ACOUSTIC SCENE CLASSIFICATION WITH SPECTROGRAM PROCESSING STRATEGIES

Helin Wang[1], Yuexian Zou[1,2,], Dading Chong[1]*

[1]ADSPLAB, School of ECE, Peking University, Shenzhen, China
[2]Peng Cheng Laboratory, Shenzhen, China
{wanghl15,zouyx,1601213984}@pku.edu.cn

ABSTRACT

Recently, convolutional neural networks (CNN) have achieved the state-of-the-art performance in acoustic scene classification (ASC) task. The audio data is often transformed into two-dimensional spectrogram representations, which are then fed to the neural networks. In this paper, we study the problem of efficiently taking advantage of different spectrogram representations through discriminative processing strategies. There are two main contributions. The first contribution is exploring the impact of the combination of multiple spectrogram representations at different stages, which provides a meaningful reference for the effective spectrogram fusion. The second contribution is that the processing strategies in multiple frequency bands and multiple temporal frames are proposed to make fully use of a single spectrogram representation. The proposed spectrogram processing strategies can be easily transferred to any network structures. The experiments are carried out on the DCASE 2020 Task1 datasets, and the results show that our method could achieve the accuracy of 81.8% (official baseline: 54.1%) and 92.1% (official baseline: 87.3%) on the officially provided fold 1 evaluation dataset of Task1A and Task1B, respectively.

Index Terms— Acoustic scene classification, convolutional neural networks, spectrogram processing strategies

1. INTRODUCTION

Acoustic scene classification (ASC) aims to classify audio as one of a set of categories such as home, street, and office [1], and has become an important research in the fields of acoustic signal processing. Detection and Classification of Acoustic Scenes and Events (DCASE) challenges organized by IEEE Audio and Signal Processing (AASP) Technical Committee are one of the biggest competitions for ASC task [2], which attract an increasing number of participants each year. ASC is still a challenging task because of continuous, periodic or aperiodic acoustic signals that interfere with the understanding of the scene [3].

Recently, deep learning has accomplished many achievements in audio, image, and natural language processing. The algorithms based on the convolutional neural networks (CNN) are dominant in ASC tasks [4, 5, 6], which show powerful ability to extract the robust time-frequency information from audio. In order to obtain more discriminative information, different features for input were exploited. Many researchers transform the audio data into two-dimensional spectrogram representations to be fed into the back-end networks, and different spectrogram representations have been studied, such as Mel Frequency Cepstral Coefficients

(MFCC) [7], constant-Q transform (CQT) [8], Gammatone spectrogram (Gamma) [9] and log-Mel spectrogram (Log-Mel) [7]. Among them, Sakashita *et al.* [10] proposed an ensemble of spectrograms based on adaptive temporal divisions, and Seo *et al.* [11] proposed ensemble systems of CNN with various pre-processed features. Pham *et al.* [3] presented an encoder-decoder architecture to parallelly map three spectrograms and combine them in the middle layers of the networks. In addition, Ngo *et al.* [12] provided a comprehensive analysis on five common types of spectrograms, and made spectrogram fusion in the first layer of the networks. Several other works [13, 14] focused on how to better use a single representation (*e.g.* Log-Mel). With the inspiration that different frequency bands in a spectrogram contain distinct features, Phaye *et al.* [13] proposed a SubSpectralNet network which was able to extract discriminative information from sub-spectrograms. Qiao *et al.* [14] truncated the whole spectrogram into different sub-spectrograms, and adopted a score level based fusion mechanism to jointly improve the classification accuracy.

In this paper, we study the effective methods to take advantage of the spectrogram representations. Based on the convolutional neural networks (CNN), we propose several spectrogram processing strategies to obtain more discriminative information for ASC, including the spectrogram processing strategies in multiple representations (SPSMR), multiple frequency bands (SPSMF), and multiple temporal frames (SPSMT). More specifically, log-Mel spectrogram (Log-Mel), constant-Q transform (CQT), Gammatone spectrograms (Gamma) and Mel Frequency Cepstral Coefficients (MFCC) are used as the input to the networks. Instead of the feature-level fusion for the spectrogram representations, four independent networks with different representations are applied before the decision-level fusion, which is the SPSMR. In addition, we exploit to make decision on the subparts of a spectrogram rather than the whole spectrogram to improve the robustness, *i.e.* the spectrogram processing strategy in multiple frequency bands (SPSMF) and the spectrogram processing strategy multiple temporal frames (SPSMT). Under the official train/test split of DCASE 2020 Task1 development set [15], our system could achieve 81.8% accuracy with 0.694 log loss in the Task1A evaluation set, and 92.1% accuracy with 0.312 log loss in the Task1B evaluation set.

