IMPROVING ACOUSTIC SCENE CLASSIFICATION VIA SELF-SUPERVISED AND SEMI-SUPERVISED LEARNING WITH EFFICIENT AUDIO TRANSFORMER

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

2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW) | 979-8-3503-7981-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICMEW63481.2024.10645355

ABSTRACT

In response to the challenges posed by the abundance of unlabeled acoustic scene data in the real world, along with the domain differences in acoustic scenes, the ICME 2024 Grand Challenge has introduced the task of "Semi-supervised Acoustic Scene Classification under Domain Shift." To tackle this issue, we propose a multi-stage semi-supervised framework that utilizes the self-supervised learning (SSL) model – Efficient Audio Transformer (EAT) and the self-learning fine-tuning method. This framework employs self-supervised learning to train on a wealth of unlabeled acoustic scene data, thereby obtaining the capability of extracting audio representations. It then leverages semi-supervised fine-tuning with pseudo-labels and utilizes a test-time adaptation strategy to optimize inference. Our approach achieved a Macro-accuracy of 0.752 across ten categories on the final evaluation dataset, ranked second, only 0.006 lower than the first-place system.

Index Terms— Acoustic Scene Classification, Self-supervised Learning, Semi-supervised Learning, Test-time Adaptation

1. INTRODUCTION

The realm of audio scenes is remarkably broad, encompassing a diverse range of recording devices, geographical areas, and the varied cultural and linguistic contexts they represent. This diversity poses a significant challenge in gathering extensive datasets for training an Acoustic Scene Classification (ASC) model[1]. Relying on limited labeled data can hinder the model's ability to generalize effectively when applied in real-world scenarios. Furthermore, there has been a rising need to deploy ASC applications on mobile devices, which impose constraints on model size, adding another layer of complexity to the development of efficient and robust ASC systems.

In practice, obtaining labeled acoustic scene data is challenging due to the high costs and low accuracy of manual labeling[2], as humans struggle to discern subtle differences in audio. Meanwhile, unlabeled acoustic scene data is more readily available. Therefore, finding ways to develop effective models with limited samples or utilizing unlabeled data to enhance a model's ability to adapt across different scenarios is crucial. While self-supervised learning and supervised learning dominate the field of acoustic scene classification, semi-supervised techniques[3, 4] that leverage unlabeled data are gaining attention for their potential to improve model training and generalization. Addressing these challenges has become a key focus for researchers and practitioners in the field.

To tackle the issue of limited labeled data in acoustic scene classification, we propose a multi-stage training framework that capitalizes on the strengths of self-supervised learning (SSL). Initially, we employ the advanced SSL model, the Efficient Audio Transformer (EAT) [5], leveraging its effectiveness and efficiency on audio SSL. This approach allows us to harness the intrinsic structure of audio data by reconstructing the representation of the audio spectrogram, thereby enhancing the model's generalization capabilities.

Upon completing the pre-training stage, the model gains a foundational understanding of audio representation. To further tailor the model to the specific domain of the dataset provided for the competition, we proceed with fine-tuning using labeled samples. This step aims to address the domain shift problem, making the model more adept at handling the dataset's unique characteristics.

Additionally, we incorporate fine-tuning using labeled data from the TAU dataset [6] to boost the model's generalization performance. We employ weighted sampling techniques during this process to maintain label balance and prioritize the target dataset's relevance.

After fine-tuning, the system preliminarily acquires the capability to accurately label the target dataset. To optimally utilize the abundance of unlabeled data, we implement a pseudo-labeling strategy. This involves generating a probability distribution for the unlabeled data and setting a high probability threshold to extract reliable pseudo-labels for fur-

ther iterative training. This comprehensive approach seeks to maximize the model's adaptability and effectiveness in acoustic scene classification with limited labeled data.

Also, we incorporate a test-time adaptation strategy[7] into our ASC system. During inference, the model explicitly compares the features of testing data with the aggregated knowledge of the training data before outputting the predictions. This adaptation process allows the model to account for the domain shift and improves the robustness of the ASC model.

2. DATA PREPROCESSING AND AUGMENTATION

2.1. Datasets

In compliance with the challenge rules, we utilize the ASC challenge development dataset [8], TAU Urban Acoustic Scenes (UAS) 2020 Mobile development dataset [6] and the CochlScene dataset [9] for model pre-training. No more datasets are used for model training.

