Fusing Recalibrated Features and Depthwise Separable Convolution for the Mangrove Bird Sound Classification

Chongqin Lei LEICHONGQIN@CQU.EDU.CN
Weiguo Gong WGGONG@CQU.EDU.CN
Zixu Wang 201808021045@CQU.EDU.CN
Key Lab of Optoelectronic Technology and Systems of Education Ministry, College of Optoelectronic Engineering, Chongqing University, Chongqing, China

Editors: Wee Sun Lee and Taiji Suzuki

Abstract

The bird community in the mangrove areas is an important component of the mangrove wetlands ecosystem and an indicator species for the assessment of the environmental health status of mangrove wetlands. The classification of bird species by the sound of bird in the mangrove areas has the advantages of less interference to the environment and wide monitoring range. In this paper, we propose a novel method that combines the feature recalibration mechanism with depthwise separable convolution for the mangrove bird sound classification. In the proposed method, we introduce Xception network in which depthwise separable convolution with lower parameter number and computational cost than traditional convolution can be stacked in a residual manner, as the baseline network. And we fuse the feature recalibration mechanism into the depthwise separable convolution for actively learning the weights of the feature channels in the network layer, so that we can enhance the important features in bird sound signals to improve the performance of the classification. In the proposed method, firstly we extract three-channel log-mel features of the bird sound signals and we introduce the mixup method to augment the extracted features. Secondly, we construct the recalibrated feature maps including the different scales of information to get the classification results. To verify the effectiveness of the proposed method, we build a dataset with 9282 samples including 25 kinds of the mangrove birds such as Egretta alba, Parus major, Charadrius dubius, etc. habiting in the mangroves of Fangcheng Port of China, and execute the experiments on the built dataset. Furthermore, we also validate the adaptability of our proposed method on the dataset of TAU Urban Acoustic Scenes 2019, and achieve a better result.

Keywords: mangrove bird sound classification, feature recalibration, bird sound dataset, mixup

1. Introduction

The acoustic monitoring system is very popular as a non-invasive method to study the number and community of the vocal animals. It can provide information about biodiversity and its spatial-temporal distribution changes Kelling et al. (2012). In bird sound classification, the most widely used feature-based conventional modeling techniques include Hidden Markov Model (HMM) (Potamitis et al. (2014); Chou et al. (2007); Jančovič et al. (2014)), Gauss Mixture Model (GMM) Ganchev et al. (2015), Support Vector Machine (SVM) (Fagerlund (2007); Tran and Li (2010)), and Template Matching Kaewtip et al.

(2016). The success of the bird sound classification methods based on GMM and HMM depends on the applicability of the audio parameterization process, especially the segmentation and selection of representative parts of specific species sounds. And the template-based bird sound classification algorithms, such as dynamic time warping (DTW), have very large computational costs. In general, the traditional bird sound classification often requires high applicability and computational cost of audio parameterization. However, the method of deep learning does not require high parameterization of features and the computational cost is relatively small.

With the rise of deep learning, a large number of research fields have introduced the deep learning methods such as convolutional neural network (CNN) to solve the key and difficult problems in the field. Since the introduction of AlexNet Krizhevsky et al. (2012) in 2012, mensional signal processing. The only difference is that for the sound signals, we need to extract features to obtain two-dimensional or even higher-dimensional feature vectors that can be input into CNN. In the competitions such as BirdCLEF Bird Challenge Joly et al. (2018) and DCASE Acoustic Scene Classification Plumbley et al. (2018), the best results of solutions submitted by participants before 2016 are using traditional methods such as template matching and dictionary learning Salamon and Bello (2015). Since then, the best models for acoustic classification tasks are based on CNN.

The two elements of deep convolution network applied to the bird sound classification task are effective input features and appropriate network structures. Firstly, an effective model input can maximize its representation ability. So in order to get good performance of the mangrove bird sound classification, in this paper, we set three-channel log-mel features as the input of the network. Secondly, because the appropriate network structure is the key to get good accuracy of classification and a network with enough learning ability can maximize the classification accuracy without serious overfitting, we introduce the Xception network [27] as the baseline network, which is a network that introduces depthwise separable convolution Sifre and Mallat (2014) and stacks them in a residual manner. Taking into account the fact that the bird sound data contains less effective target information, we construct a network by embedding the Squeeze-and-Excitation (SE) block Hu et al. (2018) in depthwise separable convolution for recalibrating the channels of the features extracted from each layer of the network, which enables the constructed network to actively learn useful features and discard useless features for improving the classification accuracy of the mangrove bird sound classification. In addition, we concatenate the recalibrated feature maps including the different scales of information to get the classification results. Furthermore, deep neural networks have a large number of parameters and the bird sound data has few samples or few effective target information in the samples, which are extremely easy to result in an over-fitting problem. In the field of the image processing, although the random clipping, scaling and other data augmentation methods can be applied to solve the over-fitting problem, they are not effective in the bird sound classification. So we introduce the mixup method Zhang et al. (2017) for relieving the over-fitting and augmenting the mangrove bird sound data.