The remainder of this paper is organized as follows. Section 2 gives full particulars of a series of processing strategies. Section 3 details the architectures of our networks. Section 4 presents the details of experiments and results, and Section 5 concludes this paper.

2. SPECTROGRAM PROCESSING STRATEGIES

In this section, the conventional CNN-based method and our proposed spectrogram processing strategies are introduced, which are
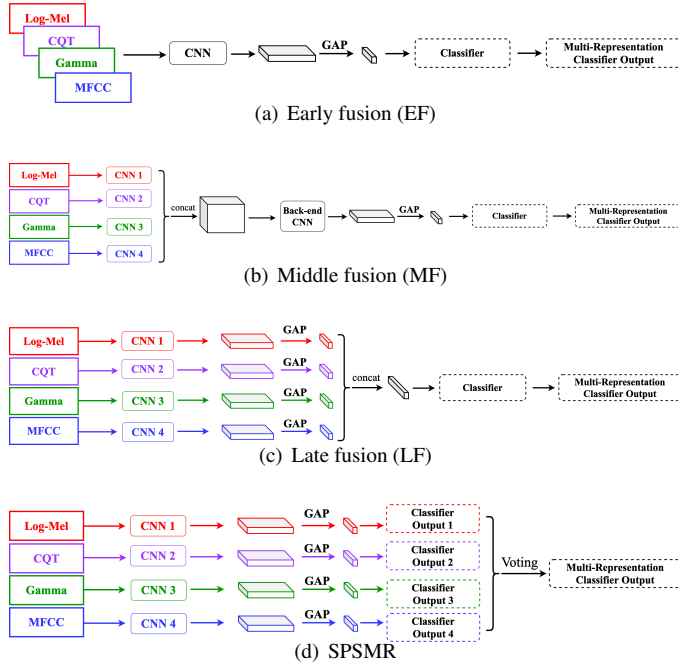
---

* Yuexian Zou is the corresponding author.

(a) Early fusion (EF)

(b) Middle fusion (MF)

(c) Late fusion (LF)

(d) SPSMR

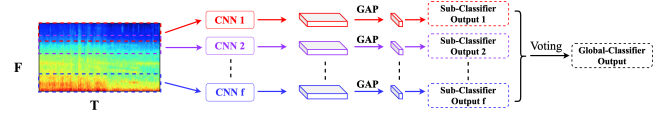Figure 1: The illustration of different spectrogram fusion methods and our proposed SPSMR.

Figure 2: The illustration of our proposed SPSMF.

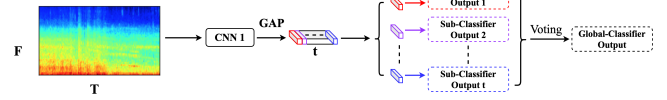

Figure 3: The illustration of our proposed SPSMT.

2.2. The Spectrogram Processing Strategy in Multiple Representations

Instead of inputting single representation to the networks, multiple representations can obtain more robust information from the raw audio. Four types of spectrogram representations (*i.e.* Log-Mel, CQT, Gamma, and MFCC) are used in our experiments. As shown in Figure 1(a), one intuitive approach [12] to apply multiple representations is inputting the multi-channel feature maps $M^* \in \mathbb{R}^{n \times T \times F}$, where n denotes the number of representations. However, different representations have different characteristics and a single CNN network cannot model the differences effectively. Other fusion methods [3, 18, 19] could be the middle fusion in Figure 1(b) or the late fusion in Figure 1(c), which jointly uses the multiple representations after several CNN layers or the whole CNN layers. However, the same regions in different feature maps reflect different frequency information, which may cause the mismatched problem. In addition, a single classifier trained on the fusion features cannot take advantage of multiple representations well.