The Chinese Acoustic Scene Dataset. The CAS 2023 dataset stands as a comprehensive resource for exploring environmental acoustic scenes, crafted through the collaborative efforts of the Joint Laboratory of Environmental Sound Sensing at Northwestern Polytechnical University's School of Marine Science and Technology. This expansive dataset encompasses a total of 10 distinct acoustic scenes, culminating in over 130 hours of audio recordings, each with segments of 10 seconds. The dataset for the competition was derived entirely from the CAS 2023 dataset, and the development dataset was approximately 24 hours long, of which 20% are labeled, including recordings from 8 cities. In the evaluation dataset, data were selected from 12 cities, with a special selection of 5 unseen cities.

TAU UAS. The TAU Urban Acoustic Scenes 2020 Mobile dataset [6] consists of 64 hours of recordings from various acoustic scenes. The recordings are captured in different cities across Europe, using four devices (A, B, C, and D) simultaneously. To enhance the diversity of the dataset, 11 simulated devices (S1 - S11) are created in the dataset, using synthetic recordings simulated from device A.

CochlScene. The Cochl Acoustic Scene Dataset [9], abbreviated as CochlScene, is an acoustic scene dataset containing 76,115 ten-second audio files from 13 different acoustic scenes. The recordings of the dataset are sourced from crowdsourcing participants in Korea and manually selected to enhance evaluation reliability.

2.2. Feature extraction

Utilizing spectrograms for various audio tasks has proven to be effective[10, 11]. In our data preprocessing pipeline, all audio files are standardized to a sample rate of 16,000 Hz. Spectrograms are then generated using a Hanning window of 25 milliseconds with a hop size of 10 milliseconds and an FFT window of 400. Subsequently, these spectrograms are transformed into 128-dimensional log-mel spectrograms.

2.3. Data augmentation

For model fine-tuning, the following data augmentation techniques are mainly employed: SpecAugment [12], Mixup [13] and Roll Augmentation.

SpecAugment. SpecAugment [12] directly modifies neural network inputs by warping features, masking frequency channels, and masking time steps. This technique, originally developed for speech recognition, helps improve model performance by creating diverse training examples.

Mixup. Mixup [13] interpolates between pairs of inputs and their labels to generate additional samples. This approach encourages the model to learn more robust features and improves generalization to unseen data. In our pipeline, Mixup enriches the dataset by creating new samples from log-mel spectrograms of audio clips.

Roll Augmentation. Roll Augmentation increases audio data diversity by rolling the audio signal and splicing segments. This technique generates samples with time-domain variations, reducing overfitting and enhancing the robustness and accuracy of acoustic scene classification models.

3. METHOD

In our study, we utilized the EAT model, a self-supervised learning framework specifically designed for audio, as the foundation for tasks related to audio scene classification. Initially, the model was pre-trained using three principal datasets: the ASC Challenge Development Dataset, the TAU Urban Acoustic Scenes Development Dataset, and the CochlScene Dataset. This pre-training stage aimed to capture a broad and generalized spectrum of audio scene representations. Subsequently, to enhance the model's performance on a validation set drawn from the ASC Challenge Development Dataset, we implemented a semi-supervised strategy—self-learning. This approach combines iterative fine-tuning with pseudo-labeling, demonstrating the effectiveness of integrating self-supervised and semi-supervised methodologies to improve audio scene classification.

3.1. Self-supervised Pre-training

Leveraging unlabeled data allows for capturing both low-level acoustic events and rich semantic information from raw audio waves or their spectrograms. This approach is crucial for understanding complex acoustic environments, enabling the extraction of nuanced features often overlooked by traditional supervised methods. Relying solely on supervised datasets can lead to overfitting and poor generalization due to their limited scope and diversity.

