

A TWO-STAGE APPROACH TO DEVICE-ROBUST ACOUSTIC SCENE CLASSIFICATION

Hu Hu¹, Chao-Han Huck Yang¹, Xianjun Xia², Xue Bai³, Xin Tang³,
 Yajian Wang³, Shutong Niu³, Li Chai³, Juanjuan Li², Hongning Zhu², Feng Bao²,
 Yuanjun Zhao², Sabato Marco Siniscalchi^{1,4}, Yannan Wang², Jun Du³, Chin-Hui Lee¹

¹School of Electrical and Computer Engineering, Georgia Institute of Technology, GA, USA

²Tencent Media Lab, Tencent Corporation, China

³University of Science and Technology of China, HeFei, China

⁴Computer Engineering School, University of Enna Kore, Italy

{huhu, huckiyang, chinhui.lee}@gatech.edu, yannanwang@tencent.com, jundu@ustc.edu.cn

ABSTRACT

To improve device robustness, a highly desirable key feature of a competitive data-driven acoustic scene classification (ASC) system, a novel two-stage system based on fully convolutional neural networks (CNNs) is proposed. Our two-stage system leverages on an ad-hoc score combination based on two CNN classifiers: (i) the first CNN classifies acoustic inputs into one of three broad classes, and (ii) the second CNN classifies the same inputs into one of ten finer-grained classes. Three different CNN architectures are explored to implement the two-stage classifiers, and a frequency sub-sampling scheme is investigated. Moreover, novel data augmentation schemes for ASC are also investigated. Evaluated on DCASE 2020 Task 1a, our results show that the proposed ASC system attains a state-of-the-art accuracy on the development set, where our best system, a two-stage fusion of CNN ensembles, delivers a 81.9% average accuracy among multi-device test data, and it obtains a significant improvement on unseen devices. Finally, neural saliency analysis with class activation mapping (CAM) gives new insights on the patterns learnt by our models.

Index Terms— Acoustic scene classification, robustness, convolutional neural networks, data augmentation, class activation mapping

1. INTRODUCTION

Acoustic scene classification (ASC) refers to the task of identifying real-life sounds into environment classes, such as metro station, street traffic, public square, etc. An acoustic scene sound contains much information and rich content, which makes accurate scene prediction difficult and an intriguing research problem at the same time. In recent years, the organizers of the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge [1, 2, 3] provided both the benchmark data and a competitive platform to promote acoustic scene research and analysis. If we are to analyze top ASC systems reported in the challenge, we will find that most of them are built on the deep neural networks (DNNs) framework, and the key ingredient of their success is the use of convolutional layers [4, 5, 6, 7, 8]. Advanced deep learning techniques, such as attention mechanism [9, 10], mix-up [11, 12], generative adversarial network (GAN) and variational auto encoder (VAE) based data augmentation [13, 14], and deep feature learning [15, 16, 17] can further enhance ASC results. Although those ASC systems work well

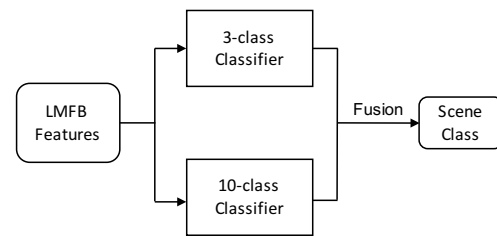


Fig. 1. The proposed two-stage ASC system.

on in-domain data, a generalization issue is observed in the presence of audio recording collected in mismatched testing conditions, e.g., audio segments recorded with devices different from those used in the training phase [2, 3]. Device robustness is an inevitable issue in a real production environment, and it is an important aspect to handle when designing an ASC system.

In this work, we have three main contributions: (i) we propose a novel two-stage ASC system; (ii) we investigate three different fully CNN models in the proposed two-stage system; (iii) we explore novel data augmentation strategies to reduce the device dependency of our models. Our experimental results are gathered on the DCASE 2020 task1a development data set, and the proposed CNN models achieve competitive results. Specifically, we attained a 79.4% overall ASC accuracy by ensemble of our three CNN models, and it obtains a significant improvement on unseen devices when compared with the baseline. The two-stage approach further boosts the accuracy of the ensemble model, and an ASC classification accuracy of 81.9% is attained. To better understand the key characteristics of our CNN based systems, we carry out a neural saliency analysis based on the class activation mapping (CAM) [18] on input data and the final output. Such analysis shows that our CNN models actually focus more on the meaningful audio segments containing acoustic events for the final classification task, such as bird sounds in a park or car horn sounds from street traffic, rather than leveraging on the background sound, as instead reported in [19].

