IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 22 July 2022, accepted 5 August 2022, date of publication 11 August 2022, date of current version 18 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3198104

RESEARCH ARTICLE

Design of Bird Sound Recognition Model Based on Lightweight

FAN YANG, YING JIANG, AND YUE XU
College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing 210037, China
Corresponding author: Ying Jiang (yj15562022@163.com)

ABSTRACT Bird sounds recognition is of great significance in bird protection. With appropriate sound classification, research can automatically predict the quality of life in the area. Nowadays, the deep learning model is used to classify bird sound data with high classification accuracy. However, the generalization ability of most existing bird sound recognition models is poor, and the complicated algorithm is applied to extract bird sound features. To address these problems, a large data set containing 264 kinds of birds is constructed in this paper to enhance the generalization ability of the model, and then a lightweight bird sound recognition model is proposed to build a lightweight feature extraction and recognition network with MobileNetV3 as the backbone. By adjusting the depthwise separable convolution in the model, the recognition ability of the model is improved. A multi-scale feature fusion structure is designed, and the Pyramid Split Attention (PSA) module is added to the multi-scale feature fusion structure to improve the adaptability of the network to scale extraction of spatial information and channel information. To improve the refinement ability of the model towards the global information, the channel attention mechanism and ordinary convolution are introduced into Bneck module which makes the Bneck module become the Bnecks module. The experimental results show that the accuracy of Top-1 and Top-5 of the model in identifying 264 kinds of birds on the self-built data set is 95.12% and 100%, which are higher than that of MobileNetV1, MobileNetV2, MobileNetV3 respectively. Although the accuracy is lower than ResNet50, the number of parameters and floating-point operations (FLOPs) of the model is only 2.6M and 127M respectively. The accuracy is only reduced by 2.25% while saving costs.

INDEX TERMS Attention mechanism, bird sound recognition, deep learning, lightweight, multi-scale feature fusion.

I. INTRODUCTION

More than 10,000 species of birds are found in almost every environment, from unspoiled rainforests to suburbs and even cities [1], [2]. Nowadays bird species all over the world are extinct to varying degrees. For example, Hawaii, as the extinction capital of the world, has lost 68% of bird species, which may destroy the entire food chain and thus the ecological environment of Hawaii. Using population monitoring, researchers can understand how local birds respond to changes in the environment and conservation efforts. Being able to monitor bird movements in real-time is the first step in this work [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang.

At present, many professionals begin to observe birds for a long time to conserve their species [4]. However, most of the monitoring tasks are manual by professionals. As birds fly fast and are difficult to observe, and when they live on land, they are easily frightened by human activities and cannot be recorded by the camera quickly. Therefore, using image recognition to recognize birds in real-time is both difficult and expensive [5]. What's more, many birds are isolated in inaccessible high-altitude habitats. Due to the difficulties in physical monitoring, more and more professionals generally recognize the bird species by hearing [6] and recording. This method, called bioacoustics monitoring, can provide a passive and cost-effective strategy for the study of endangered bird populations. Nevertheless, if a manual surveillance program is performed, this monitoring process is

time-consuming and laborious, and real-time monitoring of birds in areas such as ecological protection zones can't be carried out.

Most people in related fields tend to use Internet of things devices to remotely online monitor bird populations. Since most of the bird protection habitats are in the wild, it is difficult for the online monitoring system to transmit the sound of birds back to the server for data processing, recognition and feedback under good network conditions. If the off-line monitoring is carried out in the bird reserve, the low-cost embedded equipment cannot carry the high complexity sound feature extraction algorithm and high-precision sound recognition algorithm. Therefore, aiming at this point, this paper wants to design a lightweight bird voice recognition algorithm, which can not only achieve high accuracy by using simple and single features, but also make the model small enough to run in low-cost embedded devices.

A. PRIOR WORK

There is a lot of work for bird sound recognition. In the traditional field of machine learning, Ramirez *et al.* [7] used Mel frequency cepstral coefficients (MFCC) and inverted Mel frequency cepstral coefficients (IMFCC) as sound features to recognize the sound of birds and found that IMFCC achieved better recognition accuracy. Lucio *et al.* [8] adopted the method of multi-feature fusion and fused the sound features with three texture feature operators: local binary, Gabor filtering, and local phase quantization. Finally, the support vector machine was used to obtain 77.65% accuracy in 46 kinds of birds. Salamon *et al.* [9] generated a feature dictionary from logarithmic scale Mel spectrum species and achieved 93.96% accuracy in 43 species of birds using the support vector machine (SVM). Pahuja *et al.* [10] generated a statistically evaluated short-term Fourier trans-form spectrogram-based feature matrix as characterization of vocalization patterns of bird species, and attain enhanced recognition accuracy (96.1%) using a multi-layer perceptron artificial neural network. In the above machine learning model, classifier algorithms are often relatively simple and easy to implement, but in order to improve the accuracy of classifiers, most experts and scholars will use complex feature fusion extraction algorithms. Although these feature extraction algorithms do effectively improve the classi-fication accuracy, due to their high complexity, the cost of implementation is often high.