In this paper, we collect and build the mangrove bird sound dataset including 25 kinds of birds in Beilun River Estuary Mangrove National Nature Reserve of Guangxi, China, and execute the experiments on the built sound dataset to verify the effectiveness of the proposed method. Furthermore, we use DCASE 2019 acoustic scene classification dataset

to validate the effectiveness of the proposed method. The main contributions of this paper are highlighted as follows:

- We build a sound dataset with 25 kinds of birds, which is specially for the task of bird species identification in the mangrove areas.
- We propose a novel method that fuses the feature recalibration and the depthwise separable convolution to effectively classify the bird sound in the mangrove areas, and demonstrate the good adaptability in acoustic scene classification tasks.
- We introduce the mixup method for augmenting the mangrove bird sound data, and construct the multi-scale recalibrated feature maps of the network to further improve the mangrove bird sound classification accuracy.

The rest of this paper is organized as follows. Section 2 gives a brief review of the related work. Then we formulate the problem and present the proposed method in section 3. Section 4 is the experiments and discussions. Finally, Section 5 is our conclusions.

2. Related works

The traditional modeling methods include GMM and HMM, template matching and so on. Kwan et al. (2006) summarized the HMM and GMM methods in bird sound classification. It can be concluded that both modeling methods need a lot of parameterized calculation of features, and the requirement of parameterized applicability is extremely strict. Kaewtip et al. (2016) proposed a template-based algorithm for bird sound classification, which can be applied in limited training data or noisy environment. This algorithm used DTW and the prominent region (i.e. high-energy) of the training data acoustic spectrum to obtain the template. Ruiz-Muñoz et al. (2018) proposed a random projection dictionary learning approach for a bioacoustics application, which combines the power of estimating spectratemporal patterns given by the convolutive model and the computational complexity savings associated with the random projection approach. However, modeling based on deep learning method does not require very high requirements for feature parameterization, and the computational overhead is not too high under reasonable circumstances. Kiskin et al. (2018) presented that a CNN outperforms generic recordings, where the wavelet-trained CNN outperforms traditional classification algorithms with no hyper parameter re-tuning of either approach. Tóth and Czeba (2016) proposed a method of bird sound classification based on CNN to fine-tuned the classification of 1500 species for BirdCLEF 2018, a challenge of bird classification based on audio recording. Chakraborty et al. (2016) extracted Mel Cepstrum Coefficient (MFCC) features from bird sound recordings in the lower Himalayas and inputs them into SVM and DNN classifiers. Xie et al. (2018) designed a bird sound recognition model based on transfer learning, which uses VGG-16 model (pretrained on ImageNet) to extract features, and then adds a classifier composed of two fully connected hidden layers and a SoftMax layer.

The CNN extracts abstract features by merging spatial information on a channel-by-channel basis using local receptive fields Jacobsen et al. (2016). It is much difficult to train a performance-efficient network using features with less effective target information in bird

sound classification tasks. Generally speaking, the performance optimization of classification networks can be considered form two aspects. Firstly, from the perspective of spatial dimension, for example, the Inception structure Szegedy et al. (2015) embeds multi-scale information and aggregates features of different receptive fields to improve performance. Furthermore, considering the characteristics of bird sound data, we want to use lightweight networks as much as possible for our work, so we consider introducing a depthwise separable convolution model. Combined with the above analysis, we choose Xception (Extreme Inception) Chollet (2017), which is a network that introduces depthwise separable convolution and stacks them in a residual manner based on Inception V3, as the baseline model of this paper. And we improve the classification performance of the network by combining feature maps of different scales as input to the classifier. Secondly, according to the relationship of feature channels, the SE Network selectively enhances the useful features and compresses useless features by using global information. Inspired by this thought, we introduce the SE block and embed it into the backbone structure of Xception, which can explicitly model the channel correlation between convolution layer features to improve the representation ability.

3. Proposed method

In this section, we first review the related works of the feature recalibration and the depthwise separable convolution, and then propose our method.

3.1. Feature Recalibration

Hu et al. (2018) proposed the SE block to enhance the accuracy by modeling the correlation between feature channels and strengthening the important features. The SE block is the representative of the channel-wise recalibration of the feature maps. Figure 1 gives a detailed description of the SE block.

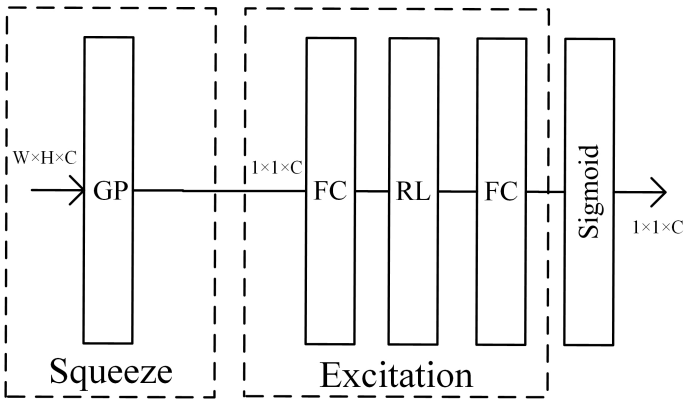

Figure 1: The detailed structure diagram of SE block, where "GP" indicates the global pooling operation, "FC" represents full connection layer, and "RL" means activation function of ReLu.

As shown in Figure 1, the Global Average Pooling operation is performed on the input feature maps of size $W \times H \times C$, which is the Squeeze process in the figure. Then the output data of size $1 \times 1 \times C$ is fed into two fully connected layers, which is the Excitation process in the figure. Finally, the output is limited to the range of [0, 1] by the activation function Sigmoid.