2. TWO-STAGE ASC SYSTEM DESIGN

2.1. Two-Stage Classification Procedure

Figure 1 shows the proposed two-stage ASC system. The overall system consists of two independent classifiers and outputs the class of the input audio scene choosing among ten classes. Different from

a more conventional one-stage ASC systems, which directly feed acoustic features into the 10-class classifier and get final scene class, an extra general classifier, namely a 3-class classifier is introduced, which we expect it to enhance the classification process and leverage the over-fitting issue of 10-class classifier.

In our setup, the 3-class classifier classifies an input scene audio into one of three broad classes: in-door, out-door, and transportation. This 3-class classification way is from our prior knowledge that scene audios can be roughly categorized into such three classes. The 10-class classifier is actually the main classifier, which assigns a given input audio clip into one of ten target acoustic scene classes, including airport, shopping mall, metro station, pedestrian street, public square, street traffic, tram, bus, metro, park. Each audio clip should belong to one of the three / ten classes. The final acoustic scene class is chosen by the score fusion of those two classifiers. If we let \mathcal{C}^1 and \mathcal{C}^2 denote the set of three broad classes, and ten classes, respectively, and let F^1 and F^2 indicate the output of the first and second classifier, respectively. The final predicted class $Class(x)$ for the input x is:

$$Class(x) = \operatorname{argmax}_{q, (p \in \mathcal{C}^1, q \in \mathcal{C}^2, p \supset q)} F_p^1(x) * F_q^2(x),$$

where $p \supset q$ means that p can be thought of a super set of q . For example, transportation class is the super set for bus, tram, and metro classes. Therefore, the probability of an input audio clip to be from the public square scene is equal to the product of the probability of out-door place, $F_p^1(x)$, and that of public square, $F_q^2(x)$.

2.2. CNN based Classifiers and Ensemble

Three CNN based models are investigated to build our ASC system, namely Resnet, FCNN and fsFCNN. Resnet is a deep residual neural network [20] empowered by residual learning. Residual learning is a major efficient method to avoid gradient vanish issues while increasing the depth of networks to allow high-level feature learning. Our Resnet structure is based on the model in [21], which has no frequency sub-sampling throughout the whole network. Each input feature map is divided into two sub-feature mapping along the frequency dimension. To be specific, if we have N frequency bins, the first $N/2$ and the second half are processed by two parallel stacked residual layers. A global pooling layer and 10-way soft-max are used to get the final utterance level prediction results.

The FCNN (fully convolutional neural network) model is a VGG [22]-like model which is built with 9 stacked convolutional layers with small-size kernel. Each convolutional layer is followed by a batch normalization operation and ReLU activation function. Dropout is also used in order to alleviate over-fitting issues. Before the final global average pooling layer, channel attention [23] is applied to each output channel of the last layer. fsFCNN (frequency sub-sampling FCNN) is an extension of FCNN model which mainly has 2 more convolutional layers and reduces max-pooling size in the frequency axis. In our experiments we notice that it can help to leverage over-fitting issue.

Finally, we build an ensemble using those CNN models and make the final predictions. The output probabilities after softmax are summed by equal weights, and then the scene class with the maximum posterior represents the final classification decision. For the 3-class classifier, only Resnet and FCNN are used; whereas, for the 10-class classifier, all three CNN models are used. The two-stage score fusion is performed after model ensemble of classifiers in each stage.

2.3. Data Augmentation Strategies for Training

When training the CNN models, 9 different data augmentation methods are investigated. The training for 3-class models and 10-class models adopt the same data augmentation strategy. They can be categorized into two main classes: generating extra data or not.