In recent years, deep convolution neural network has made great progress in sound recognition and other aspects [11], [12], [13]. Zhang *et al.* [14] used short time Fourier transform (STFT) and other methods to convert birds sound into the spectrum and used convolutional neural network to classify bird sounds. Different from using a simple convolutional neural network, Sankupellay *et al.* [15] used 50 layers residual neural network (Resnet50) to classify the time spectrum of bird sounds. Huang *et al.* [16] used densely connected networks (Densenet) to extract time spectrum features and classify them, which improved the classification effect. To further improve the recognition accuracy. Sheng *et al.* [17] used 1-dimensional CNN-LSTM, 2-dimensional vgg-style, and 3-dimensional densenet121 model as feature extractors to extract advanced features, and then used a shallow classifier to recognize 43 kinds of bird sounds, achieving a balanced accuracy of 93.89%. The methodology [18] deviates from the existing approaches by integrating transfer learning. Using such as ResNet50, DenseNet201, InceptionV3, Xception, and EfficientNet can effectively extract and recognize the audio signals from different bird species with significant prediction accuracy. In the above deep learning model, the complex feature ex-traction algorithm is replaced by various deeper and high-precision models with many parameters, but this also faces the same problem. A large number of parameters will reduce the computing speed of the device, and complex model pairs cannot be applied to low-cost CPU. It is still unrealistic to run the models in low-cost embedded devices.

In addition, although most of the studies on bird sound recognition have achieved high recognition accuracy, the amount of data set used in the research is small [17], [18], [19], [20], [21], [34], [35], [36], [37]. Most studies are limited to identifying a single bird species, and the number of bird species in the data set used is only 20 to 30 (in the following, this paper will list some comparative data), so the proposed model does not have generalization ability.

Therefore, in order to apply the recognition model to low-cost embedded devices to realize offline real-time bird popu-lation monitoring, it is necessary to improve the generaliza-tion ability of the model, reduce the complexity of feature extraction algorithm and design a lightweight model.

B. CONTRIBUTION

In order that overcoming the above shortcomings, this paper first collects a large number of bird sound data and constructs a data set of 264 kinds of birds. Then, a single Mel spectrum is used as the sound data feature. Finally, a lightweight recognition model is designed to recognize the bird sound feature map, and the classification result is obtained. The contributions of the paper can be summarized as follows:

1) *Built a huge bird data set*: In this paper, a large data set containing 264 species of birds is constructed, which can effectively improve the generalization ability of the model;
2) *Lightweight bird recognition model based on improved MobileNet design*: This paper designs a light-weight bird sound recognition model to improve the accuracy of bird sound recognition. The multi-scale feature fusion structure is proposed, and then a PSA (pyramid split attention) module is added to the multi-scale feature fusion structure to enhance the adapt-ability of the network to scale extraction of spatial information and channel information. The Bnecks block is designed, and the channel attention mechanism and ordinary convolution are introduced to improve the refinement ability of the model to the global information;

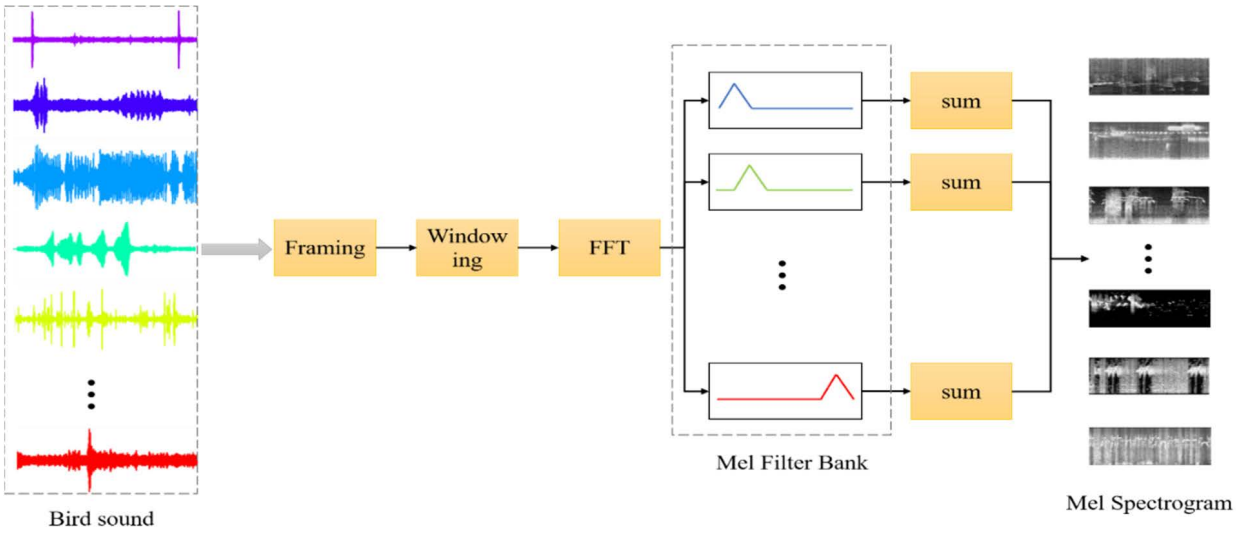

FIGURE 1. Bird audio signal feature extraction process.

