Deep Semantic Encoder-Decoder Network for Acoustic Scene Classification with Multiple Devices

Xinxin Ma*, Yunfei Shao†, Yong Ma*⁕ and Wei-Qiang Zhang†
* School of Physics and Electronic Engineering, Jiangsu Normal University, Xuzhou 221116, China
⁕ Jiangsu Normal University Kewen College, Xuzhou 221116, China
† Beijing National Research Center for Information Science and Technology,
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
E-mail: mxxjsnu@163.com, shaoyf@tsinghua.edu.cn, may@jsnu.edu.cn, wqzhang@tsinghua.edu.cn

Abstract— **In this paper, we proposed Mini-SegNet, a simplified encoder-decoder SegNet model to capture deep semantic information in sound events. The semantic information can effectively discriminate the acoustic segments in different scenes. We also applied spectrum correction to combat mismatched frequency response. In order to prevent over-fitting, we adopted mixup augmentation, ImageDataGenerator and temporal crop augmentation for data augmentation. Our best single system achieved an average accuracy of 65.15% on different devices in the DCASE2020 Development dataset, more than 10% improvement over the baseline system. The results indicate that our approach can achieve good classification performance, without use of any supplementary data from outside the official challenge dataset.**

I. INTRODUCTION

Sounds carry a great deal of information about our environment, from individual physical events to sound scenes as a whole. The problem of sensing and understanding the environment in which a sound is known as Acoustic Scene Classification (ASC) [1]. It is a multi-class classification task recognizing the recorded environment sounds specific acoustic scenes that characterize either the location or situation such as park, metro station, tram, etc. ASC has been applied in many fields, such as context-aware [2], surveillance [3], and robotic navigation [4]. For example, if a self-driving automobile "hears" children yelling from blind spot, it can slow down to avoid a possible accident. Intelligent Bluetooth headset can automatically reduce noise and adjust the volume of headset according to the environment of users.

Acoustic Scene Classification is one of the core research problems in the field of Computational Sound Scene Analysis. It has been a major task in the IEEE Audio and Acoustic Signal processing (AASP) Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) since 2013. In the 2013 DCASE Challenge [5], the organizer established and released the open datasets, and provided the benchmark to evaluate different approaches for acoustic scene classification, which motivating researchers to further work in this area. Subsequently, many methods have been applied to acoustic scene classification, such as signal processing and machine learning, including dictionary learning [6], matrix factorization [7][8], wavelet filterbanks [9]. And recently popular deep learning, such as Convolutional Neural Networks (CNNs) [10], Gated Recurrent Neural Networks (GRNN) [11], Convolutional Recurrent Neural Networks (CRNN) [12], and Attention-based Atrous Convolutional Neural Networks [13].

As one of the substantial tasks, acoustic scene classification has been extensively practiced in every challenge. In the past few years, the audio data of acoustic scene classification came from a kind of high-quality acquisition equipment. In this case, we call it regular acoustic scene classification. However, in real life, environmental sound is mostly collected by different recording devices. In order to study acoustic scene classification more widely, DCASE 2018 and 2019 proposed the mismatch in different recording devices A, B, C and D. DCASE 2020 [14], ASC challenge consists of 2 subtasks, Acoustic Scene Classification with Multiple Devices, and Low-complexity Acoustic Scene Classification. In this paper, we focus on the first subtask. This task contains 10 classes of sounds recorded with multiple devices. The dataset contains a certain amount of examples from a high quality devices (referred to as A), but only a limited number from the targeted low quality devices (referred to as B and C) and simulated devices (referred to as S1-S6). Especially, a part of the evaluation dataset is a compressed version of recorded audio data from device D and simulated devices S7-S11. This brings ASC closer to real-world conditions, but also presents a huge challenge.

The general framework of regular acoustic scene classification usually contains two steps. Firstly, obtain 2D time-frequency representation of audio data, and extracting relevant features. Second employ these features to learn and achieve classification. And the different features are used in acoustic scene classification, such as log-mel energies, their nearest neighbor filtered version [15], mel-spectrograms from harmonic percussive source separation (HPSS) audio [16][17], and spectrograms of Gammatone filters and Constant Q Transform (CQT) [18]. After computing the 2D time-frequency representation, some methods based on CNNs have achieved good performance for acoustic scene classification.