Data augmentation strategies, which do not generate any extra data, include: (i) Mixup [24], which randomly mixes data batches with corresponding labels; (ii) Random cropping [21], which randomly crops input features into ones with a smaller length along the time axis; (iii) SpecAugment [25], which randomly masks input features by zero along both time and frequency axes. SpecAugment is performed on batch level, and the parameter of mask is set to 10% of the time and frequency dimensions, respectively.

Strategies that generate extra data include: (i) Spectrum correction [26], which aims at transforming a given input spectrum to that of a reference, possibly ideal, device. Different from the original idea, we newly employ spectrum correction as a data augmentation technique, where we create a reference device spectrum, by averaging the spectrum from all training devices except that from device A, and then correct the spectrum of each training waveform collected with device A to obtain extra data. (ii) Reverberation with Dynamic Range Compression (DRC), where we simulate more audio clips from ‘new devices’ by introducing reverberation and applying DRC - this data-augmentation solution is also new for ASC. Room impulse responses [27] are used to create reverberation into the device A data, and DRC is then applied to dynamically compress the amplitude range. Room impulse responses and DRC settings are randomly chosen to generate new training waveforms. (iii) Pitch shift, where we randomly shift the pitch of each audio clip based on the uniform distribution. (iv) Speed change, where we randomly change the audio speed based on the uniform distribution. If output waveform is longer than the original one, extra samples are dropped from the end, otherwise, padding is applied till attaining the same input length. (v) Random noise, we randomly add Gaussian noise to each training waveform. (vi) Mix audios, where we randomly mix two audios from the same acoustic scene class. It’s device-independent so it may help simulate a ‘new device’. Although the solutions from (iii) to (vi) are not new strictly speaking, their application for ASC device robustness has never been evaluated before.

3. EXPERIMENTS & ANALYSIS

3.1. Experimental Setup

All ASC systems are evaluated on the DCASE 2020 task1a development data set [3], which consists of $\sim 14\text{K}$ 10-second single-channel train audio clips and $\sim 3\text{K}$ test audio clips recorded by 9 different devices, including real devices A, B, C, and simulated device s1-s6. Only device A, B, C, s1-s3 are in the training set; whereas, devices s4-s6 are not available in the training phase. The greatest amount of training audio clips are recorded with device A, namely over 10K audio clips. In the test set, the number of audio clips from each device is the same. For each input audio clip, a short-time Fourier transform (STFT) with 2048 FFT points is applied, using a window size of 2048 samples and a hop length of 1024 samples. The Librosa [28] library is employed to extract log-mel filter bank (LMFB) features. Log-mel deltas and delta-deltas without padding are also computed, which finally generates an input feature tensor with size of $423 \times 128 \times 3$. Before feeding the input tensors into the CNN classifier, we perform utterance-level scaling operation to scale LMFB coefficients into $[0,1]$. All deep models in this work are built with Keras [29]. Stochastic gradient descent (SGD) with a cosine-decay-restart learning rate scheduler is used to train all deep models. Max-

Table 1. Classification accuracy of different 3-class models.

3-class Model	Resnet	FCNN	Ensemble
Acc. %	91.4	92.9	93.2

Table 2. Overall results on DCASE 2020 task1a development set. All proposed data augmentation strategies are used in all our models.

System	A %	B&C %	s1-s3 %	s4-s6 %	Avg. %
Baseline [3]	70.6	61.6	53.3	44.3	54.1
Suh et al. [30]	-	-	-	-	74.4
Gao et al. [31]	-	-	-	-	72.5
Liu et al. [32]	-	-	-	-	72.1
Koutini et al. [33]	-	-	-	-	73.3
Resnet	83.0	76.1	73.6	71.0	74.6
FCNN	87.3	79.5	75.7	73.0	76.9
fsFCNN	83.9	78.6	75.4	72.8	76.2
Ensemble	87.0	81.5	78.0	76.9	79.4
2-stage Resnet	84.5	78.6	76.2	76.4	77.7
2-stage FCNN	89.1	82.9	78.5	76.9	80.1
2-stage fsFCNN	83.9	81.2	78.6	76.4	79.0
2-stage Ensemble	87.9	84.1	80.4	79.9	81.9

imum and minimum learning rates are 0.1, and 1e-5, respectively.¹

3.2. Experimental Results

Table 1 shows the accuracy for 3-class classifiers, namely Resnet, FCNN and Ensemble of these two models. The 3-class classification task is relatively easy since an accuracy above 90% is attained by each of the three models. FCNN shows a better accuracy than Resnet, and the ensemble of these two models allows a further boost of the classification result. In the following experiments with the two-stage system, we use Ensemble system as the 3-class classifier.