In order to overcome the above problems and make better use of the representations, we propose a novel spectrogram processing strategy in multiple representations (SPSMR), which trains several independent CNN models based on different representations and then combines them by the decision-level fusion. As shown in Figure 1(d), the predicted scores of the four models are averaged to obtain the final predicted score, which is also known as the average voting strategy.

$$\hat{y} = \frac{1}{4} \left(\hat{y}_1 + \hat{y}_2 + \hat{y}_3 + \hat{y}_4 \right) \tag{3}$$

where $\hat{y}_1$, $\hat{y}_2$, $\hat{y}_3$, $\hat{y}_4$ denotes the predicted score of CNN model with the input representation of Log-Mel, CQT, Gamma, and MFCC, respectively. In this case, each independent model can focus on different representations and be more discriminative by the decision-level fusion.

the spectrogram processing strategy in multiple representations (SPSMR), the spectrogram processing strategy in multiple frequency bands (SPSMF), and the spectrogram processing strategy in multiple temporal frames (SPSMT). In addition, several other spectrogram fusion methods are also introduced.

2.1. The Conventional CNN-based Method

CNN-based methods were widely used in ASC task, and provided the state-of-the-art performance [16, 17]. To be specific, given an audio clip, the two-dimensional time-frequency representation (*e.g.* Log-Mel) is first extracted. Convolutional layers are then applied to the time-frequency representation $M \in \mathbb{R}^{T \times F}$ to obtain the deep representation $M' \in \mathbb{R}^{c \times t \times f}$, where c denotes the number of the output channels.

$$M' = f_{\text{cnn}} \left(M; \theta_{\text{cnn}} \right) \tag{1}$$

Here, f_{cnn} denotes the operation of the convolutional layers and θ_{cnn} denotes the model parameters of the convolutional layers. The global pooling layer (*e.g.* global average pooling) and fully-connected layers are then applied to obtain the predicted score of the classification. Let f_{gp}, f_{fc} be the operations of the global pooling layer and the fully-connected layers, respectively. The predicted score $\hat{y} \in \mathbb{R}^N$ (where N denotes the number of categories) can be obtained by

$$\hat{y} = f_{\text{fc}} \left(f_{\text{gp}} \left(M' \right); \theta_{\text{fc}} \right) \tag{2}$$

where θ_{fc} denotes the model parameters of the fully-connected layers.

2.3. The Spectrogram Processing Strategy in Multiple Frequency Bands

The spatial regions of the feature maps are treated equally in the conventional CNN-based methods, however, different acoustic scenes show different activity on the frequency bands [13]. Therefore, we take the sub-spectrograms [13] as input and train several

classifiers. Different from [13], the final decision is made by the average voting strategy rather than training a global classifier, which shows better performance in our experiments. As shown in Figure 2, for f sub-spectrograms, the final score is obtained by

$$\hat{\boldsymbol{y}} = \frac{1}{f} \sum_{i=1}^{f} \hat{\boldsymbol{y}}_i \qquad (4)$$

2.4. The Spectrogram Processing Strategy in Multiple Temporal Frames

Several temporal devisions have been studied in [10] for ASC task to promote the generalization and robustness, including non division, non-overlap division and overlap division. Among them, overlap division shows the best performance. However, such division method only focused on the information within the window size of the division (*e.g.* 2s). In this paper, a spectrogram processing strategy in multiple temporal frames (SPSMT) is proposed, which feeds the whole audio clip to the network and makes decision on each temporal frame after CNN. Thus, the decision made by each frame could take into account the information of neighboring frames. As shown in Figure 3, for the final feature map $\boldsymbol{M}^{'} \in \mathbb{R}^{c \times t \times f}$, global average pooling is applied to the frequency bands and the classifier is then applied to each temporal frames.

$$\hat{\boldsymbol{y}} = \frac{1}{t} \sum_{i=1}^{t} \hat{\boldsymbol{y}}_i \qquad (5)$$

Note that all the temporal frames share the same classifier in SPSMT, so that there are no extra parameters needed.