The corresponding author are Yong Ma and Wei-Qiang Zhang. This work was supported by the National Natural Science Foundation of China under Grant No. U1836219 and the Natural Science Foundation of College and Universities in Jiangsu Province (No.17KJB510018).

Mun et al. [19] addressed the problem of data insufficiency and proposed to use the Generative Adversarial Network (GAN) to augment training data. Phaye et al. [20] used sub-spectrograms by giving intuitive and statistical analyses, developed a sub-spectrogram based on CNN architecture for ASC. McDonnell [21] proposed deep residual network with late fusion of separated high and low frequency paths. Even though the previous methods have improved performance a lot, there are still a lot of basic problems worth exploring. For example, many scenes are quite confusing between each other and have high similarity in time. Moreover, CNN-based approaches are hard to capture the correlation of sound events in different scenes.

Compared with the regular acoustic scene classification, the acoustic scene classification with multiple devices still need some processing to adjust the different recording devices. For this response, spectrum correction [22], aggressive regularization and augmentation [23], domain adaption [24] and feature transform [25] are gradually used in acoustic scene classification with multiple devices.

In this paper, we propose a concept of semantic segmentation for acoustic scene classification. Therefore, we designed encoder-decoder network similar to SegNet [26] for acoustic scene semantic segmentation, which we term Mini-SegNet. And, to evaluate our network model, we participated in DACSE 2020 task1a: Acoustic Scene Classification with Multiple Devices. To deal with mismatched data, we learn from the work of Michal Kosmider et al. [22], try to apply spectrum correction to adjust the varying frequency response of the recording devices. Our system consists of two important stages. Firstly, mono audio signals are converted to time-frequency representations, scaled by spectrum correction, and zero mean and unit variance normalization. Secondly, the log-mel feature are fed to Mini-SegNet models for feature learning. Besides, we adopted mixup, ImageDataGenerator and temporal crop augmentation for data augmentation.

The rest of the paper is organized as follows. Section II presents the proposed ASC systems, including audio preprocessing and spectrum correction, the convolutional neural networks, and data augmentation. Section III provides experiments and the performance of the proposed approach. Section IV discussion the experimental in detail. Finally, conclusion is provided in Section V.

II. THE SYSTEM

This section introduces the proposed audio preprocessing methods. It also describes the details utilized process flow and model architecture.

A. Audio preprocessing and spectrum correction

The spectrum correction can scale the frequency response of the recording devices, which was descried and demonstrated in [22]. It is mainly implemented in two steps. First, the correction coefficients are calculated from the spectrum of *n* aligned pairs of recordings. Second, all recordings are then transformed using the calculated coefficients. In view of our experimental comparison, we only use 750 samples of data from each device A, B, C to determine the reference spectrum and the coefficients of each device. The spectrum coefficients are expressed as vectors, i.e. one coefficient per frequency bin. Then use the corresponding coefficient to scale the spectrum bin of each device. The correction is applied by multiplying the Short Time Fourier Transform of the signal by the correction coefficients on the frequency axis of each time point.

After spectrum correction, further present the spectrum in the log-mel domain. The data are mono audio files with 44.1 kHz sample rate. We transformed them into power spectrogram by skipping every 1024 samples with 2048 length Hanning window. A spectrum of 431 frames was yields from 10 seconds audio file, and each spectrum was compressed into 128 bins of mel frequency scale. Then, zero mean and unit variance normalization is applied to the log-mel feature. Therefore, we extract the log-mel energy of 128 frequency bins and 431 temporal frames per segment.

B. Proposed Model

CNNs have powerful feature extraction capabilities, which realizes feature extraction and dimensionality reduction through operations such as convolutional and pooling. The previous classification methods only used lasted feature map, the feature map followed by some Fully Connected layers (FC). It not only has a large number of parameters, but also has a large amount of calculation. In image classification, using the last feature map's information only, can achieve great performance. But in our case, its performance is not satisfactory.