3) *Simple bird sound feature extraction process*: By extracting the Mel spectrogram and stacking it as a three-dimensional feature into the recognition model, a better recognition result can be obtained.

The organization of the remainder of this study is as follows:

In Section 2, the related work is shown. Then section 3 explains how to construct the bird sound recognition model. In Section 4, the ablation experimental results, the comparison of results between different models and the comparison of the result between the scheme proposed in this paper and the previous are given. Finally, Section 5 concludes the research.

II. RELATED WORK

A. DATA SET CONSTRUCTION

The bird sound data used in this paper comes from various bird recognition competitions of Kaggle [23], [24], [25] and some bird sounds in rural areas of Baguazhou Qixia, District, Nanjing City, Jiangsu Province, China. The collected bird sound data are sorted and labeled respectively. There are 264 bird categories. Table 1 shows bird sounds in the data set and the number of audio clips contained in this paper. Due to the large amount of data, we only listed a s bird sound data information.

TABLE 1. Data information for each category of the dataset.

Category Number	Category Name	Data Volume	Duration(s)
0	aldfly	102	15
1	ameavo	87	37
2	bkpwar	100	20
3	blujay	100	20
4	brebla	100	20
5	eargre	68	20
6	fiespa	132	20

B. DATA PREPROCESSING

The data source, data format and the sampling rate of the different bird sound data in the data set constructed in this paper are different, so before extracting features of the bird sound, corresponding pre-treatment is needed to eliminate the differences of input data in data source, data format, and sampling rate. In addition, the duration of each bird sound segment in the dataset constructed in this paper is different, but overall, the duration of each sample data is more than 10 seconds, therefore this paper intercepts the sample data at 5 seconds interval so that the duration of each sample data is the same. To eliminate the effect of the amplitude difference in bird audio data on model training, this paper standardizes min-max for each intercepted bird sample data as follows:

$$S(n)_t = \frac{S(n)_t - \min\{s(n)\}}{\max\{s(n)\} - \min\{s(n)\}} \tag{1}$$

$S(n)_t$ denotes the input signal after normalization at t-time, $s(n)_t$ presents the original input signal at x-time, $min\{\cdot\}$, $max\{\cdot\}$ are the minimum and maximum values respectively. In order to verify the influence of standardized data on the experimental results, this paper will prove it in the ablation experiment in Section 4.

C. FEATURE EXTRACTION

Different from human voice recognition, bird sound recognition in this paper focuses more on the characteristics of bird sound than the content of bird sound. In order to simplify the complexity of the feature fusion algorithm and reduce the computational load of the model, the Mel spectrum, which is widely used in speech recognition systems, is selected as the feature of the bird audio signal. The process of extracting the

feature is shown in Figure 1. The Mel spectrum of the bird audio signal obtained in this paper is defined as follows:

$$feature(m) = \sum_{k=0}^{N-1} E(k)H_m(k) \tag{2}$$

Here *feature(m)* is the corresponding energy characteristic of the Mth Mel filter, $E(k)$ is the signal energy spectrum, $H_m(k)$ is the response of the Meier filter, and N is the length of the FFT. The feature is fused on the channel dimension to get a 3-D feature map. Furthermore, the difference between the standardized data and the original data are compared by calculating the feature extraction time, as shown in Figure 2. The result shows that under the same machine, standardization can speed up the speed of feature extraction. Figure 3 appear that the standardized data is more distinctive while non-standardized data is a noisy, featureless signal.

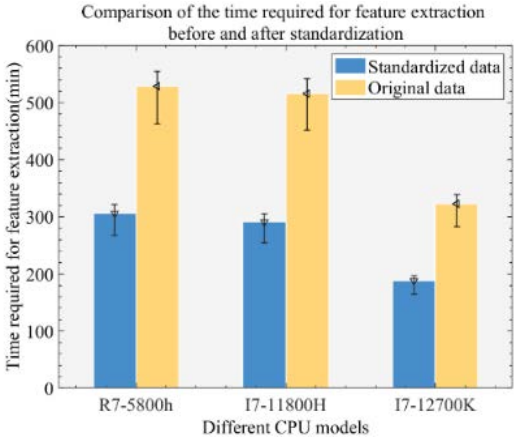

FIGURE 2. Comparison of time required for feature extraction between standardized and original data by different types of CPU.