The principle of SE block is to enhance the important features and weaken the unimportant ones by controlling the channel weights of the features, so as to make the extracted features more directional. For the mangrove bird sound data, the features extracted from 5 seconds audio clips contain many redundant background noises, so we need to suppress these insignificant noises and enhance the target bird sound information to improve the classification performance. Therefore, it is necessary to introduce a feature recalibration mechanism.Figure 2 shows in detail the feature spectrograms extracted from 25 kinds of the bird sound signals in this paper. As can be seen from the Figure 2, the proportion of the highlights representing the sounds of target bird in each feature map is very small.

3.2. Depthwise Separable Convolution

The depthwise separable convolution was first proposed by Sifre and Mallat (2014) in 2013. It decomposes the traditional convolution into a depthwise convolution and a 1×1 channel convolution. As shown in Fig 3, (a) is the traditional convolution operation, (b) and (c) correspond to the depthwise convolution and the 1×1 channel convolution of the depthwise separable convolution respectively. As shown in Fig 3, (a) shows a conventional convolution operation in which the convolution kernel size is $D_K \times D_K \times M$, and the number of the output feature maps is N. Traditional convolution is a one-step operation, while the depthwise separable convolution integrates traditional convolution into two-step operations as shown in (b) and (c) respectively. First, as shown in (b), a spatial convolution with convolution kernel size of $D_K \times D_K \times 1$ is performed, and then as shown in (c), it shows a channel convolution with convolution kernel size of $1 \times 1 \times M$.

3.3. Feature Recalibrated Depthwise Separable Convolution

Inception structure is designed to achieve the highest classification precision in classification tasks. Since its first introduction, Inception has been one of the best performing models for both ImageNet datasets and Google internal datasets, especially JFT-300M. Therefore, when selecting baseline models in the mangrove bird sound classification tasks, we focus on various improved networks of Inception structure. Furthermore, the dataset of the mangrove bird sound classification task has the characteristics with less data and less effective target information in the features extracted from one 5-second segment, and the advantages of Xception network, such as easy migration, less computation, adaptability to the different tasks and high accuracy, can solve these problems. So we choose the Xception network as the baseline and add the SE block to it to enable the network to automatically acquire the important information of each feature channel through training. Figure 4 shows the overall network architecture. There are 12 Blocks in the backbone network, each one contains similar convolution and pooling operations. As shown in Figure 4, the connection marked red in the enlarged Block structure is the SE block that we introduce. The details of the SE block are further explained in the upper right enlarged Block structure. We add the SE