3. NETWORK ARCHITECTURES

Our base network architectures are shown in Table 1. For DCASE 2020 Task1A, the network is a VGG [20] style network, similar to [21]. Batch normalization (BN) [22] and Rectified Linear Units (ReLU) [23] are used following the convolutional operations. Global pooling is applied after the last convolutional layer to obtain fixed-length vectors, which is operated by the global average pooling in the frequency axis and the global max pooling in the temporal axis [21]. Two fully-connected layers followed by a softmax function are then applied to obtain the prediction for classification. Dropout with a ratio of 0.5 is applied between the fully-connected layers. While for DCASE 2020 Task1B, a tiny CNN model is employed to achieve the low complexity, and other setups are the same as Task1A.

4. EXPERIMENTS

4.1. Datasets and Metrics

Our proposed method is evaluated on two novel datasets for ASC, *i.e.* TAU Urban Acoustic Scenes 2020 Mobile, Development dataset (DCASE2020 1A) [15] and TAU Urban Acoustic Scenes 2020 3Class, Development dataset (DCASE2020 1B) [15]. DCASE2020 1A contains totally 64h data from 10cities and 9 devices. The dataset is provided with a training/test split in which 70% of the data for each device is included for training, 30% for testing. The task targets generalization properties of systems across a number of different devices. DCASE2020 1B contains data with a single device from 10 cities. The total amount of audio is 40 hours, and
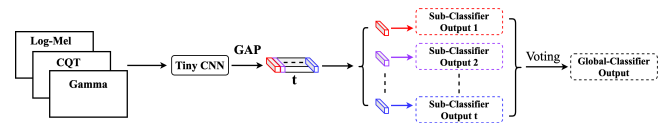

Figure 4: The illustration of DCASE2020 1B network.

Table 1: Network Architectures

DCASE2020 1A	DCASE2020 1B
Conv 3×3 @ 64, BN, ReLU Conv 3×3 @ 64, BN, ReLU	Conv 7×7 @ 32, BN, ReLU
Avg Pooling 4×2	Avg Pooling 4×2
Conv 3×3 @ 128, BN, ReLU Conv 3×3 @ 128, BN, ReLU	Conv 7×7 @ 32, BN, ReLU
Avg Pooling 4×2	Avg Pooling 4×2
Conv 3×3 @ 256, BN, ReLU Conv 3×3 @ 256, BN, ReLU	Conv 3×3 @ 64, BN, ReLU
Avg Pooling 2×2	Avg Pooling 2×2
Conv 3×3 @ 512, BN, ReLU Conv 3×3 @ 512, BN, ReLU	Conv 3×3 @ 64, BN, ReLU
Global Pooling	Global Pooling
FC 512, ReLU	FC 200, ReLU
FC 10, softmax	FC 3, softmax

audio is provided in binaural, 48kHz 24-bit format. The task targets low complexity solutions for the classification problem.

Macro-average accuracy and multiclass cross-entropy (log loss) are used as our metrics [15]. In addition, we also report the model size.

4.2. Experimental Setups

For DCASE2020 1A, all the raw audios are resampled to 44.1kHz and fixed to the certain length of 10s by zero-padding or truncating. Log-Mel, CQT, Gamma, and MFCC are then extracted with window size 2048 (46ms) and hop length 512 (11.6ms). The number of frequency bands are 40, 64, 64 and 40 respectively, which are the optimal numbers for each representation according to our experimental results. We test the single representation methods (*i.e.* Log-Mel CNN, CQT CNN, Gamma CNN, and MFCC CNN), and our proposed methods (*i.e.* SPSMR, SPSMF, and SPSMT).

For DCASE2020 1B, in order to achieve the low complexity, early fusion (EF) is applied instead of SPSMR. In addition, SPSMT is applied because of no extra parameters. The overall architecture is shown in Figure 4. Three representations are used, *i.e.* Log-Mel, CQT and Gamma, and the number of frequency bands are all 64. Other experimental setups are the same as DCASE2020 1A.

