files across 13 different acoustic scenes, contributed by 831 participants.

3. METHOD

3.1. Data augmentation

For model training, we primarily employ three data augmentation techniques: SpecAugment [11], Mixup [12] and Freq-MixStyle [13, 14].

SpecAugment. SpecAugment [11] was initially crafted for speech data improvement, and can also enhances audio by applying frequency and time masking to log mel spectrograms. It randomly hides frequency bins and time segments, thereby increasing model robustness to frequency and temporal variations. This dual-masking approach effectively guards against audio distortions.

Mixup. Mixup [12] creates new dataset entries by blending the inputs and targets of two audio clips. Given two audio inputs x_1 and x_2 with their corresponding targets y_1 and y_2, the augmented input x and the target y are formed as $x = \lambda x_1 + (1 - \lambda)x_2$ and $y = \lambda y_1 + (1 - \lambda)y_2$, with λ being drawn from a Beta distribution. Typically, this technique is applied to the log mel spectrogram of the audio clips from one batch.

Freq-MixStyle. Freq-MixStyle (FMS) [13, 14] is an adaptation of the original MixStyle [15] concept but tailored for frequency. It first normalizes the frequency bands within a spectrogram, then reintroduces variability by denormalizing them using the combined frequency statistics from two different spectrograms. The application of FMS to any given batch occurs with a probability determined by the hyperparameter p_{FMS}, with mixing coefficients drawn from a Beta distribution shaped by α.

3.2. Pretraining ASC Model

3.2.1. Network architecture

Our ASC model employs the CNN10 configuration from PANNs [16], adapted for the audio tagging task. This architecture consists of 10 layers, including 4 convolutional blocks. Each block contains 2 convolutional layers with 3x3 kernels. Batch normalization is incorporated between convolutional layers to enhance training efficiency and stability,

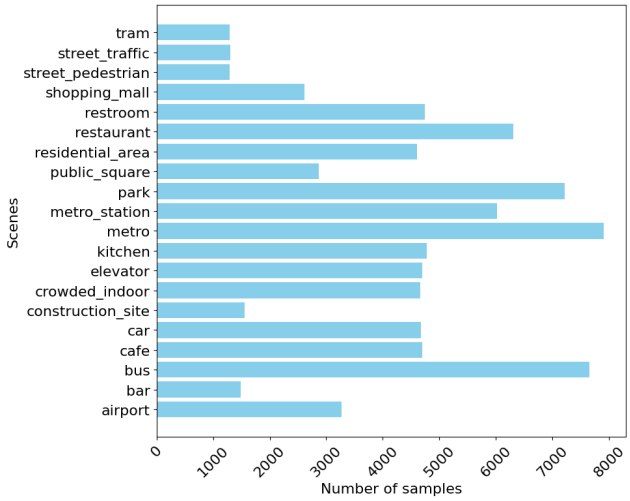

Fig. 2. The count of samples per scene in the newly generated pre-training dataset, adjusted by a weighted sampling strategy. Each scene's sample count is multiplied by its corresponding dataset weighted ratio to reflect the strategy's impact on the dataset composition.

along with the ReLU activation function. For downsampling, average pooling with a 2x2 kernel size is applied within each convolutional block. The model consists of 6.037M parameters in total.

3.2.2. *Training strategy*

To train the ASC model, we use data from the challenge development set, TAU, and CochlScene. These datasets vary in both classes and quantities, requiring us to reorganize them. We combine identical classes from each dataset and introduce new ones, resulting in a total of 20 classes. To ensure each dataset contributes equally, we apply weighted sampling for data from the three datasets, setting the weights at a 10:1:1 ratio.

Despite the adjustments, as shown in Figure 2, disparities in the number of audio clips among various scene classes persist. Such variation can lead to overfitting in classes with an abundance of training clips and underfitting in those with fewer. To mitigate these issues, we implement a strategy that ensures an equal representation of audio clips from all sound classes in each minibatch.

For additional robustness, our training includes data augmentations like SpecAugment, Mixup, and Freq-MixStyle, improving the model's performance across various acoustic scenes.

3.3. Two-Stage Finetuning

After pretraining, we finetune the model on the challenge development dataset in two stages. In the first stage, we use FixMatch, a semi-supervised algorithm, to finetune the model with both labeled and unlabeled data. In the second stage, we generate pseudo labels for all unlabeled data using the stage 1 model. Then, we finetune the model further using either labeled data or data with these pseudo labels.

3.3.1. *Stage 1: FixMatch*

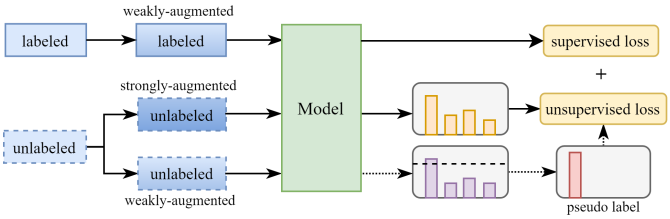

Fig. 3. Diagram of the FixMatch algorithm.