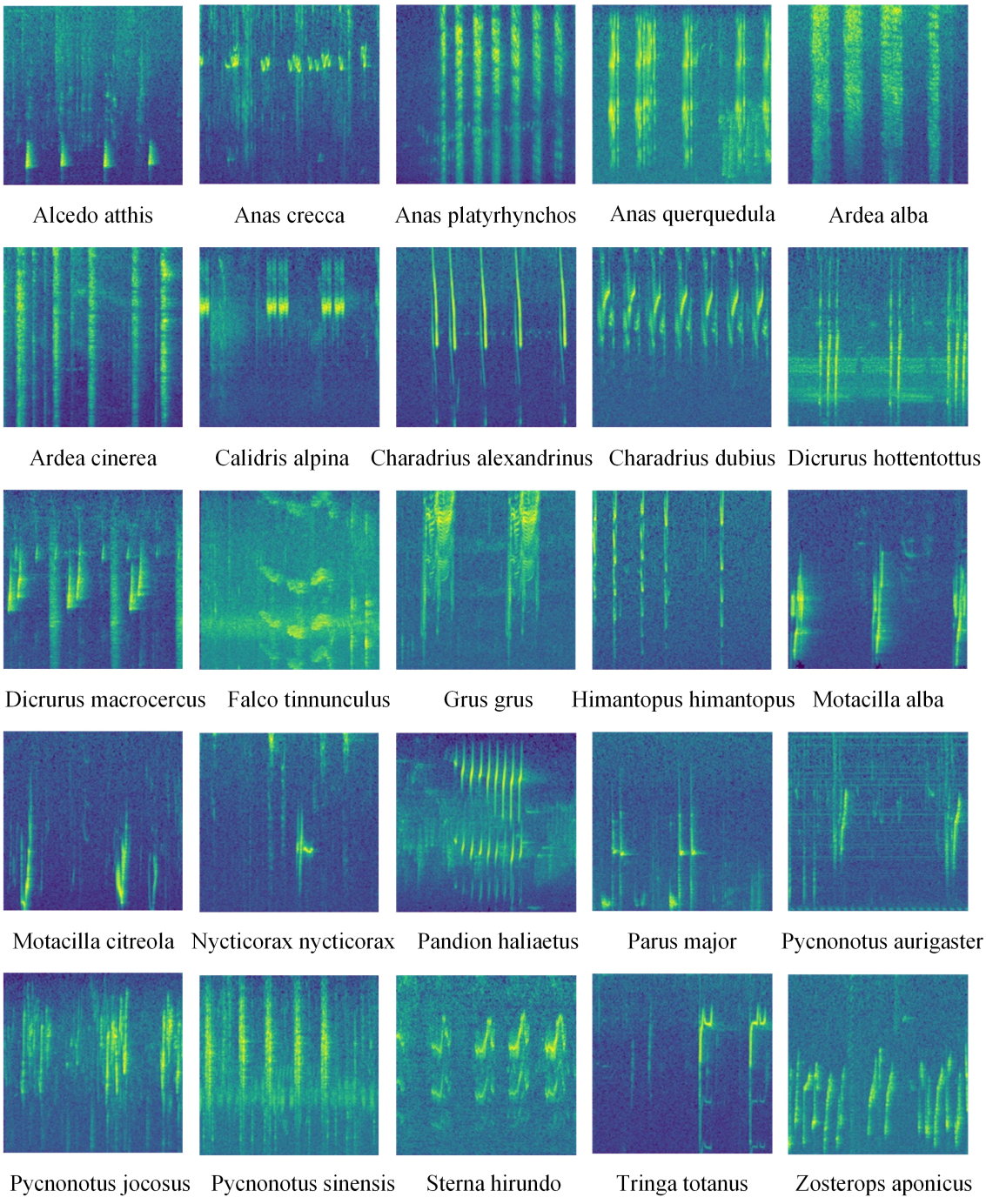

Figure 2: The Log-mel feature Spectrograms extracted from 25 mangrove bird sound signals.

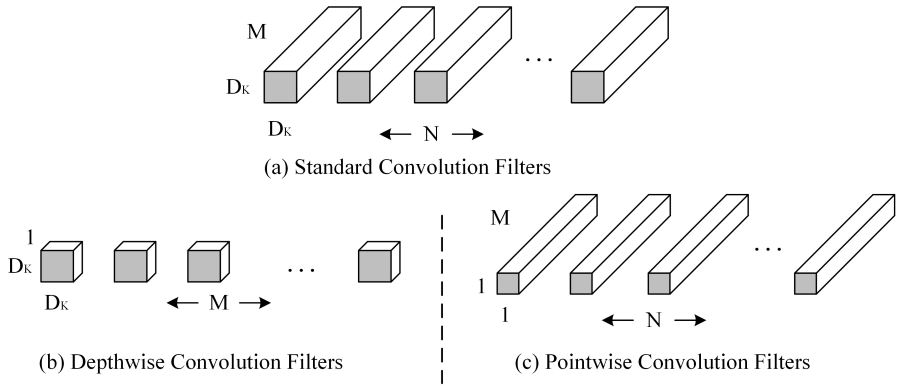

Figure 3: The difference between traditional convolution operation and depthwise separable convolution operation.

block in front of the Maxpooling layers of each Block in Xception network, and its output is added to the output of the residual connection and then used as the input to the next Block.

4. Experiments and Discussions

4.1. Building of Mangrove Bird sound Dataset

The collected bird sounds involve a lot of background noises. For example, the sounds collected in the remote mountains may contain buzzing insect sounds, while the sounds recorded in the edge of human settlements may contain the noise of motor vehicles. Other non-bird sound, such as wind sound and rain sound, are also common background noises. And the studies of birds are mostly based on geographical or environmental distribution, such as the study of birds in the Amazon rainforest. So it is always difficult to collect these data, whether in image classification or sound classification of birds. In our work, we collected the related mangrove bird sound through Xeno-canto in the early stage, and we selected 25 kinds of bird sound in Guangxi mangrove areas on Xeno-canto to build the dataset for verifying the effectiveness of the proposed method. The data collected on Xeno-canto mainly include the records with subjective evaluation of A or B grade. We did not distinguish the location and altitude of records in detail, and we tried to select the audio with better quality and no longer than 90 seconds. In this way, we chose the records with large proportion of bird sound time as far as possible to facilitate later processing. The records collected from Xeno-canto come from the different parts of the world, but we have not distinguished them in detail. In this paper, we try our best to obtain the sounds of birds in the mangrove areas under various circumstances, only to identify which kind of birds make sounds. We set up the mangrove bird sound dataset by assigning the same label to these slightly different records based on the species. Table 1 provides the detailed information on the birds sounds in the mangrove bird sound dataset.