In the field of target detection, some researchers proposed semantic pixel-wise segmentation of images, such as SegNet [26], using multi-scale feature mapping to improve detection performance. The key component of SegNet is the decoder network, which consists of a decoder layer corresponding to each encoder. Of these, the appropriate decoder performs nonlinear up sampling on its input feature map using the max-pooling index received from the corresponding encoder. We think that the acoustic scene is composed of some basic units (acoustic events), just as language governs the syntax of phonemes and words. We know that bird chirping sound is recorded in the park, the sound of aircraft engines is recorded in the airport. Bird chirping and aircraft engines are what we call acoustic events. These acoustic events contain some semantic information, which has a certain internal relationship with the discrimination of acoustic scene. CNN-based models have been widely utilized to encode complicated scene utterances into high-level semantics representations. Inspired by these ideas, we use an encoder/decoder architecture to learn the acoustic scene to events mapping.

Therefore, we like to test the idea of Mini-SegNet and use pooling indices to inform the up-sampling layers to extract acoustic features by the pooling layers in the encoding process. This makes it easier for the decoder to localize the acoustic events in frequency. In our work, we modified the original SegNet to extract the multi-granularity abstract features, as shown in Fig. 1. In encoder module, convolution and pooling are used to extract features and reduce dimensions.

In the decoding process, the position and frequency band information are recovered by convolution of the corresponding encoding module sampled on up-sampling to make up for the missing pixel information. This method makes full use of the semantic information of sound events in the acoustic scene through the encoding and decoding process, and uses the rules of "acoustic scene based on sound events" to provide a preliminary basis for future work. We give more details of the network below.

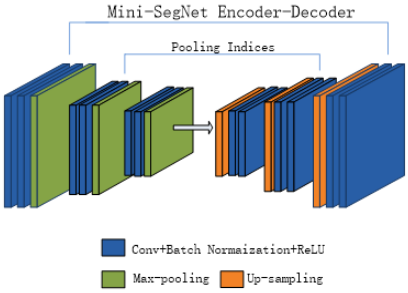

Fig. 1: Illustrate how to extract the multi-scale features

C. *Mini-SegNet for ASC*

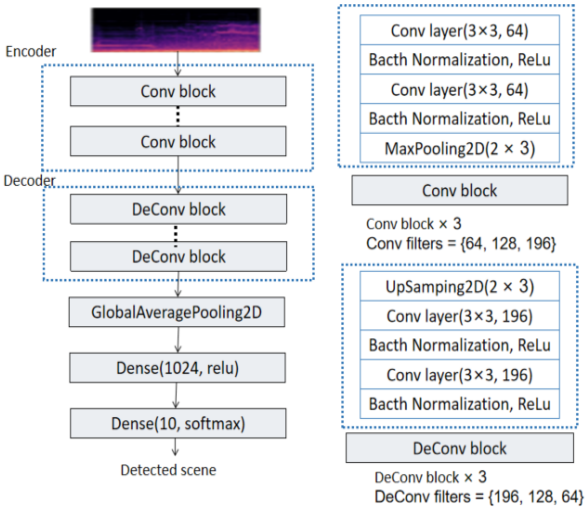

Fig. 2: Details of the Mini-SegNet model for ASC

The design of the encoder-decoder network represents the key technical contribution of this paper. The details of Mini-SegNet as shown in Fig. 2, it describes the acoustic scene classification under Mini-SegNet model. In this network, we use a simpler and smaller convolution/de-convolution. In the SegNet, the final decoder output is fed to a multi-class softmax classifier to produce class probabilities for each pixel independently. In our modified Mini-SegNet, we first change the FC to the global-average pooling. Because FC has many parameters, this reduces the number of parameters and can reduce the occurrence of over fitting. And use the softmax layer to achieve acoustic scene classification.

As shown in Fig. 2, it is mainly composed of encoder and decoder module. The number maps in decoder are 64, 128, and 196. The number of features maps in the decoder are 196, 128, and 64. In the encoder module, consists of three Conv block. Each block contains two Convolutional layers whose convolution kernel is 3 × 3, followed by a batch normalization, a ReLU non-linearity, and a maxpooling. The output of the encoder is taken as the input of the decoder module. The decoder module consists of three DeCov block and similar to encoder. Each DeConv block, up-samping is performed first, then followed by Convolutional layers, batch normalization, ReLU. Finally, global max pooling is applied, and two dense layers are utilized to output final predictions.