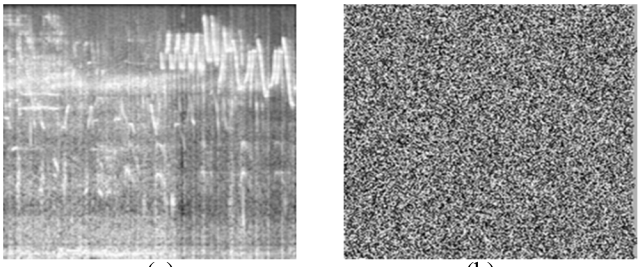

FIGURE 3. Comparison of the feature map of standardized and with that of original data. (a) Feature map of standardized data; (b) Feature map of original data.

III. MODULE CONSTRUCTION

For purpose of making the deep learning model can be rapidly deployed and run on the mobile terminal, Howard *et al.* [26] proposed the depthwise separable convolution (DSC) for mobile devices. Compared with the traditionnal convolution neural network, DSC can improve the training speed of the model, reduce the parameters, calculation of the model and also can infer at a faster speed at the moving terminal. The DSC consists of a depthwise(DW) convolution and a point-wise(PW) convolution, in which the DW convolution works as shown in Figure 4(a) and the PW convolution as shown in Figure 4(b).

DW convolution performs convolution operations on the input images in their respective channels, and the output feature map has the same number of channels as the input images. It can effectively obtain the channel information of the input image, but cannot use the feature information of different channels at the same position. To address this point, PW convolution is required to spatially combine the feature maps output by the DW convolution, expand the output channel, and extract spatial information. The combination of DW convolution and PW convolution results in a DSC that takes only one-T of the traditional convolution, as follows:

$$\frac{1}{T} = \frac{1}{N} + \frac{1}{D_K^2} \times 100\% \tag{3}$$

where N is the number of output channels of the convolution operation, and D_K is the size of the input image (it is assumed that the size of the input image is $D_K \times D_K$).

Although DSC can reduce the number of parameters and computations, the sequential combination of the DW convolution and the PW convolution limits its feature extraction capabilities. Due to the initial module of the feature data is always transmitted in low-dimensional form, and DW convolution cannot expand the output channel. This will result in the loss of the original features. Not only that, the ReLU activation function is usually used after DW convolution to introduce nonlinearity and speed up training.

For traditional images, because an image has rich features, these disadvantages can be overcome by relying on rich features. However, for bird sound spectrogram features, low-dimensional data will lose a large number of features after passing through the activation function ReLU, resulting in the collapse of low-dimensional data. Therefore, if the PW convolution is performed first and the DW convolution is followed, the low-dimensional feature data can be converted into high-dimensional data by PW convolution, so that a large amount of spatial information will be stored in the feature map, and then the feature information of each channel can be extracted by DW convolution using the high-dimensional features after the PW convolution. Through the above adjustment, the bird sound recognition model proposed in this paper can speed up the inferring time and improve the accuracy at the same time.

In order to extract features in low dimensions to the greatest extent, this paper redesigns the activation function used by the DSC, and adopts the Mish function with a smoother gradient, which is defined as follows:

$$f(x) = x \tanh(\ln(1 + e^x)) \tag{4}$$

Through the above methods, the improved DSC can enhance the extraction of low-dimensional features without

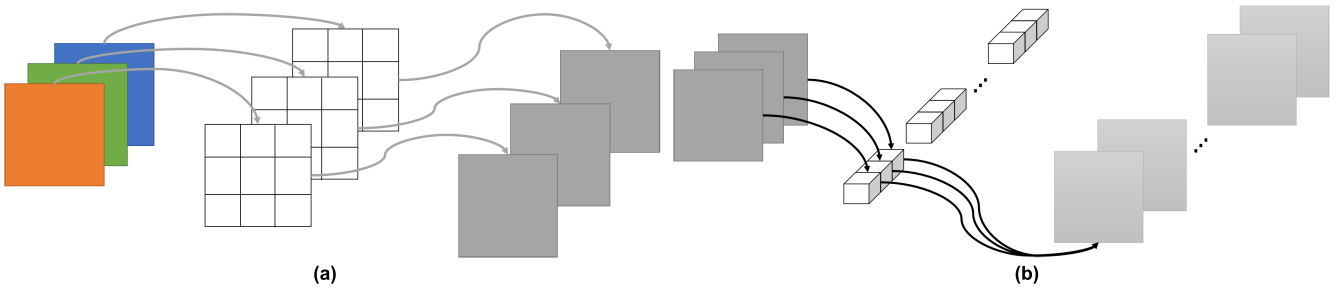

FIGURE 4. Process of depthwise separable convolution. (a) The process of depthwise convolution; (b) The process of pointwise convolution.

introducing too many parameters and calculations, so as to speed up the training and inferring time of the model.