In the training phase, the Adam algorithm [24] is employed as the optimizer with the default parameters. The model is trained end-to-end with the initial learning rate of 0.001 and the exponential decay rate of 0.91 for each 200 iterations. Parameters of the networks are learned using the categorical cross entropy loss. Batch size is set to 64 and training is terminated after 12000 iterations. Data augmentation methods Mixup [25] is applied on the audio sample level in our experiments to prevent the system from over-fitting and improve the performance.

Table 2: Comparison of accuracy and log loss on DCASE2020 1A dataset

Model	Accuracy	Log loss	Model size
DCASE2020 1A Baseline [15]	51.4%	1.365	19.1 MB
Log-Mel CNN	72.1%	0.879	18.9 MB
CQT CNN	71.2%	0.870	18.9 MB
Gamma CNN	76.1%	0.762	18.9 MB
MFCC CNN	63.6%	1.029	18.9 MB
SPSMR	79.4%	0.696	75.5 MB
Log-Mel CNN + SPSMF	75.5%	1.135	94.4 MB
CQT CNN + SPSMF	74.5%	1.185	94.4 MB
Gamma CNN + SPSMF	78.8%	1.169	94.4 MB
MFCC CNN + SPSMF	60.9%	1.801	94.4 MB
SPSMR + SPSMF	80.9%	0.737	377.6 MB
Log-Mel CNN + SPSMT	74.5%	0.987	18.9 MB
CQT CNN + SPSMT	73.3%	1.032	18.9 MB
Gamma CNN + SPSMT	78.2%	0.866	18.9 MB
MFCC CNN + SPSMT	67.6%	1.081	18.9 MB
SPSMR + SPSMT	79.7%	0.701	75.5 MB
SPSMR + SPSMF + SPSMT	**81.8%**	**0.694**	453.1 MB

Table 3: Comparison of different spectrograms fusion methods on DCASE2020 1A dataset

Model	Accuracy	Log loss	Model size
Log-Mel CNN	72.1%	0.879	18.9 MB
CQT CNN	71.2%	0.870	18.9 MB
Gamma CNN	76.1%	0.762	18.9 MB
MFCC CNN	63.6%	1.029	18.9 MB
EF	75.2%	0.852	18.9 MB
MF	77.0%	0.777	20.2 MB
LF	76.4%	0.698	72.5 MB
SPSMR	**79.4%**	**0.696**	75.5 MB

4.3. Experimental Results

Table 2 demonstrates the experimental results of different models on DCASE2020 1A dataset. Among them, SPSMR + SPSMF + SPSMT achieves the highest accuracy and the lowest log loss, which shows that our proposed SPSMR, SPSMF, SPSMT can obviously improve the performance for ASC. As for single representation methods (*i.e.* Log-Mel CNN, CQT CNN, Gamma CNN and MFCC CNN), Gamma CNN beats others. SPSMR outperforms all the single representation methods, which evaluates the effectiveness of using multiple representations. In addition, SPSMF and SPSMT can be directly applied to a single representation method or SPSMR, and both show accuracy gains. It is worth mentioning that SPSMT can improve the performance for ASC without extra parameters.

Furthermore, we compare different spectrogram fusion methods with our proposed SPSMR on DCASE2020 1A dataset, and show the results in Table 3. Apart from SPSMR, we report the results of four single representation methods (*i.e.* Log-Mel CNN, CQT CNN, Gamma CNN and MFCC CNN) and three feature-level fusion methods (*i.e.* EF, MF and LF), which are introduced in Section 2.2. EF performs worse than the other fusion methods and even worse than a single representation method (*i.e.* Gamma CNN). This is because the less discriminative representation (*e.g.* MFCC) for ASC may interfere other representations by EF. In addition, MF performs better than LF, and becomes the best feature-level fusion method in our experiments. It can be seen that our proposed SPSMR

Table 4: Comparison of accuracy and log loss on DCASE2020 1B dataset

Model	Accuracy	Log loss	Model size
DCASE2020 1B Baseline [15]	87.3%	0.437	450 KB
Log-Mel CNN	87.5%	0.428	468 KB
Log-Mel CNN + SPSMT	90.8%	0.371	468 KB
EF + SPSMT	**92.1%**	**0.312**	491 KB

outperforms all the other methods, which shows the powerful ability of combining multiple representations. We attribute this to the four independent networks in SPSMR rather than making fusion in the feature level, which allows each representation to be more discriminative. The decision-level fusion by SPSMR also leads to the improvement of robustness when a single representation shows poor discriminative for ASC.