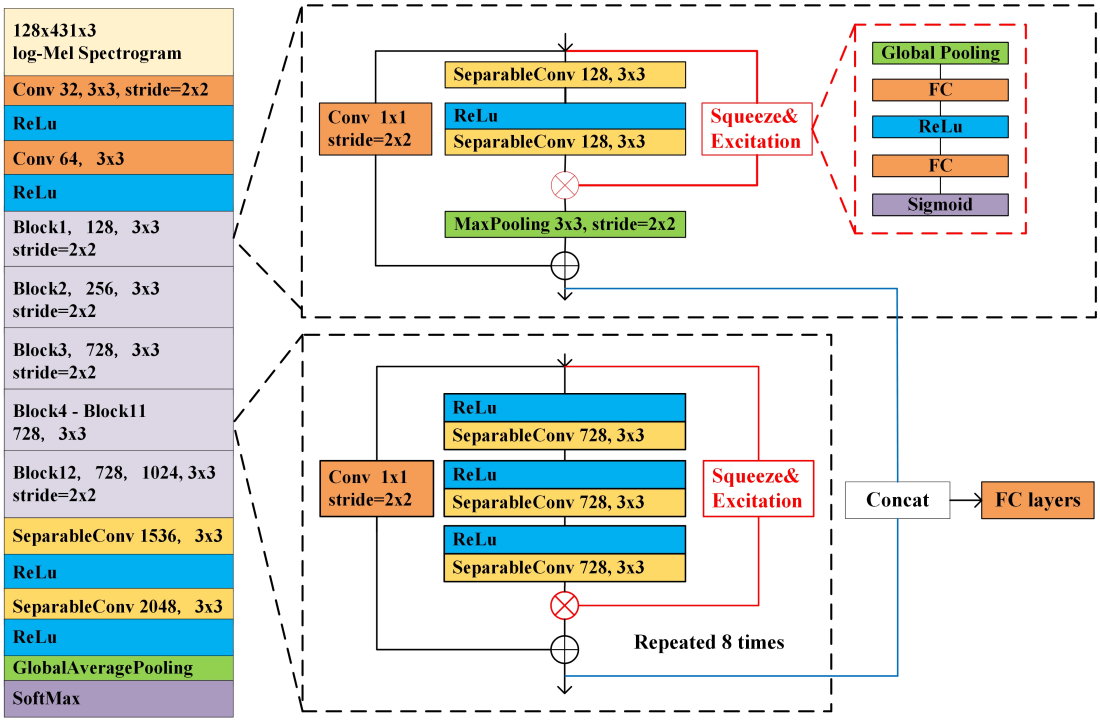

Figure 4: The overall architecture of the network. The left structure is a backbone network containing 12 Blocks, where Block4 to Block11 are the same Block, and the rest are similar Blocks. The dotted box on the upper right is the detail of Block1 (the red dotted box shows the detail of the SE block), and the dotted boxes at the lower right is the detail of Block4, the red connection part in the figure is the SE block introduced in this paper and the concatenated operation of recalibrated features is a connection marked as blue.

In the mangrove bird sound dataset, we preprocess the original records to audio segments 5 seconds and remove the audio segments only with the background noise from the original records to preserve the segments with the target bird sound. In the built sound dataset, there are 9282 recorded segments, of which the smallest one has 190 recorded clips and the largest one has 500 recorded clips. For each bird sound clip, we give a single label, regardless of the complex background sounds in the audio. We preprocess all the sounds data into a mono and 5-second clips with the sampling rate of 22050 Hz. And we collect these records from the different contributors on Xeno-canto, and the detailed information on each bird sound is given in the Table 1.

In addition, in order to further verify the effectiveness of our method, we also carry out experiments on the TAU Urban Acoustic Scenes 2019 dataset, which extends the TUT Urban Acoustic Scenes 2018 dataset with other 6 cities to a total of 12 large European cities. The dataset consists of 10-seconds audio segments from 10 acoustic scenes, and each acoustic scene has 1440 10-second segments (48 kHz / 24bit / stereo, 240 minutes of audio). The dataset is recorded in 12 large European cities. The dataset contains audio material from 10 cities, whereas the evaluation dataset contains data from all 12 cities. The dataset is perfectly balanced at acoustic scene level, with very slight differences in the number of segments from each city.

4.2. Data augmentation

In the classification tasks, the data augmentation often determines the performance of the classification results. As mentioned above, the data augmentation methods in the image processing field are not suitable for the mangrove bird sound classification task. So we need to find a simple and effective way to augment the data. Zhang et al. (2017) proposed the Mixup method which combines prior knowledge, that is, linear interpolation of eigenvectors has linear interpolation of related labels to extend the training distribution, to effectively alleviate the over-fitting problem. In general, the Mixup method makes no reference to the data type and enables data augmentation by generating virtual data.The linear interpolation of the samples can be demonstrated using Equations (1) and (2).

$$\bar{x} = \lambda x_i + (1 - \lambda) x_j \tag{1}$$

$$\bar{y} = \lambda y_i + (1 - \lambda) y_j \tag{2}$$

where $(x_i\ ,\ y_i)$, $(x_j\ ,\ y_j)$ are two samples randomly selected from the one-batch training data, $(\overline{x},\ \overline{y})$ is the generated virtual sample and $\lambda \in [0,1]$.

As shown in equations 1 and 2, two samples are randomly selected from the one-batch training data for random weighted summation, and the labels of the samples are also randomly weighted and summed. This method can increase the generalization ability of the model while reducing the computational cost. In the experiment, two samples randomly selected from the same batch of the training data are linearly interpolated to generate a new sample.

Table 1: Sound signal details of MBSD25

Order	Family	Species	Abb.	Recordings	Clips/5s
Ciconiiformes	Ardeidae	Ardea cinerea	ARCI	39	371
		Egretta alba	EGAL	90	307
		Nycticorax nycticorax	NYNY	90	330
Anseriformes	Anatidae	Anas crecca	ANCR	39	217
		Anas platyrhynchos	ANPL	39	371
		Anas querquedula	ANQU	45	190
Falconiformes	Falconidae	Falco tinnunculus	FATI	75	358
	Pandionidae	Pandion haliaetus	PAHA	47	340
Charadiiformes	charadriidae	Charadrius alexandrinus	CHAL	82	352
		Charadrius dubius	CHDU	61	416
	Scolopacidae	Tringa totanus	TRTO	82	468
		Calidris alpina	CAAL	72	340
	Recurvirostridae	Himantopus himantopus	HIHI	54	402
	Sternidae	Sterna hirundo	STHI	82	330
Passeriformes	Motacillidae	Motacilla alba	MOAL	64	500
		Motacilla flava	MOFL	63	392
	Pycnonotidae	Pycnonotus jocosus	PYJO	65	454
		Pycnonotus sinensis	PYSI	64	484
		Pycnonotus aurigaster	PYAU	47	317
	Dicruridae	Dicrurus macrocercus	DIMA	46	389
		Dicrurus hottentottus	DIHO	42	332
	Paridae	Parus major	PAMA	70	453
	Zosteropidae	Zosterops japonicus	ZOJA	42	452
Coraciiformes	Alcedinidae	Alcedo atthis	ALAT	101	395
Gruiformes	Gruidae	Grus grus	GRGR	68	322