A. MODEL OVERALL DESIGN ARCHITECTURE

The backbone part of the lightweight model designed in this paper refers to MobileNetV3-Small [27]. This paper has adjusted and improved the problems existing in MobileNetV3-Small and the situation of the actual data from data set in this paper. The overall model architecture is shown in Figure 5 and Table 2. The lightweight model consists of one Inception block, two Bnecks blocks and 17 Bneck blocks stacked together. Data is extracted in the form of high dimension in the backbone and transferred between blocks in the form of low dimension.

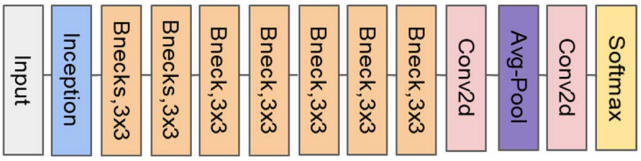

FIGURE 5. Lightweight model network architecture.

TABLE 2. Overall architecture of The lightweight model (c is the output channel, n represents the quantity, s represents the number of stride, SE represents whether to use Squeeze-and-Excitation module and activate represents the activation function used).

Input	Operator	c	n	s	SE	activate
$216^2 \times 3$	*Inception*	64	1	1	—	*ReLU*
$108^2 \times 64$	*Bnecks*	32	1	1	*SE*	*ReLU*
$108^2 \times 32$	*Bnecks*	32	1	1	*SE*	*ReLU*
$108^2 \times 32$	*Bneck*	16	1	1	—	*Mish*
$108^2 \times 16$	*Bneck*	24	2	2	—	*Mish*
$54^2 \times 24$	*Bneck*	32	3	2	—	*H-swish*
$27^2 \times 32$	*Bneck*	64	4	1	—	*H-swish*
$27^2 \times 64$	*Bneck*	96	3	2	*SE*	*H-swish*
$14^2 \times 96$	*Bneck*	160	3	2	*SE*	*H-swish*
$7^2 \times 160$	*Bneck*	320	1	1	—	*H-swish*
$7^2 \times 320$	*Con2d 1x1*	1280	1	1	—	—
$7^2 \times 1280$	*Avgpool 7x7*	—	1	—	—	—
$1^2 \times 1280$	*Conv2d*	264	1	1	—	—

B. MULTI-SCALE FEATURE FUSION STRUCTURE DESIGN

In order to enhance the feature extraction of sound data. This paper is inspired by the fact that neurons can process and collect muti-scale spatial information at the same state due to the different sizes of receptive fields when stimulating the human brain. While avoiding the introduction of too many parameters and computations [28], so it is only improved in the initial module of the network architecture, and the improved multi-scale feature fusion structure in this paper - Inception block [29] is added. The improved Inception block architecture is shown in Figure 6.

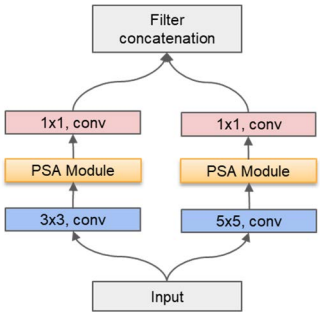

FIGURE 6. Multi-scale feature fusion structure.

In the initial stage of the model, the features of input data are rich, so it is necessary to design a multi-scale feature fusion structure to fully extract the features of the original data. In this paper, two parallel branches are used for data feature extraction. The two parallel branches are 3×3 and 5×5 multiscale feature extraction. After the multi-scale feature extraction of each branch, the PSA (Pyramid Split Attention, PSA) module [30] is introduced, which can fully capture the spatial information of different scales to enrich the feature space, establish a long-distance spatial attention dependence mechanism and extract channel features of different scales, and the model architecture is shown in Figure 7. The 3×3 convolution is used to extract the subtle features of the original sound data, and the 5×5 convolution is used to extract the overall characteristics of the original sound data. Considering the amount of computation and the introduction of the PSA module, it does not use larger and more convolution kernels for the initial feature extraction operation.

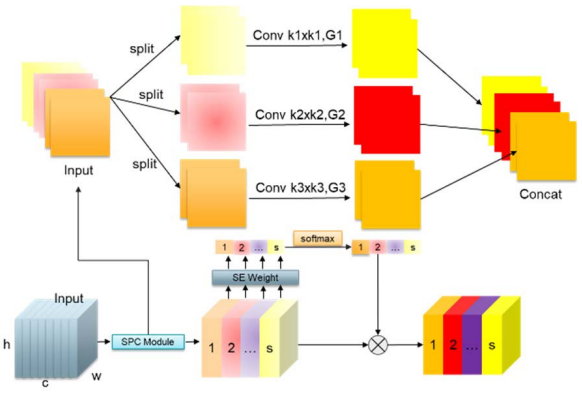

FIGURE 7. PSA module architecture.