All main experimental results concerning the 10-class ASC task are shown in Table 2. In the training set, device A data accounts for around 75%, and device B, C, s1-s3 accounts for around 5%, respectively. Thus, we can think of device A as a source device, and the remaining ones as target devices. Based on the device information of data, we divide the test set into four different subsets, which represent real source data (device A), real target data (device B & C), target seen-simulated data (device s1-s3), and target unseen-simulated data (device s4-s6). The first row in Table 2 gives ASC accuracy of the official baseline system [3], which uses a two-fully connected layer neural classifier, and OpenL3 [15] to extract input audio embedding. From baseline results, we can argue that good ASC accuracy can be attained on real source data (device A), but a severe performance drop is observed on other devices. Specifically, a severe ASC accuracy drop is reported on unseen devices (s4-s6), where the ASC accuracy is as low as 44.3%. Rows from 6th through 9th in Table 2 display ASC results attained with a 10-class classifier. Resnet attains a significant improvement on all test devices over the baseline system. FCNN and fsFCNN outperform Resnet on all testing scenarios, and FCNN achieves slightly better results than fsFCNN. Finally, the ensemble of these three models attains a 79.4% overall ASC accuracy, which is a meaningful improvement compared to any of the single-stage models reported in Table 2.

The experimental results of two-stage systems are shown in the

¹Code available: https://github.com/MihawkHu/DCASE2020_task1

Table 3. Accuracy comparison of different data augmentation strategies on Resnet. Mixup and random cropping are always employed. ‘sa’ indicates specAugment. ‘sc’ indicates spectrum correction. ‘r’ indicates reverberation with DRC. ‘aug’ indicates another four augmentation methods, including pitch shift, speed change, random noise and mix audios.

System	A %	B&C %	s1-s3 %	s4-s6 %	Avg. %
Resnet	78.8	72.1	69.3	69.5	71.0
+sa	80.3	73.5	71.4	67.7	71.6
+sa+sc	79.1	75.0	70.7	68.9	72.0
+sa+sc+r	80.3	74.7	71.4	70.3	72.8
+sa+sc+r+aug*	83.0	76.1	73.6	71.0	74.6

last four rows of Table 2. When comparing single-stage 10-class classifier with the two-stage system, a significant improvement can be attained. For example, comparing ‘Resnet’ with ‘2-stage Resnet’, 3.1% absolute improvement on overall accuracy is observed (74.6% vs. 77.7%). Accuracies on all the test sets have indeed improved, especially on the unseen-simulated test set (device s4-s6), where a 5.4% absolute improvement is obtained. The overall absolute accuracy gains with the proposed two-stage technique are 3.2% for FCNN, 2.8% for fsFCNN, and 2.5% for the ensemble. The best performance is attained by the two-stage of three models’ ensemble, which achieves 81.9% overall ASC accuracy. Although the top performance is still delivered on the source device A (87.9%), the average ASC accuracy on the remaining test devices is above 79%, which compares favourably against the baseline system. When comparing ASC accuracy between seen-simulated test set (device s1-s3) and unseen-simulated test set (device s4-s6), we can argue that the gap in performance is mild (80.4% vs. 79.9%), and they are close to the overall average accuracy (80.4% vs. 79.9% vs. 81.9%). Those results confirm the effectiveness of the proposed two-stage system with respect to improving device robustness.

To better compare our results, the results of top-5 ranking systems (excluding ours) [30, 31, 32, 33] are also listed from 2st to 5th rows in Table 2. The experimental results clearly demonstrated that the proposed two-stage system achieves a competitive performance with all benchmark models.

3.3. Evaluation of Data Augmentation Strategies

We now investigate the effectiveness of different data augmentation strategies in combination. We test only the Resnet-based ASC system to keep the experimental setup consistent. Table 3 shows the classification accuracy attained with different data augmentation schemes. It should be noted that mixup and random cropping are always employed in the training phase.