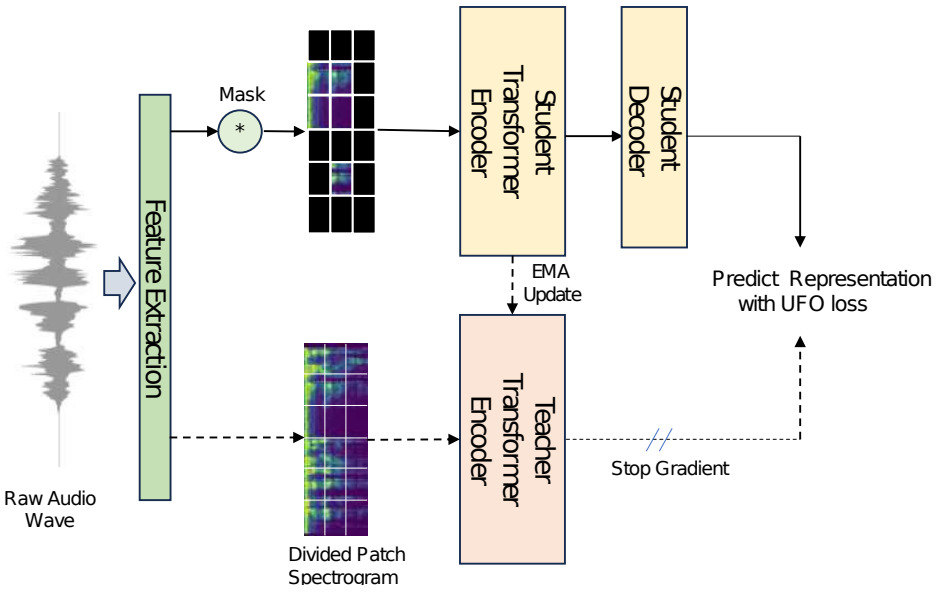

Fig. 1. **Architecture of EAT in Self-supervised Pre-training with Acoustic Scene Audio.** EAT employs a bootstrap framework, iteratively learning acoustic scene features through a self-teaching method. The student model updates using UFO loss, while the teacher model is refreshed through exponential moving average (EMA) updates.

To address the problem of limited labeled data, audio self-supervised learning (SSL) models use pretext tasks like masked autoencoders (MAE)[14] for pre-training, leveraging vast amounts of unlabeled data to learn audio features across various scenes and devices. This pre-training enables superior performance in downstream acoustic scene classification tasks. We employed the EAT model, which utilizes a bootstrap self-supervised training paradigm within the audio domain. The Transformer-based EAT model uses the Utterance-Frame Objective (UFO) as a loss function, integrating global utterance-level and local frame-level losses to predict audio scene representations.

The EAT framework employs a teacher-student structure where the teacher model has access to full spectrograms while the student model sees a masked version. This setup allows the student to learn from the teacher's representations[15]. The model dynamically refines the teacher using exponential moving average (EMA) adjustments based on the student's updates, ensuring efficient learning and superior performance in acoustic scene classification.

In our experiments, we pre-trained the EAT framework on the ICME ASC challenge dataset, TAU, and CochlScene datasets. We used a weighted pre-training approach with a 1:1:10 ratio, prioritizing the ASC data. This strategy enhanced the model's adaptability and performance in target acoustic scene classification tasks, demonstrating the efficacy of a fine-tuned, self-supervised learning approach in audio scene analysis.

3.2. Semi-supervised Learning

Given the limited labeled data in the ASC Development Dataset (1,740 labeled instances versus 6,960 unlabeled instances), relying solely on supervised learning poses challenges in enhancing the model's generalization capabilities. To address this, we implemented a self-learning-based semi-supervised learning method that leverages the abundant unlabeled data in both the TAU dataset and the ASC Development Dataset.

Our semi-supervised learning method unfolds iteratively in two main stages: fine-tuning and pseudo-labeling. Initially, the pre-trained EAT model is fine-tuned on the labeled ASC data using the standard cross-entropy loss function:

$$L = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \tag{1}$$

After fine-tuning, the EAT model predicts the classes of the unlabeled data from both the TAU and ASC Development datasets. A confidence threshold is applied to generate pseudo-labels, ensuring that only predictions above this threshold are retained as hard labels. This approach enhances the quality of pseudo-labels, creating an augmented dataset that combines both labeled data and high-confidence pseudo-labeled data.

The model is then re-trained on this augmented dataset, adjusting and improving based on the broader set of examples, including pseudo-labeled data. This iterative cycle of pseudo-labeling and fine-tuning continues, with each iteration