Our proposed method is also evaluated on DCASE2020 1B dataset. As presented in Table 4, EF+SPSMT outperforms the official baseline with the similar model size, which shows that both EF and SPSMT are the powerful methods to improve the performance for ASC with few extra parameters. Different network structures are employed in DCASE2020 1A and 1B, which shows that our proposed spectrogram processing strategies can fit different networks.

5. CONCLUSION

In this paper, three spectrogram processing strategies (*i.e.* SPSMR, SPSMF, and SPSMT) have been proposed to make better use of the spectrogram representations and greatly improved the performance for ASC task. These strategies were designed to fit the audio characteristics, and can be directly applied to other neural networks. We believe that the proposed strategies can offer good generalization properties for other audio processing tasks. The code for the experiments is available.[1]

6. ACKNOWLEDGMENT

7. REFERENCES

[1] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational analysis of sound scenes and events.* Springer, 2017.

[2] http://dcase.community/workshop2020/.

[3] L. Pham, H. Phan, T. Nguyen, R. Palaniappan, A. Mertins, and I. McLoughlin, "Robust acoustic scene classification using a multi-spectrogram encoder-decoder framework," *arXiv preprint arXiv:2002.04502*, 2020.

---

[1]https://github.com/WangHelin1997/
DCASE-2020-Task1A-Code
[2]https://github.com/qiuqiangkong/dcase2019_
task1

[4] 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.

[5] Y. Han, J. Park, and K. Lee, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," *the Detection and Classification of Acoustic Scenes and Events (DCASE)*, pp. 1–5, 2017.

[6] 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.

[7] I. V. McLoughlin, *Speech and audio processing: a MATLAB-based approach*. Cambridge University Press, 2016.

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

[9] D. P. Ellis, "Gammatone-like spectrograms," 2009. [Online]. Available: http://www.ee.columbia.edu/dpwe/resources/matlab/gammatonegram

[10] Y. Sakashita and M. Aono, "Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions," *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2018.

[11] H. Seo, J. Park, and Y. Park, "Acoustic scene classification using various pre-processed features and convolutional neural networks," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), New York, NY, USA*, 2019, pp. 25–26.

[12] D. Ngo, H. Hoang, A. Nguyen, T. Ly, and L. Pham, "Sound context classification basing on join learning model and multi-spectrogram features," *arXiv preprint arXiv:2005.12779*, 2020.

[13] S. S. R. Phaye, E. Benetos, and Y. Wang, "Subspectralnet–using sub-spectrogram based convolutional neural networks for acoustic scene classification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 825–829.

[14] T. Qiao, S. Zhang, Z. Zhang, S. Cao, and S. Xu, "Sub-spectrogram segmentation for environmental sound classification via convolutional recurrent neural network and score level fusion," in *2019 IEEE International Workshop on Signal Processing Systems (SiPS)*. IEEE, 2019, pp. 318–323.

[15] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions," *arXiv preprint arXiv:2005.14623*, 2020.

[16] H. Chen, Z. Liu, Z. Liu, P. Zhang, and Y. Yan, "Integrating the data augmentation scheme with various classifiers for acoustic scene modeling," *arXiv preprint arXiv:1907.06639*, 2019.

[17] K. Koutini, H. Eghbal-zadeh, and G. Widmer, "Cp-jku submissions to dcase'19: Acoustic scene classification and audio tagging with receptive-field-regularized cnns."

[18] 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," *arXiv preprint arXiv:1912.10211*, 2019.

[19] J. Huang, H. Lu, P. Lopez-Meyer, H. A. C. Maruri, and J. A. d. H. Ontiveros, "Acoustic Scene Classification using Deep Learning-Based Ensemble Averaging," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE)*, New York, NY, USA, 2019, pp. 94–98.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[21] Q. Kong, Y. Cao, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, "Cross-task learning for audio tagging, sound event detection and spatial localization: Dcase 2019 baseline systems," *arXiv preprint arXiv:1904.03476*, 2019.

[22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[23] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.