When comparing the first two rows, we can see that specAugment can boost the overall ASC accuracy from 71.0% to 71.6%, but on unseen data the accuracy becomes worse (from 69.5% to 67.7%). Next, we add spectrum correction, which results in an increase in the amount of training data. As described before, device A data is used to generate data from other devices, in which we get the same size extra data with device A data. From the results in Table 3, we can observe that spectrum correction has a beneficial effect on overall classification results, but the accuracy on device A decreases. That outcome is not unexpected since there is now much more training data from other devices than from device A. New extra data is generated by reverberation with DRC, which aims to simulate data from more different devices. From the results we can see when extra data with

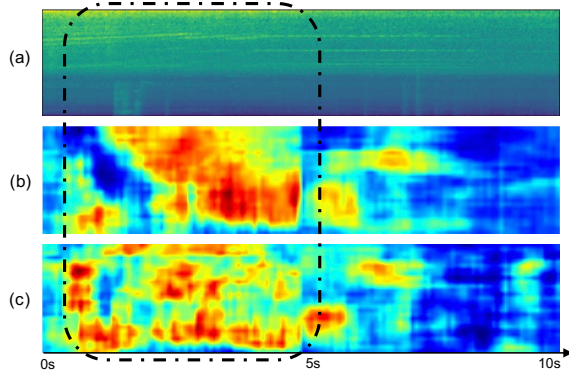


Fig. 2. Neural saliency analysis via CAM of an audio clip from a metro station (*metro_station_vienna_87_2389_a.wav*). The three plots are: (a) spectrogram, (b) CAM of 3-class classifier, (c) CAM of 10-class classifier. In this 10-second audio, brake and horn sound starts from 0s to around 8s and only reverberation remains after 5s.

reverberation and DRC is added in the training set, the overall ASC accuracy is further boosted (72.8%), where most of the gain is imputable to an increase of the accuracy on the unseen test data. Finally, we employ four more data augmentation approaches, including pitch shift, speech change, random noise and mix audios. As shown in the last row of Table 3, performance can be further boosted. Using all strategies, a 74.6% accuracy on Resnet is attained².

3.4. Neural Saliency Analysis via Class Activation Mapping

Neural saliency methods [18, 34] aim to provide interpretable analyses on the weight distribution over hidden neurons of a well-trained artificial neural network model. Benchmark saliency methods, such as class activation mapping (CAM) [18], have shown a correlation between model performance and ability in spotting physical pattern in DNN-based acoustic models [34].

In this work, we use CAM [18] to gain a better understanding about what sound patterns are found by our CNNs to accomplish the ASC task. Indeed, CAM can highlight the class-specific discriminative regions in the input feature map. In other words, CAM helps generating a two-dimension activation map that can be used to better interpret the prediction decision made by deep architectures. In our case, where an audio clip is transformed into a time-frequency representation, CAM reveals whether the ASC decision is triggered by CNN’s attention posed on time and frequency regions of input audio clip having a meaningful semantic content with respect to the target acoustic scene. Examples of some analysis results obtained with CAM are shown in Figures 2 and 3, where the spectrogram of the input audio clip (top panel), and the CAM for the 3-class classifier (center panel), and the 10-class classifier (bottom panel) are shown. Regions in CAM with deep red color imply that the CNN pays more attention to them, and those regions have a higher class-specific discriminative power. As the size of FCNN and fsFCNN’s output feature map is too small to map back to the input, only Resnet model is used for neural saliency analysis.

In Figure 2, a 10-second audio clip referring to a metro station scene is visualized; the audio clip contains brake and horn sounds from 0s to around 5s. After 5s, only reverberation remains in the audio. The CAM result indeed reveals that the CNN pays more attention to the segment between 0s and 5s. Moreover, 3-class classifier and 10-class classifier have similar activation regions, and the

²We also increased amount of Resnet parameter, i.e., the channel number, when all data-augmentation schemes are used.