Figure 3 illustrates that during each training step of FixMatch [7]. For every iteration, a data batch is retrieved, encompassing both the labeled data, denoted as $\mathcal{X} = \{(x_i, p_i) : i \in (1, \ldots, N)\}$, and the unlabeled data, represented by $\mathcal{U} = \{u_j : j \in (1, \ldots, M)\}$. Here, x_i and u_i signify the labeled and unlabeled training instances, respectively, while p_j stands for the one-hot encoded label corresponding to x_i. Furthermore, N and M indicate the counts of labeled and unlabeled instances in the current batch, respectively. We apply two distinct augmentation strategies to the data: weak augmentation, symbolized by $\alpha(\cdot)$, and strong augmentation, represented by $\mathcal{A}(\cdot)$. In our framework, SpecAugment is utilized for the weak augmentation. To constitute the strong augmentation, we extend SpecAugment by integrating an additional technique, namely Freq-MixStyle.

Training comprises the computation of two varieties of cross-entropy loss: supervised, denoted as L_s, and unsupervised, denoted as L_u. The supervised loss, L_s, is determined by calculating the conventional cross-entropy for weakly augmented labeled data, where $p_m(y|x)$ represents the model's predicted class distribution:

$$L_s = \frac{1}{N} \sum_{i=1}^{N} \mathrm{H}(p_i, p_m(y \mid \alpha(x_i))) \tag{1}$$

In the calculation of the unsupervised loss, L_u, we commence by obtaining the predicted class distribution from the model for the weakly augmented data, indicated as q_j. If these predictions exceed the confidence threshold τ, they are utilized as pseudo labels, $\hat{q}_j$, for the strongly augmented data in the computation of the cross-entropy loss. This methodology ensures alignment between the representations derived from weakly and strongly augmented data, thereby reinforcing the model's robustness and consistency across varied augmenta-

tions.

$$L_u = \frac{1}{M}\sum_{j=1}^{M} \mathbb{1}(\max(q_j) \geq \tau)\mathrm{H}(\hat{q}_j, p_m(y \mid \mathcal{A}(u_j))) \quad (2)$$

Thus, the total loss for this stage is defined as follows:

$$L = L_s + \lambda_u L_u \quad (3)$$

Here, λ_u represents the weighting factor for the unsupervised loss component, which, in our training regimen, is assigned a value of 0.5.

3.3.2. Stage 2: Pseudo Labeling

During the initial stage of fine-tuning, the number of unlabeled sample predictions chosen as pseudo labels increases as the model's performance improves. Consequently, in the second stage, we assume the model has developed the capability to accurately predict labels for unlabeled data. Therefore, we utilize the model from stage 1 to generate pseudo labels for the remaining unlabeled training samples, followed by fine-tuning the model on this newly labeled dataset. Additionally, to elevate the challenge of this stage, we apply the strong augmentation techniques, initially used in stage 1, across all data.

3.4. Test-time Adaptation

A test-time adaptation method [9] based on k-nearest neighbor (KNN) is adopted to bridge the gap between the development and the evaluation sets. The embeddings of all labeled samples of the development set are pre-extracted to form a memory bank for KNN. During inference, the embedding of each query sample is compared with the memory bank via cosine similarity, and the distances to top-k neighbors are utilized as the scoring coefficient. Specifically, let $\mathcal{M}_L$ denote the set of embeddings of all labeled samples in the development set. For each query embedding x_i, we search $\mathcal{M}_L$ for a subset of top-k neighbors $N_{\mathcal{M}_L}(x_i)$ by means of cosine similarity:

$$w_{ij} = \frac{x_i^T x_j}{\|x_i\|_2 \|x_j\|_2} \quad (4)$$

Then the final prediction of the model can be given by:

$$\eta(x_i) = \text{Softmax}\left(\sum_{x_j \in N_{\mathcal{M}_L}(x_i)} w_{ij}\mathbf{1}\{y_j\}\right) \quad (5)$$

where y_j is the label of x_j, and $\mathbf{1}\{y_j\}$ denote the one-hot vector of x_j. It is noted that the adopted method neither fine-tunes the model on the evaluation set, nor utilizes the statistics of the evaluation set, which is in compliance with the challenge rules.

Table 1. The performance of the proposed ASC system for each scene class on the evaluation set.