D. *Data augmentation*

In order to prevent over-fitting, we combined mixup [27], ImageDataGenerator and temporal crop augmentation.

In mixup, we randomly select a pair of samples from training data. Let x_1, x_2 be the features, and y_1, y_2 be the one-hot labels respectively, the data is mixed as follows:

$$x = \lambda x_1 + (1 - \lambda) x_2 \quad (1)$$

$$y = \lambda y_1 + (1 - \lambda) y_2 \quad (2)$$

where the parameter λ is a random variable with Beta distribution B (0.4, 0.4).

In addition, we tried to use ImageDataGenerator in this task. It is an image generator, mainly used in image classification. At the same time, it can also enhance the data in batch, expand the size of data set, and enhance the generalization ability of the model. In our work, it is implemented with width shift, height shift. We additionally used crop augmentation in the temporal axis: each of the two samples combined using mixup were first cropped independently and randomly from 431 dimensions down to 400. In our work, data augmentation does improve performance, and we make a detailed comparison in next section.

III. EXPERIMENT

A. *Experiment setup*

All trainings were done on GPU, with a batch size of 32, with the cross-entropy loss function, and with stochastic gradient descent with momentum of 0.9 for the optimizer. At the same time, we using a warm restart learning rate schedule, its maximum value of 0.1 after 2, 6, 14, 30, 126 and 254 epochs, and then decays according to a cosine pattern to 1×10^{-5}. In our work, each network has trained for 510 epochs. Experiments show that this method can improve the accuracy of acoustic scene classification.

B. *Dataset*

To evaluate our system, we use the task1a data from the official dataset of TAU Urban Acoustic Scene 2020 Mobile Development dataset. The dataset consists of 10 acoustic scenes: airport, park, metro, street_pedestrian, street_traffic, tram, metro_station, bus, public_square, shopping_mall. The

development set contains data from 10 cities and 9 devices: 3 real devices (A, B, C) and 6 simulated devices (S1-S6). The total amount of audio in the development set is 64 hours.

The development dataset comprises 40 hours of data from device A, and smaller amounts from the other devices. Audio is provided in single channel 44.1 kHz 24-bit format, and were split into 10s segments that are provided in individual files. As shown in Table I, the development dataset is provided with a train/test in which 70% of the data for each device is included for training, 30% for testing. Some simulated devices (S4, S5, S6) appear only in the test subset.

Table I
TAU Urban Acoustic Scenes 2020 Mobile Development dataset

Device	Total segments	Train segments	Test segments
A	14400	10215	330
B, C	1080	750	330
S1, S2, S3	1080	750	330
S4, S5, S6	1080	—	330

C. *Comparison with baseline*

The DCASE2020 Task1A challenge [14] is evaluated using accuracy calculated as the average of the class-wise accuracy, also known as "macro-average accuracy". Because the data sets come from different devices, as shown in the Table I. Our experimental results are mainly evaluated by the average accuracy, that is, the average accuracy of scene classification under various devices.

Table II: Details parameters and results (All-accuracy: %, train/test) of the SegNet in various configurations. All of these, trained with same data augmentation and no spectrum correction, with 254 epochs.

	SegNet		Mini-SegNet
	Original	Smaller	Our proposed
Encoder	(64*2, 128*2, 256*3, 512*3, 512*3)	(64*2, 128*2, 256*3)	(64*2, 128*2, 196*2)
Decoder	(512*3, 512*3, 256*3, 64*2, 64*2)	(256*3, 128*2, 64*2)	(196*2, 128*2, 64*2)
Train params	31, 803, 338	4, 063, 178	2, 086, 954
Time(s)/Epoch	387s	295s	198s
All-accuracy	56.63%/93.21%	62.89%/92.68%	63.97%/91.89%