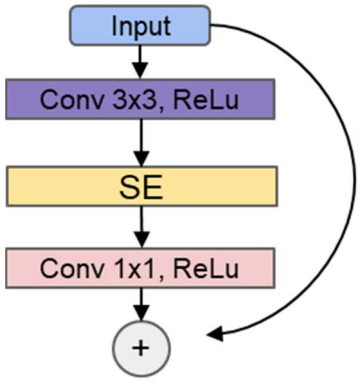

FIGURE 9. Structure of Bnecks where SE modules are used.

C. NETWORK BACKBONE DESIGN

In order to reduce the number of parameters and computations, this paper reduces the number of backbone layers of MobileNetV3, and the kernel size of the depthwise convolution is 3×3. Referring to the reverse residual structure proposed by MobileNetV2 [31]. On this basis, this paper proposes two block structures—Bneck and Bnecks block. The Bneck block structure is shown in Figure 8, which draws on the residual connection idea of ResNet [32], and designs the structure of the reverse residual. The 1×1 convolution is used

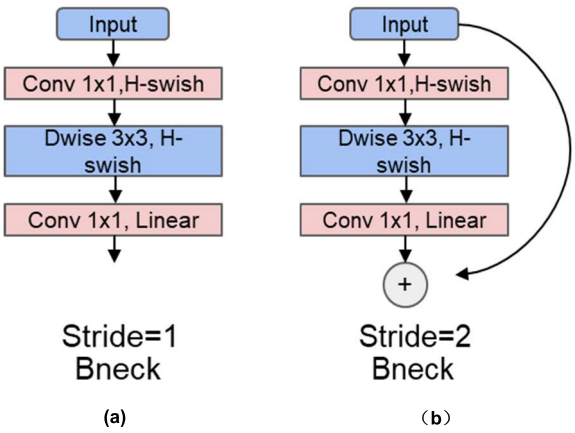

FIGURE 8. Two structures of Bneck. (a) The structure with the stride of 1; (b) The structure with the stride of 2.

The Bnecks block adds the channel attention mechanism on the basis of the Bneck block, and at the same time replaces the DSC in the Bneck block with ordinary convolution and introduces a residual structure, as shown in Figure 9. In order to avoid ordinary convolution causing a surge of computations and parameters, this paper has tried many experiments and found that only adding a small number of Bnecks blocks after the initial Inception block can improve the effectiveness of the model, and the computation and parameters of the model will not be significantly improved.

When the data is sent into the model, the input end of the model has the most abundant data. A large amount of thinning data exists at the bottom of the model. If the global information cannot be extracted at the input end, the classification accuracy of the model cannot be improved. Referring to the design idea of EfficientNet [28] and the attention mechanism [39], the Bnecks module is added after the multi-scale feature extraction module of the model. The attention mechanism can enhance the extraction of different channel information in data, while common convolution integrates the channel weights learned by the attention mechanism to extract global information emphatically and effectively. So that in the Bnecks block, the irrelevant information in the global information will be removed and the effective information will be retained to the maximum extent, so that the model can refine the global information to the maximum extent, thus improving the refinement ability of the model to the global information.

IV. RESULT

A. EXPERIMENTAL ENVIRONMENT

The feature extraction of the data is completed in the environment based on python3.9, the model recognition and classification part are completed in the environment based on python3.9 and pytoch1.8, the hardware configuration is 5GHz Intel i7 12700K processor, 32GB 3200Mhz DDR4 memory, Nvidia GeForce RTX3070 and Nvidia GeForce RTX3070Ti graphics cards. The total number of birds Mel spectrogram samples after feature extraction is 229164, 183690 samples are selected as the training set and 45924 samples are used as the test set. In the experiment, the learning rate is set to 0.025, and the batch size is which is set is 32, the epoch is set to 300, the model optimizer is Stochastic Gradient Descent (SGD, Stochastic Gradient Descent), the loss function uses the cross-entropy loss function, and the learning rate descent strategy uses Cosine Annealing [33].

B. ALGORITHM COMPARISON AND ANALYSIS

In order to verify that each improvement point of the proposed model contributes to the improvement of model performance,