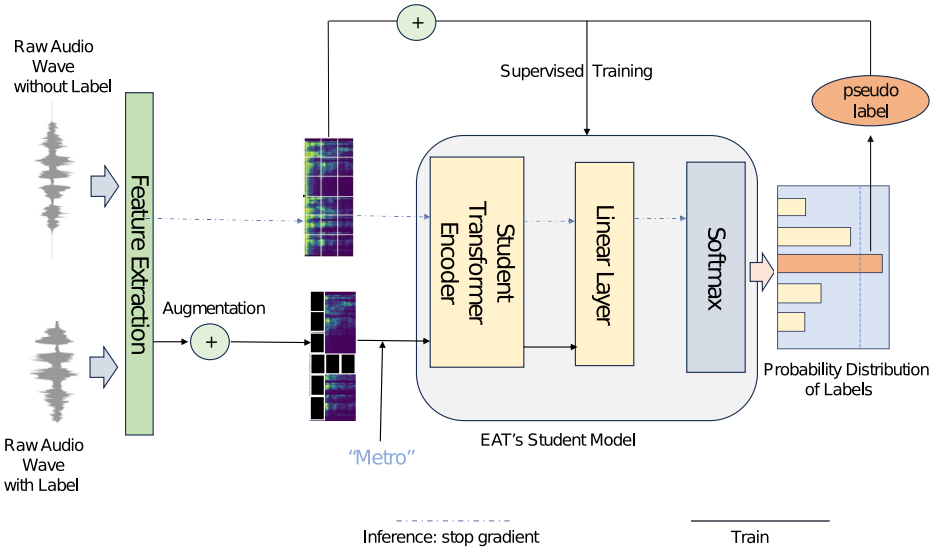

Fig. 2. **Architecture of EAT in Semi-supervised Learning with Labeled and Unlabeled Data.** Within the semi-supervised framework, labeled data undergo data augmentation before being used to fine-tune the student model. Unlabeled data is first subjected to inference, with softmax applied to obtain a probability distribution for the labels. If these probabilities exceed a certain threshold, the model is then fine-tuned through supervised training.

aiming to enhance the model's generalization ability by leveraging insights from the expanded training dataset.

To improve initialization parameters, we iteratively trained models that had already been fine-tuned, carefully controlling the number of iterations to avoid reinforcing incorrect labels. This strategy helps refine the model's performance by continually incorporating high-quality pseudo-labels into the training process.

3.3. Test-time Adaptation

To address the domain shift between the development and evaluation datasets, our approach employs a test-time adaptation (TTA) strategy using the k-nearest neighbor (KNN) methodology, inspired by Zhang et al. [16]. We start by extracting embeddings from all labeled samples in the development set to create a memory bank. During inference, each evaluation sample's embedding is compared to those in the memory bank using cosine similarity to identify the k-nearest neighbors.

For an embedding x_i from the evaluation set, its similarity to a neighboring embedding x_j in the memory bank $\mathcal{M}_L$ is quantified using the cosine similarity metric:

$$w_{ij} = \frac{x_i^\top x_j}{\|x_i\|_2 \|x_j\|_2}. \tag{2}$$

This allows for a weighted aggregation of the neighbors' labels. The final output for x_i, denoted by $\eta(x_i)$, is calculated by applying the softmax function to the combined weighted sum of the one-hot encoded labels of its nearest neighbors:

$$\eta(x_i) = \text{Softmax}\left(\sum_{x_j \in N_{\mathcal{M}_L}(x_i)} w_{ij} \mathbf{1}\{y_j\} \right). \tag{3}$$

Our methodology ensures that the model's predictions adapt to domain-specific characteristics, enhancing accuracy and reliability. This strategic use of KNN, supported by a robust memory bank, improves the model's capability to handle domain shifts.

4. EXPERIMENT

4.1. Dataset Split

In the dataset provided by the challenge, a significant number of highly similar audio recordings are present, which raises the likelihood of the model overfitting on the validation set allocated within the ICME ASC challenge dataset. Through experimental analysis, we discovered that allocating a larger portion of the dataset for training tends to make the model overly specialized to data resembling the validation set, resulting in an artificially high validation accuracy, often stabilizing at 100%. This scenario masks the true effectiveness of the model.

To devise a more effective training strategy, we adopted a small-sample training method. This involves using a small

fraction of the data for training while reserving the rest of the labeled data for validation. This approach helps in assessing the impact of various components and hyperparameters on the model's performance. We set the ratio of the training set to the validation set at 1:9 for the subsequent experiments. Ultimately, the final model is trained using the entire dataset, applying the optimal strategy derived from the small-sample experiments to the full-sample training process.

4.2. Training details

In the pre-training stage, we utilized 4 GeForce RTX 3090 GPUs, with a batch size of 12, and conducted 20,000 updates using the EAT model framework. The model was optimized with an Adam optimizer set to a learning rate of 0.0005, Adam betas of [0.9, 0.95], a weight decay of 0.05, and employed a cosine learning rate scheduler. For the exponential moving average (EMA) method, the EMA decay was set to 0.9998, and the EMA end decay to 0.99999. The student model input spectrogram featured a mask scale of 0.8.

In the fine-tuning stage, we carried out 3,000 updates using a single GTX 3090 Ti. This was followed by iterating on the inference of unlabeled data, filtering out instances with confidence above a certain threshold as pseudo-labels, and then reintroducing them into the model for further training. Mixup and SpecAugment techniques were incorporated into the training. Additionally, a portion of the TAU dataset that overlaps with the ICME challenge was included, with the training of this data weighted to ensure that the labels remained relatively balanced.