In our work, we first change the final decoder output layer of SegNet to the global-average pooling. Then on this basis, we tested the original SegNet network and other architectures. In Table II, we show some of our configuration parameters and results. In order to save the training time, we only set 254 epochs. In Table II, 64 * 2 means two convolution layers with 64 output maps. The original SegNet, each encoder network has a corresponding decoder layer and hence the encoder network has 13 convolutional layers. The train parameters are 31, 803, 338, and the training time of each epoch is 387 seconds. The train parameters (Train params) are big and the train time (Times(s)/Epoch) is too long. The all-accuracy is 93.21% on the train set and 56.63% on the test set. The results show that its performance is poor, especially in the train set and test set there is a large over-fitting. The main reason is that original SegNet has a deep network depth, which cannot be fully used when our data is limited. Therefore, we have made many attempts to modify it from the depth of the network. As shown in Table II, after adjusting the network depth and making some parameters comparison, we propose the Mini-SegNet. Compared with others structures, the Mini-SegNet has a better accuracy, and less train parameters and less train time, and over-fitting has also been alleviated.

We report the performance of our system using this train-test setup in order to allow comparison of different system on the development dataset. First, repeat the baseline system, it is a modification of the baselines from previous DCASE challenge editions of acoustic scene classification, built on the same skeleton. It replaces use of mel energies with use of OpenL3 embeddings and replaces the CNN network architecture with two fully-connected feed-forward layers (size 512 and 128) as in the original OpenL3 publication [28]. In DCASE 2020 Task1A baseline system, OpenL3 as audio embeddings, two fully connected layers as classifier. The baseline system has an average accuracy of 54.1% (+-1.4) on different devices in the development dataset. Compared with the baseline system, our proposed method achieves a relative improvement or more than 10% on task1a.

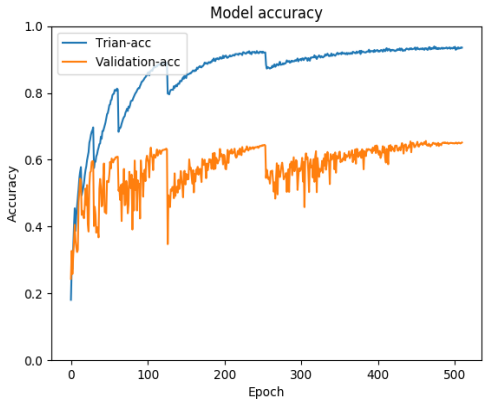

Fig. 3: Accuracy of proposed system with warm restart learning rate schedule (510 epochs)

Our proposed system achieved of 65.14% on the different devices in the development dataset. As shown in Fig. 3, we proposed system with warm restart learning rate schedule achieve a high performance in the train-set, but only 65.14% accuracy in the validation-set. We chose the average accuracy under various recording devices (all-accuracy) as the main indicator because the task targets generalization properties of systems across a number of different devices. It is found that the system has a certain degree of over fitting, which is main due to the data inconsistency in various recording devices, as well as the unbalanced distribution of data on the training set

Table IV: Detail results of the best Mini-SegNet on different devices accuracy (%)

Scene	Accuracy	Dev. A	Dev. B	Dev. C	Dev. S1	Dev. S2	Dev. S3	Dev. S4	Dev. S5	Dev. S6
airport	55.22	72.73	57.57	75.76	45.45	63.64	54.54	60.61	45.45	21.21
bus	78.79	93.94	84.85	90.91	69.69	72.73	75.76	72.27	57.57	90.91
metro	70.71	81.82	69.70	63.63	66.67	66.67	63.64	78.79	75.76	69.67
metro_station	68.01	84.85	81.82	54.55	72.73	63.63	75.76	54.54	72.73	51.51
park	77.78	87.88	87.87	81.82	78.78	75.76	78.79	75.76	63.63	69.70
public_square	54.88	63.64	51.51	69.70	51.51	54.54	72.73	48.48	51.51	30.30
shopping_mall	53.20	66.67	66.67	72.73	36.36	42.42	51.51	48.48	45.54	48.48
street_pedestrian	49.83	75.76	69.70	57.58	51.52	48.48	39.39	36.36	21.21	48.48
street_traffic	80.80	87.89	84.85	87.89	87.89	78.78	84.85	87.88	72.73	54.54
tram	63.29	75.76	54.54	81.82	87.88	42.42	75.76	42.42	57.57	42.42
Average	65.15	79.09	70.91	73.64	64.85	60.91	67.27	60.61	56.36	52.73

and the validation set. It also reminds us that we have a lot of work to do in the future.