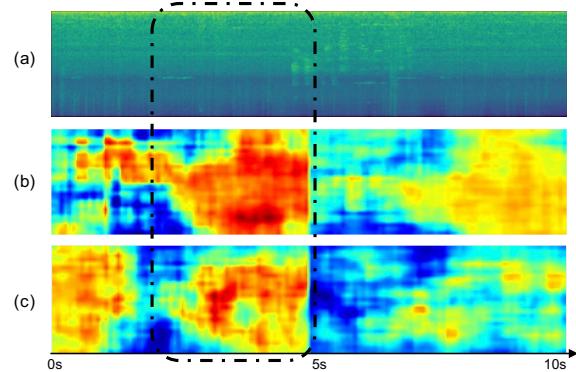


Fig. 3. Neural saliency analysis via CAM of an audio clip from a bus station (*bus_prague_1102_42431_a.wav*). The three plots are (a) spectrogram, (b) CAM of 3-class classifier, (c) CAM of 10-class classifier, respectively. In this 10-second audio, brake sound starts from around 2s to around 5s, human talk starts from around 5s to 7s.

3-class classifier seems to give more weight to the information in that region. Indeed, if we check the confidence in making the prediction, i.e., scaled class posterior probability from the CNN, the proposed two-stage fusion can increase it from 70.7% to 85.0%. The same pattern can be observed in the results of Figure 3, in which it contains a 10-second audio from a bus station. This audio has brake sound between around 2s and 5s, and human speech starts around 5s till 7s. From the CAM results, we can see that CNN pays more attention to the segment from 2s to 5s, which contains a brake sound. However, the CNN is not interested in the human speech although it is clearly audible and related to the bus arriving announcement. This result is indeed interesting: Human beings would use the bus-station announcement to make ASC decision; however, the CNN cannot perform speech recognition, so it may simply classify that region as belonging to an ideal human speech class, which however appears in many scenes and has therefore a low discriminative power.

We can thus argue that CNNs use acoustic events to classify the input scene audios. Specifically, CNNs pay more attention to audio segments containing acoustic events, such as birds sound from park, or car horn sound from street traffic. It should be noted that our conclusion with CAM is different from that in [19], where the authors claim that CNNs use background sound rather than acoustic events. By comparing the experimental setups, the main difference is that each audio clip is chunked into 10 1-second segments in their setup; in contrast, we use the whole 10-second audio clip as the input. Since acoustic events usually last for several seconds, it’s difficult for CNNs to capture them when inputs are very short.

4. CONCLUSION

This work focuses on device robustness in acoustic scene classification. We propose a novel two-stage ASC framework based on CNNs. Three different fully CNN models and several novel data augmentation strategies are investigated to improve device robustness. A general 3-class classifier and a specific 10-class classifier are combined through score fusion. Experiments on the DCASE 2020 task1a development set show the effectiveness of our solution. Specifically, our best system, a two-stage fusion of a CNN-based ensemble, obtains a state-of-the-art 81.9% average ASC accuracy. Moreover, our CAM-based neural saliency analysis demonstrates that CNNs pay particular attention to audio segments strictly related acoustic events rather than simply fetching information from the background environmental sound.