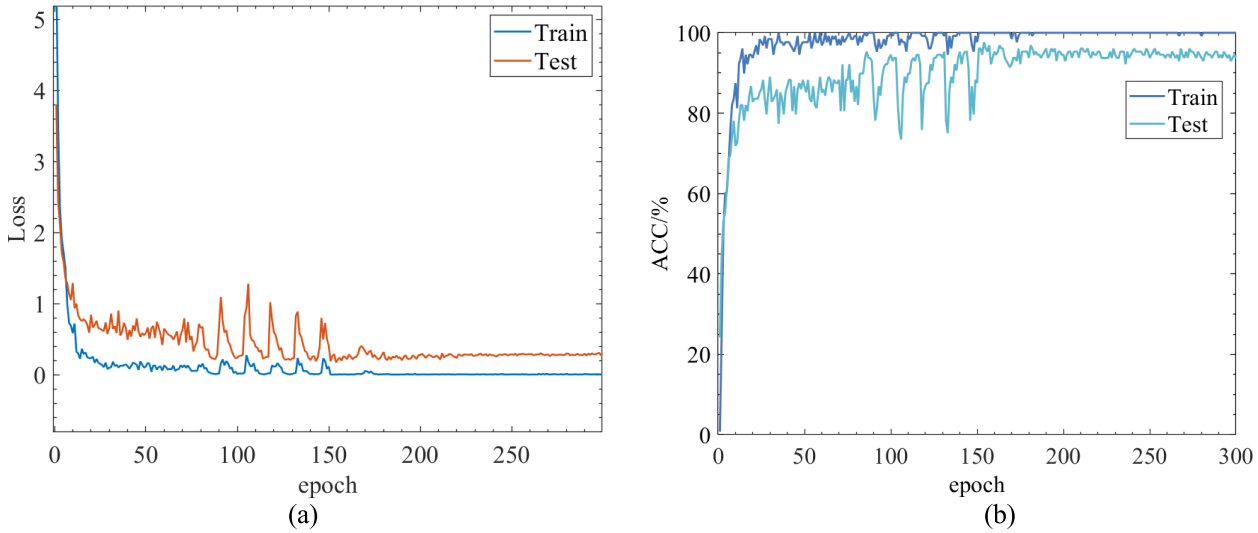

FIGURE 10. Schematic diagram of the training results of the model proposed in this paper. (a) The loss curve in the training set and test set; (b) The accuracy curve of the training set and test set.

this paper has conducted a series of ablation experiments. In the ablation experiments, the TOP-1 accuracy of the recognition model on the test set is used as the benchmark. Ablation experiments include whether to use multi-scale feature fusion module, whether to use Bnecks module with attention mechanism, whether the depth separable convolution is adjusted as described in Section 3 and whether to carry out standardization. The results of the ablation experiments are shown in Table 3.

TABLE 3. Comparison of ablation experimental results (1 represents whether multi-scale feature fusion is carried out, 2 represents whether there is Bnecks module, 3 represents whether the DSC can be adjusted for separable convolution, and 4 indicates whether standardization is carried out).

Compared	1	2	3	4	Acc/%	Weights/ M	FLOPS/ M
1	—	✓	✓	✓	*91.22%*	*2.4M*	*100M*
2	✓	—	✓	✓	*94.78%*	*2.3M*	*113M*
3	✓	✓	—	✓	*84.62%*	*2.0M*	*96M*
4	✓	✓	✓	—	*45.64%*	*2.6M*	*127M*
This paper	✓	✓	✓	✓	*95.12%*	*2.6M*	*127M*

C. ALGORITHM COMPARISON AND ANALYSIS

At present, there are a large number of types of deep learning models proposed at home and abroad. In order to show the effectiveness of the model in this paper, the current classic deep learning models such as ResNet, DenseNet, VGG, etc. and lightweight deep learning models MobileNet, ShuffleNet, EfficientNet and other models are selected respectively. Using the above models to train the data set built in this paper, record the test set accuracy and training loss of different models, and compare with the model proposed in this paper.

Figure 10 shows the performance of the training set and test set of the model in this paper. The convergence speed of the model is fast. The model converges when it approaches 200. The accuracy of the model is high, which can reach 100% in the training set and 95.12% in the test set. The model has good learning ability. The loss of the model on the training set is close to 0, and the loss on the test set is about 0.2. Therefore, the overall performance of the model proposed in this paper is better.

As shown in Figure 11, it is the training result curve of each model on the bird audio feature map data set, where epoch is the iteration period of training, ACC is the accuracy of the test set, and Loss is the training loss. As can be seen from Figure11 (a), the training loss of the model proposed in this paper decreases more quickly than other previous models, and the most convergent value is close to 0, indicating that the model has a fast-learning ability and can learn the key characteristics of bird sound data more quickly. At the same time, it can be concluded from Figure 11 (b), that the model presented in this paper also has a good classification accuracy. Although the model proposed in this paper adopts a lightweight architecture, it still achieves good results, the training effect is close to ResNet50, it converges faster than ResNet50 in the training process, and the accuracy rate is better than that of MobileNet and ShuffleNet.

In this paper, the statistical results of different models are tabulated, as shown in Table 4.