4.3. Multi-Scale Feature Fusion

n the mangrove bird sound classification task, the collected audio data is clipped into the segments for an equal duration, generally according to the standard of 5s or 10s per segment. The clipped audio segment may contain only a few target birds sounds, so the features extracted from the clipped audio segment only contain a small amount of available information. In order to make better use of the feature information, we combine the recalibrated features from several different scales and input them into the classifier, so that we can make more effectively use of the different layer features with multi-scale information to compensate the disadvantage of the input audio data with less useful information.

In the Figure 4, we only give the corresponding schematic diagrams of the concatenated operations of Blcok1 and Block11 which marked as a blue connection. Specifically, we concatenate the outputs of Block1, Block 3 and Block 11, and the input of the last Maxpooling layer, and then input the fused vectors to the full connection layers. This simple and effec-

tive mechanism improves the classification results of the model by combining the features of the different scales, but it does not increase too much computational cost.

4.4. Experimental Settings

The preprocessing and extracting features of raw audio in this paper rely on LibROSA, which is a python package for music and audio analysis. And all the experiments in this paper were performed on two NVIDIA GTX 1080Ti GPUs in the PYTORCH environment of the Ubuntu 16.04 system.

Our network is trained for 200 epochs in batches of 16 samples by optimizing the categorical cross-entropy and Adabound Luo et al. (2019), and we apply 40 percent dropout to the full connection layers. The learning rate, mini-batch size, and decay were respectively set to 0.0001, 16, and 0.0001. The strategy of cosine annealing is used in training, which is one cycle for every 50 epochs with initial learning rate of 0.0001, and the learning rate decreases twice in one cycle. The scoring of the mangrove bird sound classification will be based on classification accuracy: the number of correctly classified segments among the total number of segments.

4.5. Experiment Results and Discussions

The results of the comparative experiments on the mangrove bird sound dataset are shown in Figure 5, Figure 6 and Figure 7, and the Figure 8 shows the results of comparative experiments on TAU Urban Acoustic Scenes 2019 dataset. The results show that our method can effectively improve the accuracy of the mangrove bird sound classification by setting up the different comparative experiments on whether to use SE block, multi-scale feature fusion and Mixup method.

The Figure 5 illustrates and compares the performances of different experimental conditions when SE block was introduced into Xception network. Although the classification performance is good when we only add the SE block to the backbone network, we can further improve the classification results by adding the mixup and multi-scale feature fusion methods. And the experimental results show that the performance of our method can be improved by using Adabound optimizer. From the results, we can see that after adding SE block (SE) and multi-scale features fusion operation (MS) to the network, the experimental results are much lower than only adding SE. We think that the operation of MS increases the parameters of the whole network, but it is not friendly to the simple bird sound dataset. So we further add Mixup method to the network, and the experimental results have been significantly improved, which confirms our analysis.

The Figure 6 shows the results under different experimental conditions. We compare the experimental results of different conditions with or without SE in Xception network, and conclude that adding the SE to the Xception network can effectively improve the experimental results and the best result was 0.8734 on the mangrove bird sound dataset. From the experimental results, the accuracy of adding SE to Xception is improved by more than 2 percent, and adding Mixup to network can improve by 1.7 percent, while adding these two methods together result in a higher improvement. In addition, as mentioned above, when only MS is added to Xception, the accuracy will be greatly reduced, while the accuracy of using MS and Mixup at the same time will be significantly improved.

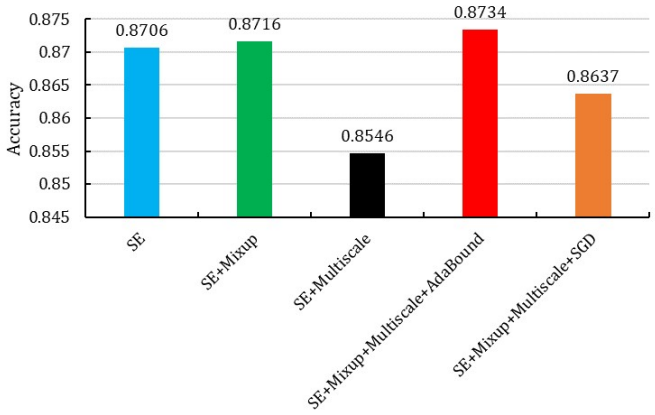

Figure 5: The experimental results under different conditions of embedding the SE block into the backbone network. "SE" means the SE block is added to the Xception network, "MS" represents the multi-scale features fusion operation, and "+" indicates that the method is used in Xception network.

Further, we compare the experimental results of different optimizers. It can be seen that the Adabound optimizer can make the model get better results, but the convergence time is almost the same as SDG.