5. REFERENCES

- [1] A. et al. Mesaros, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *DCASE*, 2017, pp. 85–92.
- [2] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *DCASE*, 2018, pp. 9–13.
- [3] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions," in *DCASE2020*, 2020.
- [4] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the dcase 2016 challenge," *IEEE/ACM T-ASLP*, vol. 26, no. 2, pp. 379–393, 2018.
- [5] Y. Han and K. Lee, "Acoustic scene classification using convolutional neural network and multiple-width frequency-delta data augmentation," *arXiv preprint arXiv:1607.02383*, 2016.
- [6] D. Battaglino, L. Lepauloux, and N. Evans, "Acoustic scene classification using convolutional neural networks," Tech. Rep., DCASE2016 Challenge, 2016.
- [7] H. Hu, C.-H. Yang, X. Xia, X. Bai, X. Tang, Y. Wang, S. Niu, L. Chai, J. Li, H. Zhu, F. Bao, Y. Zhao, S. M. Siniscalchi, Y. Wang, J. Du, and C.-H. Lee, "Device-robust acoustic scene classification based on two-stage categorization and data augmentation," Tech. Rep., DCASE2020 Challenge, June 2020.
- [8] Lam Pham, Ian McLoughlin, Huy Phan, and Ramaswamy Palaniappan, "A Robust Framework for Acoustic Scene Classification," in *Proc. Interspeech 2019*, 2019, pp. 3634–3638.
- [9] J. Wang and S. Li, "Self-attention mechanism based system for dcase2018 challenge task1 and task4," in *DCASE*, 2018, pp. 1–5.
- [10] J. Guo, N. Xu, L.-J. Li, and A. Alwan, "Attention based cldnns for short-duration acoustic scene classification," in *Interspeech*, 2017, pp. 469–473.
- [11] K. Xu et al., "Mixup-based acoustic scene classification using multi-channel convolutional neural network," in *Pacific Rim Conference on Multimedia*, 2018, pp. 14–23.
- [12] T. Nguyen and F. Pernkopf, "Acoustic scene classification with mismatched devices using cliquenets and mixup data augmentation," in *Interspeech*, 2019.
- [13] H. Chen, Z. Liu, Z. Liu, P. Zhang, and Y. Yan, "Integrating the data augmentation scheme with various classifiers for acoustic scene modeling," Tech. Rep., DCASE2019 Challenge, June 2019.
- [14] S. Mun, S. Park, D. K. Han, and H. Ko, "Generative adversarial network based acoustic scene training set augmentation and selection using svm hyper-plane," *DCASE*, pp. 93–97, 2017.
- [15] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *ICASSP*, 2019, pp. 3852–3856.
- [16] S. Mun, S. Shon, W. Kim, and H. Ko, "Deep neural network bottleneck features for acoustic event recognition," in *Interspeech*, 2016, pp. 2954–2957.
- [17] P. Sharma, V. Abrol, and A. Thakur, "Ase: Acoustic scene embedding using deep archetypal analysis and gmm," in *Interspeech*, 2018, pp. 3299–3303.
- [18] B. Zhou, A. Khosla, Lapedriza, A., A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," *CVPR*, 2016.
- [19] Y. Wu and T. Lee, "Enhancing sound texture in cnn-based acoustic scene classification," in *ICASSP*, 2019, pp. 815–819.
- [20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
- [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 *ICASSP*, 2020, pp. 141–145.
- [22] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [23] Q. Bi, K. Qin, Z. Li, H. Zhang, K. Xu, and G.-S. Xia, "A multiple-instance densely-connected convnet for aerial scene classification," *IEEE Transactions on Image Processing*, vol. 29, pp. 4911–4926, 2020.
- [24] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.
- [25] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.
- [26] T. Nguyen, F. Pernkopf, and M. Kosmider, "Acoustic scene classification for mismatched recording devices using heated-up softmax and spectrum correction," in *ICASSP*, 2020, pp. 126–130.
- [27] Keisuke Kinoshita, Marc Delcroix, Sharon Gannot, Emanuël AP Habets, Reinhold Haeb-Umbach, Walter Kellermann, Volker Leutnant, Roland Maas, Tomohiro Nakatani, Bhiksha Raj, et al., "A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 7, 2016.
- [28] B. McFee, C. Raffel, D. Liang, D.P.W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, vol. 8.
- [29] F. Chollet et al., "Keras," <https://keras.io>, 2015.
- [30] S. Suh, S. Park, Y. Jeong, and T. Lee, "Designing acoustic scene classification models with CNN variants," Tech. Rep., DCASE2020 Challenge, June 2020.
- [31] W. Gao and M. McDonnell, "Acoustic scene classification using deep residual networks with focal loss and mild domain adaptation," Tech. Rep., DCASE2020 Challenge, June 2020.
- [32] L. Jie, "Acoustic scene classification with residual networks and attention mechanism," Tech. Rep., DCASE2020 Challenge, June 2020.
- [33] K. Koutini, F. Henkel, H. Eghbal-zadeh, and G. Widmer, "CP-JKU submissions to DCASE'20: Low-complexity cross-device acoustic scene classification with RF-regularized CNNs," Tech. Rep., DCASE2020 Challenge, June 2020.
- [34] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *ECCV*, 2018, pp. 631–648.