After obtaining the pseudo-labels, in order to ensure that the effect of model tuning with labeled data is not forgotten, we chose to keep the initialization parameters of the fine-tuned model and use the pseudo-labeled data to iteratively train on the original fine-tuned model.

4.3. Experimental Results

Comparison of dataset proportions. Table 1 compares the effect of different datasets and sampling weights for pre-training and fine-tuning on given datasets. The TAU dataset for fine-tuning uses the same seven classes from the CAS dataset for supervised training and weighted sampling. The best results were obtained with three datasets for sampling-weighted pre-training and fine-tuning with CAS and TAU.

Comparison of different thresholds in semi-supervised learning. Table 2 compares the effect of different thresholds for screening pseudo-labels on the results. We found that higher thresholds screened pseudo-labels have a better effect on model tuning, but if the threshold is too high then the decrease in the number of pseudo-labels has reduced the diversity of samples. In the ablation experiments, a threshold of 0.85 gives the best results. Therefore, 0.85 threshold was used as the last submitted system.

Pre-train Data	Weighted	Fine-tune Data	Acc (%)
CAS	✗	CAS	89.78
CAS+TAU+CS	✓	CAS	93.36
CAS+TAU+CS	✗	CAS+TAU	94.05
CAS+TAU+CS	✓	CAS+TAU	**94.76**

Table 1. **The accuracy score results of pre-training and fine-tuning with different datasets and different sampling strategies.** CAS, TAU, and CS stand for ICME ASC challenge dataset, TAU UAS, and CochlScene respectively. Weighted refers to whether or not weighted sampling was used in the pre-training stage, using a sampling ratio of 10:1:1 in all cases in our experiment.

The result on the ASC challenge evaluation dataset. Table 3 showcases the results of our model on the competition's evaluation dataset, where our outcomes significantly surpassed those of the baseline semi-supervised framework based on Squeeze-and-Excitation and Transformer. Ultimately, our model achieved second place in the competition.

Threshold	Accuracy (%)
0	93.74
0.5	93.86
0.7	94.53
0.8	96.52
0.85	**96.61**
0.9	95.67
0.95	93.46
1	93.36

Table 2. **Accuracy scores on the test set with pseudo-labels at different confidence thresholds. Only samples with prediction probabilities above the threshold were used for training.**

5. CONCLUSION

In the ICME-2024 challenge, we leveraged the self-supervised model EAT to capture representations rich in inherent audio information. To achieve a more robust model, we maximized the use of the provided dataset and applied a weighted method to ensure equitable balance across each dataset. Subsequently, we adopted semi-supervised learning, selecting an appropriate threshold to guarantee both the quantity and quality of pseudo-labels, thus enhancing the utilization of unlabeled data. Furthermore, we implemented a test-time adaptation strategy to boost the model's performance and generalization capability. Ultimately, our efforts culminated in achieving a Macro-accuracy of 0.752 on the evaluation dataset.

Scnene	Our (%)	Baseline (%)
Bus	76.0	40.0
Airport	94.0	54.7
Metro	99.0	90.0
Restaurant	59.0	69.0
Shopping mall	68.0	51.0
Public square	51.0	29.0
Urban park	60.7	46.0
Traffic street	76.0	65.0
Construction site	69.0	68.0
Bar	99.0	87.0
Average	75.2	60.0

Table 3. **Final results for each class on the evaluation set.**

6. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (No. U23B2018, No. 62206171, No. 62276153), and in part by Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0102 and the International Cooperation Project of PCL.

7. REFERENCES

[1] Daniele Barchiesi, Dimitrios Giannoulis, Dan Stowell, and Mark 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] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[3] 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.

[4] 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.

[5] Wenxi Chen, Yuzhe Liang, Ziyang Ma, Zhisheng Zheng, and Xie Chen, "EAT: Self-supervised pre-training with efficient audio transformer," *arXiv preprint arXiv:2401.03497*, 2024.

[6] 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.

[7] Jian Liang, Ran He, and Tieniu Tan, "A comprehensive survey on test-time adaptation under distribution shifts," 2023.

[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] 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.

[10] Yuan Gong, Cheng-I Jeff Lai, Yu-An Chung, and James Glass, "Ssast: Self-supervised audio spectrogram transformer," 2022.

[11] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao, "Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models," 2023.

[12] 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.

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

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick, "Masked autoencoders are scalable vision learners," 2021.

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo A Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al., "Bootstrap your own latent: A new approach to self-supervised learning," in *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 21271–21284.

[16] 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.