IV. Results and Discussion

A. Validation Results of Mini-SegNet

Because the official evaluation set does not provide real scene category labels, our performance evaluation can only be conducted on the development set at present. In our work, we proposed a new Mini-SegNet network for acoustic scene classification. In addition, we applied spectrum correction processing and various data enhancement for different recording equipment. And in Table III, shows results for Mini-SegNet trained in various configurations using the official test-train split. Every configuration tested on both architectures.

Table III: Accuracy on the development dataset with and without spectrum correction, mixup, ImageDataGenerator and temporal crop.

Correction	*Yes*	*No*	*Yes*	*Yes*
Mixup	*Yes*	*Yes*	*Yes*	*Yes*
ImageDataGenerator	*Yes*	*Yes*	*No*	*Yes*
Temporal crop	*Yes*	*Yes*	*Yes*	*No*
All-accuracy	65.15%	64.41%	61.35%	63.53%

We do a variety of comparative experiments on whether spectrum correction, data augmentation or not. It can be seen from Table III that spectrum correction can improve performance, but the effect is not very good. Compared with temporal crop, ImageDataGenerator does have a good performance in acoustic scene classification.

B. Discussion

For this task, the best model we train is based on spectrum correction and various data augmentation. As shown in Table IV, we calculated the classification accuracy of various acoustic scenes of various devices under the prediction of the model. We observe from our best validation results that device-wise accuracy, and find that our model has good generalization performance. Devices-wise accuracy are generally higher than that of the baseline system. Especially under a high-quality device A, the accuracy reaches 79.09%. At the same time, in the low-quality recording devices B and C, 70.91% and 73.64% respectively are achieved. The generalization ability of some classes is better, but very poorly on the same others. Such as, airport is being classified as shopping_mall, metro_station as metro and public_square as park or street_pedestrian. In the future work, we will analyze these acoustic scenes which are wrongly classified.

In addition, our accuracy is higher than that of simulated devices S1-S6 in the baseline system. According to the accuracy of device S6 with poor results, the baseline system accuracy is 39.6%, while our accuracy is 52.73%. This shows that our system has good generalization ability on unknown data. However, compared with the real equipment, the accuracy of acoustic scene classification is still very poor. This might be attribute to the simulated data itself, such as the audio quality is not clear, and typical representative sounds in acoustic scenes are not significant. In our understanding, the acoustic scene is composed of some sound events. Through the semantic sound segmentation of different acoustic scenes, we can locate the frequency, location and other information of sound events, so as to classify different acoustic scenes. However, the classification accuracy of simulated equipment is poor, which may be due to the addition of some mixed and confused audio in synthesis. This also suggests that the high recognition accuracy of acoustic scene has certain requirements on the quality of audio data. In the future, our work will focus on this part of the simulated data and a more robust system will be construct on the basis of this work.

V. CONCLUSIONS

Based on the composition of sound events in acoustic scenes, we propose a semantic segmentation encoder-decoder network Mini-SegNet for acoustic scene classification (ASC). It is effective in addressing the multiple devices in the ASC task provided by DCASE 2020. In this task, we used spectrum correction to combat mismatched frequency responses. And, we proposed Mini-SegNet to extract multi-granularity abstract features of sound events in acoustic scenes. Our experiments verify that our approach achieves good performance under single system and no data expansion. Besides, we also demonstrate that spectrum correction is able to improve the classification accuracy of multiple devices, but the effect is not very good. In closing, the DCASE 2020 task1a confirm that our proposed encoder-decoder network Mini-SegNet is useful in acoustic scene classification.

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*, 2015, 32(3):16-34.

[2] Eronen, Antti J, et al. "Audio-based context recognition," *IEEE Transactions on Audio Speech and Language Processin*g, 2006, 14(1): 321-329.

[3] R. Radhakrishnan, A. Divakaran, and P. Smaragdis, "Audio analysis for surveillance applications," in *Proc. IEEE Workshop Applications of Signal Processing to Audio and Acoustics, 2005*, pp.158-161.

[4] S. Chu, S. Narayanan, C. C. J. Kuo and M. J. Mataric, "Where am I? Scene Recognition for Mobile Robots using Audio Features," *in Proc. IEEE International Conference on Multimedia and Expo*, Toronto, Ont., 2006, pp. 885-888.