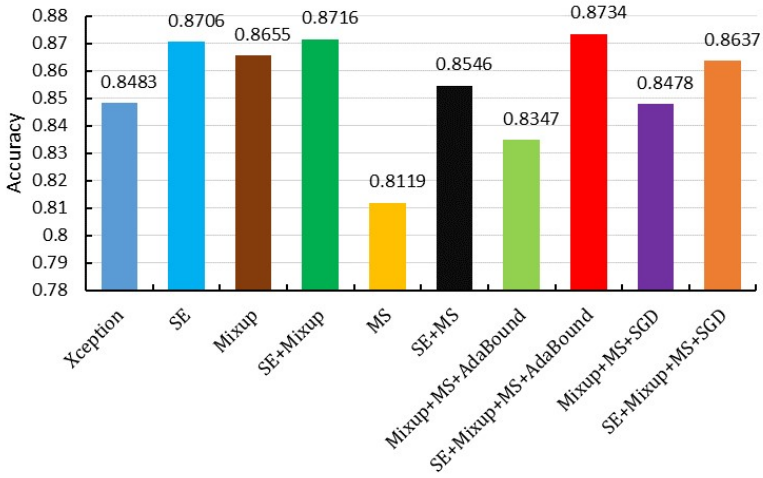

Figure 6: The experimental results under different conditions of embedded and non-embedded SE block in backbone network. The backbone model without "SE" representation in the figure is Xception.

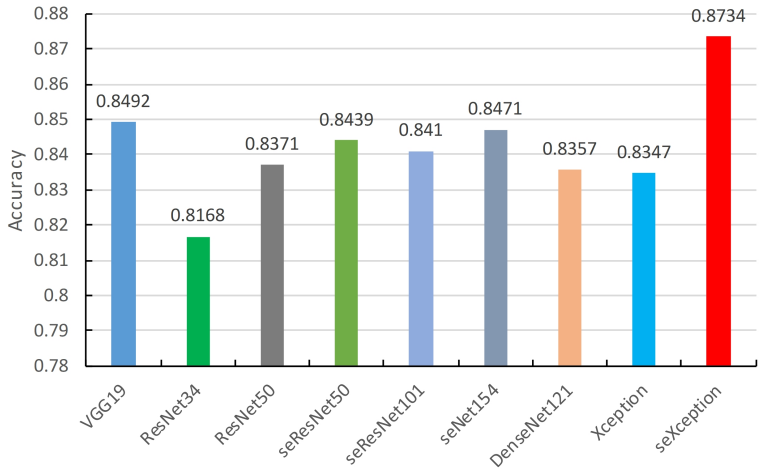

Figure 7: The comparative experimental results of mainstream classification networks on the mangrove bird sound dataset.

The Figure 7 shows the comparative experimental results with the best classification networks currently available. The Mixup and MS methods are introduced into each network. It can be seen from the figure that although the results of networks such as VGG19, seNet54, DenseNet121 are close to Xception, the parameters of these networks are very large and the training time is very expensive. Most importantly, the accuracy of our method is much higher than that of other mainstream classification models. In the training phase, the size of each mini-batch is set to 16, and the training time of each epoch of seXception is about 2 minutes.

The Figure 8 shows our experimental results on the TAU Urban Acoustic Scenes development 2019 dataset. We have done different comparative experiments, and the best results are obtained by our method, which is 20 percent higher than the official baseline. The TAU Urban Acoustic Scenes 2019 dataset is the official public dataset of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events DCASE 2019 Task 1A,which is concerned with basic problem of acoustic scene classification, in which all data (development and evaluation) are recorded with the same device, and contains only data from the 10 known acoustic scene classes. We use this public dataset to verify the applicability of our proposed method for the acoustic scene classification tasks. If our method can classify the acoustic scene well, then we can further label the existing bird sound dataset in our follow-up work, and classify the scenes of birds based on the bird sounds.

In general, our proposed method can get good classification results on the mangrove bird sound dataset we have built. At the same time, compared with other mainstream sound classification models based on deep neural network, it can be seen that our method can get the best classification results with less time overhead. Furthermore, the proposed method can also classify the acoustic scenes well.

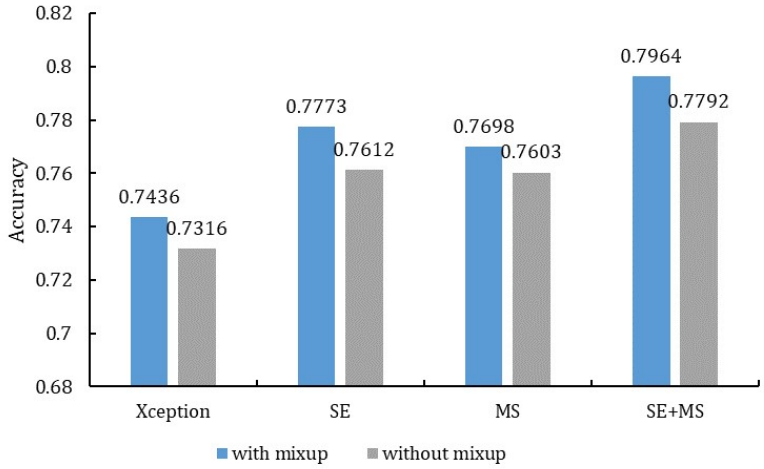

Figure 8: The results of different network structures or methods on the TAU Urban Acoustic Scenes development 2019 dataset.

5. Connclusion and Future Work

In this paper, we built a dataset containing 25 kinds of the mangrove bird sound, called mangrove bird sound dataset, and we proposed a method that combines feature recalibration mechanism with depthwise separable convolution for the mangrove bird sound classification. In our method, we use the mixup and multi-scale feature fusion tricks to get a better performance on the mangrove bird sound dataset. The experimental results demonstrate that the proposed method not only performs well on the mangrove bird sound dataset but also gets a good result on the TAU Urban Acoustic Scenes 2019 dataset. So we can conclude that our method has good adaptability in the acoustic scenes classification tasks.