Scnene	Accuracy (%)
Bus	82.0
Airport	84.0
Metro	96.0
Restaurant	74.0
Shopping mall	53.0
Public square	34.0
Urban park	59.3
Traffic street	75.0
Construction site	75.0
Bar	95.0
Average	70.0

4. EXPERIMENT

4.1. Experimental settings

Data preprocessing. In our data preprocessing pipeline, we standardize audio files to a sample rate of 44,100 Hz. The process involves generating spectrograms using a Hann window of 1024 with a 320 hop size and an FFT window of 2048. These spectrograms are then converted into log-Mel spectrograms with 64 Mel bins, ranging from 10 Hz to half the sample rate.

Data augmentation. In our augmentation strategy, we use SpecAugment with settings of 64 for time drop width, 2 for time stripes, 8 for frequency drop width, and 2 for frequency stripes. Mixup is applied with an alpha of 1.0. For Freq-MixStyle, we set a probability of 0.5 and an alpha of 0.6.

Training details. In our training process, we first pretrain the model using the Adam optimizer with a learning rate of 0.001 and binary cross-entropy loss to manage shifted target labels caused by Mixup. For finetuning, we maintain the Adam optimizer but shift to cross-entropy loss. The threshold for finetuning stage 1 is set as 0.5.

Inference details. In our test-time adaptation process, the number of neighbors k is selected as 33.

4.2. Result

Table 1 illustrates the performance of the proposed Acoustic Scene Classification (ASC) system across a range of environments. The system demonstrates exceptional accuracy in "Metro" (96.0%) and "Bar" (95.0%) scenes, highlighting its capability to accurately identify the distinct acoustic features specific to these settings. On the other hand, lower accuracies in "Public square" (34.0%) and "Shopping mall" (53.0%) indicate difficulties in accurately classifying environments that

may have a wider range of or more ambiguous acoustic characteristics. Despite these challenges, the system achieves an overall average accuracy of 70.0%, signifying its general efficacy. However, this also signifies room for improvement, especially in enhancing the system's ability to classify scenes with lower accuracies more reliably.

5. CONCLUSION

In this work, we describe our submission to ICME 2024 Grand Challenge "Semi-supervised Acoustic Scene Classification under Domain Shift". To overcome the domain shift and label scarcity challeges, we develop a semi-supervised ASC system. Our methodology involved pretraining on various datasets, finetuning with FixMatch, generating pseudo labels for further refinement, and employing test-time adaptation to alleviate the domain shift for evaluation. Our final submission achieved a third-place position with a macro accuracy rate of 70.0% on the evaluation set.

6. ACKNOWLEDGEMENT

This work was supported in part by China NSFC projects under Grants 62122050 and 62071288, in part by Shanghai Municipal Science and Technology Commission Project under Grant 2021SHZDZX0102.

7. REFERENCES

[1] Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen, "Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions," *arXiv preprint arXiv:2005.14623*, 2020.

[2] Seongkyu Mun and Suwon Shon, "Domain mismatch robust acoustic scene classification using channel information conversion," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 845–849.

[3] Truc Nguyen, Franz Pernkopf, and Michal Kosmider, "Acoustic scene classification for mismatched recording devices using heated-up softmax and spectrum correction," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 126–130.

[4] Yizhou Tan, Haojun Ai, Shengchen Li, and Mark D Plumbley, "Acoustic scene classification across cities and devices via feature disentanglement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[5] Antti Tarvainen and Harri Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[6] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *Advances in neural information processing systems*, vol. 32, 2019.

[7] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in neural information processing systems*, vol. 33, pp. 596–608, 2020.

[8] Jisheng Bai, Mou Wang, Haohe Liu, Han Yin, Yafei Jia, Siwei Huang, Yutong Du, Dongzhe Zhang, Mark D Plumbley, Dongyuan Shi, et al., "Description on ieee icme 2024 grand challenge: Semi-supervised acoustic scene classification under domain shift," *arXiv preprint arXiv:2402.02694*, 2024.

[9] Yifan Zhang, Xue Wang, Kexin Jin, Kun Yuan, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan, "Adanpc: Exploring non-parametric classifier for test-time adaptation," in *International Conference on Machine Learning*. PMLR, 2023, pp. 41647–41676.

[10] Il-Young Jeong and Jeongsoo Park, "Cochlscene: Acquisition of acoustic scene data using crowdsourcing," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2022, pp. 17–21.

[11] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[12] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[13] Florian Schmid, Shahed Masoudian, Khaled Koutini, and Gerhard Widmer, "Knowledge distillation from transformers for low-complexity acoustic scene classification.," in *DCASE*, 2022.

[14] Byeonggeun Kim, Seunghan Yang, Jangho Kim, Hyunsin Park, Juntae Lee, and Simyung Chang, "Domain generalization with relaxed instance frequency-wise normalization for multi-device acoustic scene classification," *arXiv preprint arXiv:2206.12513*, 2022.

[15] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang, "Domain generalization with mixstyle," *arXiv preprint arXiv:2104.02008*, 2021.

[16] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark 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.