[5] http://dcase.community/

[6] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in *Proc. IEEE International Conference on Acoustic, Acoustics, Speech, and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 171-175.

[7] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Acoustic Scene classification with matrix factorization for unsupervised feature learning," in *Proc. IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp.6445-6449.

[8] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Feature Learning with matrix factorization applied to acoustic scene classification," *IEEE Trans. Audio, Speech, Lang. Process*., 2017, 25(6):1216-1228.

[9] J T Geiger, K Helwani, "Improving event detection for audio surveillance using gabor filterbank features," in *23rd European Signal Processing Conference (EUSIPCO)*, Nice, France, Aug. 2015, pp.714-718.

[10] J. Salamon, J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," *IEEE Signal Processing Letters*, 2017, 24(3):279-283.

[11] Z. Ren, V. Pandit, K. Qian, et al, "Deep sequential image features on acoustic scene classification," in *Workshop on Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE 2017)*, November 2017.

[12] H. Jallet, E. Cakir, and T. Virtanen, "Acoustic scene classification using convolutional recurrent neural network," in *Workshop on Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE 2017)*, November 2017.

[13] Z. Ren, Q. Kong, J. Han, M. D. Plumbley and B. W. Schuller, "Attention-based Atrous Convolutional Neural Networks: Visualisation and Understanding Perspectives of Acoustic Scenes," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, May. 2019, pp. 56-60.

[14] http://dcase.community/challenge2020/task-acoustic-scene-classification

[15] T. Nguyen and F. Pernkopf, "Acoustic scene classification using a convolutional neural network ensemble and nearest neighbor filters," in *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2018)*, November 2018.

[16] Y. Sakashitaand M. Aono, "Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions," *Tech. Rep., DCASE 2018 Challenge*, 2018.

[17] M. Plata, "Deep neural networks with supported clusters preclassification procedure for acoustic scene recognition," *Tech. Rep., DCASE 2019 Challenge*, June 2019.

[18] H. Phan L. Pham, I. McLoughlin and R. Falaniappan, "A multi-spectrogram deep neural networks for acoustic scene classification," *Tech. Pep., DACSE 2019 Challenge*, June 2019.

[19] S. Mun, S. Park, D. Han, and H. Ko, "Generative adversial network based acoustic scene training set augmentation and selection using SVM hyper-planne," *Tech. Rep., DCASE 2017 Challenge*, September 2017.

[20] S. S. R. Phaye, E. Benetos and Y. Wang, "SubSpectralNet – Using Sub-spectrogram Based Convolutional Neural Networks for Acoustic Scene Classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, May. 2019, pp. 825-829.

[21] M. D. McDonnell and W. Gao, "Acoustic Scene Classification Using Deep Residual Networks with Late Fusion of Separated High and Low Frequency Paths," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May. 2020, pp. 141-145.

[22] Michal Kosmider, "Calibrating neural networks for secondary recording devices," *Tech. Rep., DCASE 2019 Challenge*, June 2019.

[23] Mark D, Mc Donnel, Wei Gao, "Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths," *Tech. Rep., DCASE 2019 Challenge*, June 2019.

[24] Hamid Eghbal-zadeh, Khaled Koutini and Gerhard Widmer, "Acoustic Scene Classification and Audio Tagging with Receptive-Field-Regularized CNNs," *Tech. Rep., DCASE 2019 Challenge*, June 2019.

[25] Hongwei Song and Hao Yang, "Feature Enhancement for Robust Acoustic Scene Classification with Device Mismatch," *Tech. Rep., DCASE 2019 Challenge*, June 2019.

[26] Badrinarayanan, Vijay, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *IEEE Transactions on Pattern Analysis & Machine Intelligence* (2017):1-1.

[27] H. Zahng, M. Cisse, Y. N. Dauphin, and D. Loped-paz, "mixup: beyond empirical risk minimization," arxiv preprint arxiv:1710.09412, 2017.

[28] J. Cramer, H. Wu, J. Salamon and J. P. Bello, "Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings," in P*roc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, 2019, pp. 3852-3856.