In future work, we will try our best to collect a wide range of the different mangrove bird sound and intend to continue the works of bioacoustics classification and detection, such as bird sound detection and classification, and expand the mangrove bird sound classification task to the classification of bird singing scenes which is the reason why we verify the adaptability of our method on the TAU Urban Acoustic Scenes 2019 dataset.

6. Acknowledgment

This work was supported by the Key projects of Science and Technology Agency of Guangxi province, china (Guike AA 17129002).

References

Deep Chakraborty, Paawan Mukker, Padmanabhan Rajan, and Aroor Dinesh Dileep. Bird call identification using dynamic kernel based support vector machines and deep neu-

ral networks. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 280–285. IEEE, 2016.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

Chih-Hsun Chou, Chang-Hsing Lee, and Hui-Wen Ni. Bird species recognition by comparing the hmms of the syllables. In *Second International Conference on Innovative Computing, Informatio and Control (ICICIC 2007)*, pages 143–143. IEEE, 2007.

Seppo Fagerlund. Bird species recognition using support vector machines. *Eurasip Journal on Advances in Signal Processing*, 2007(1):038637, 2007.

Todor D Ganchev, Olaf Jahn, Marinez Isaac Marques, Josiel Maimone de Figueiredo, and Karl-L Schuchmann. Automated acoustic detection of vanellus chilensis lampronotus. *Expert systems with applications*, 42(15-16):6098–6111, 2015.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

Jorn-Henrik Jacobsen, Jan van Gemert, Zhongyu Lou, and Arnold WM Smeulders. Structured receptive fields in cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2610–2619, 2016.

Peter Jančovič, Münevver Köküer, and Martin Russell. Bird species recognition from field recordings using hmm-based modelling of frequency tracks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8252–8256. IEEE, 2014.

Alexis Joly, Hervé Goëau, Christophe Botella, Hervé Glotin, Pierre Bonnet, Willem-Pier Vellinga, Robert Planqué, and Henning Müller. Overview of lifeclef 2018: a large-scale evaluation of species identification and recommendation algorithms in the era of ai. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 247–266. Springer, 2018.

Kantapon Kaewtip, Abeer Alwan, Colm O'Reilly, and Charles E Taylor. A robust automatic birdsong phrase classification: A template-based approach. *The Journal of the Acoustical Society of America*, 140(5):3691–3701, 2016.

Steve Kelling, Jeff Gerbracht, Daniel Fink, Carl Lagoze, Weng-Keen Wong, Jun Yu, Theodoros Damoulas, and Carla Gomes. ebird: A human/computer learning network for biodiversity conservation and research. In *Twenty-Fourth IAAI Conference*, 2012.

Ivan Kiskin, Davide Zilli, Yunpeng Li, Marianne Sinka, Kathy Willis, and Stephen Roberts. Bioacoustic detection with wavelet-conditioned convolutional neural networks. *Neural Computing and Applications*, pages 1–13, 2018.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Chiman Kwan, KC Ho, Gang Mei, Yunhong Li, Zhubing Ren, Roger Xu, Y Zhang, Debang Lao, M Stevenson, Vincent Stanford, et al. An automated acoustic system to monitor and classify birds. *EURASIP Journal on Advances in Signal Processing*, 2006(1):096706, 2006.

Liangchen Luo, Yuanhao Xiong, Yan Liu, and Xu Sun. Adaptive gradient methods with dynamic bound of learning rate. *arXiv preprint arXiv:1902.09843*, 2019.

Mark D. Plumbley, Christian Kroos, Juan P. Bello, Ga?l Richard, Daniel P.W. Ellis, and Annamaria Mesaros. *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*. Tampere University of Technology. Laboratory of Signal Processing, 2018.

Ilyas Potamitis, Stavros Ntalampiras, Olaf Jahn, and Klaus Riede. Automatic bird sound detection in long real-field recordings: Applications and tools. *Applied Acoustics*, 80:1–9, 2014.

José Francisco Ruiz-Muñoz, Zeyu You, Raviv Raich, and Xiaoli Z Fern. Dictionary learning for bioacoustics monitoring with applications to species classification. *Journal of Signal Processing Systems*, 90(2):233–247, 2018.

Justin Salamon and Juan Pablo Bello. Unsupervised feature learning for urban sound classification. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 171–175. IEEE, 2015.

Laurent Sifre and Stéphane Mallat. Rigid-motion scattering for image classification. *PhD thesis, Ph. D. thesis*, 1:3, 2014.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

Bálint Pál Tóth and Bálint Czeba. Convolutional neural networks for large-scale bird song classification in noisy environment. In *CLEF (Working Notes)*, pages 560–568, 2016.

Huy Dat Tran and Haizhou Li. Sound event recognition with probabilistic distance svms. *IEEE transactions on audio, speech, and language processing*, 19(6):1556–1568, 2010.

Jiang-jian Xie, Chang-qing Ding, Wen-bin Li, and Cheng-hao Cai. Audio-only bird species automated identification method with limited training data based on multi-channel deep convolutional neural networks. *arXiv preprint arXiv:1803.01107*, 2018.

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