Table 4 shows the classification effect of different models on bird sound data. The model proposed in this paper is improved based on MobileNet V3, the accuracy rate of the model is 2.94% higher than that of MobileNet V3, and the amount of network parameters is not significantly improved compared to MobileNetV3. The main reasons are as follows:

1. MobileNetV3, as the latest lightweight model, has a strong recognition ability itself, and the reverse residual

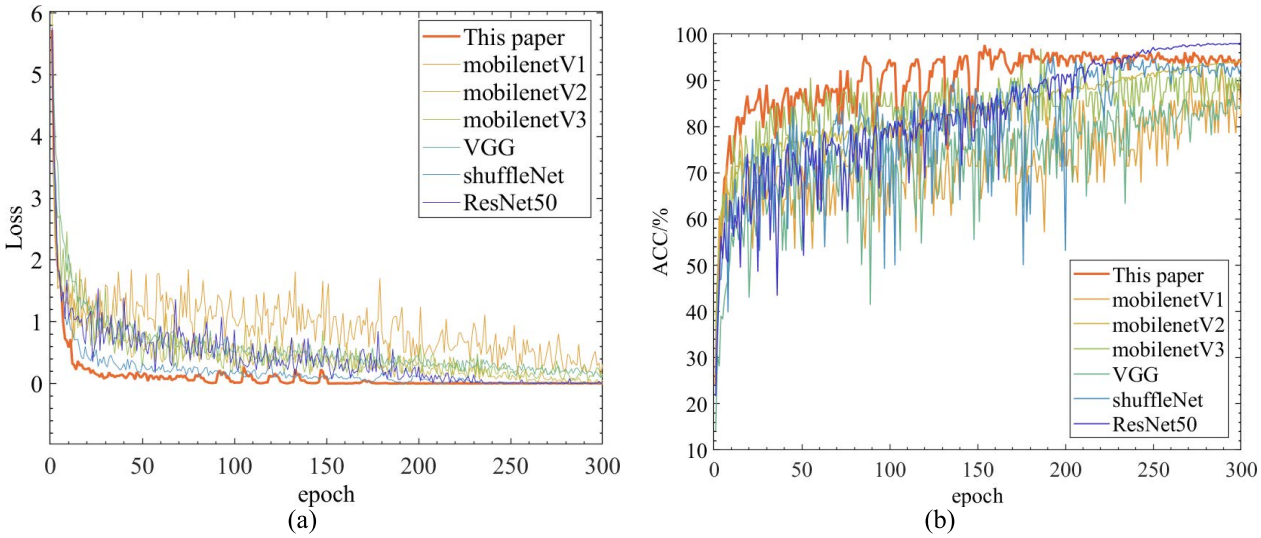

FIGURE 11. Schematic diagram of some model training results including this article. (a) The loss drop curve during the training process of some models including this paper; (b) Accuracy curve during training of some models including this paper.

TABLE 4. Comparison of the effects of each model. Top-1 Accuracy refers to how well the top-ranked category matches the actual results. Top-5 Accuracy refers to how accurately the top five categories contain actual results.

Model	TOP1-ACC	TOP5-ACC	Weights(M)	FLOPS(M)
This paper	95.12%	100%	2.6M	127M
MobileNetV1	90.15%	98.17%	1.3M	150M
MobileNetV2	94.21%	100%	2.2M	141M
MobileNetV3	92.18%	100%	2.4M	44M
ShuffleNetV1	93.72%	100%	1.3M	131M
ShuffleNetV2	89.76%	97.83%	1.3M	138M
ResNet50	97.37%	100%	25.6M	4186M
VGG16	86.74%	97.32%	15M	313M

it builds and the H-swish activation function are more conducive to model training and feature extraction;

2. This paper refers to the backbone network of MobileNet V3, but also reduces the number of backbone layers, and adds a multi-scale feature fusion structure and Bnecks structure, although the added structure introduces a large number of parameters, because the number of layers is reduced and the added structure only acts on the initial stage of the model, the parameter quantity does not change significantly;
3. The multi-scale feature fusion structure introduced by the model is aimed at the fusion of multi-scale features. In the fusion process, the PSA module is added to enhance the spatial and channel information fusion of the model. These improvements enhance the spatial and channel information fusion of the model so that the important channel information and spatial information are retained, and the unimportant information is suppressed at the same time;
4. This model introduces Bnecks into the module. In early feature enrichment phase of the model, ordinary con-volution is used instead of DSC, which can preserve the rich features and transfer them to subsequent modules to improve the final recognition accuracy.

TABLE 5. Model classification accuracy under different snr.

SNR	TOP1-ACC	TOP5-ACC
No noise	95.12%	100%
30dB	95.03%	100%
40dB	95.07%	100%
50dB	94.47%	100%

TABLE 6. Model comparison of two hardware platforms.

Device	TOP1-ACC	TOP5-ACC	Time(ms)	Price($)
Jetson Nano	94.82%	100%	472ms	$150
Jetson TX2	95.01%	100%	86ms	$1000

Then, in order to verify the robustness of the model, this paper adds white noise with SNR (Signal-Noise Ratio) of 30dB, 40dB and 50dB to the original data respectively. Then these noise mixed data are extracted according to the above processing scheme, and recognized with the proposed model. Surprisingly, when the signal-to-noise ratio is 30 dB and 40 dB respectively, the accuracy of model classification hardly changes. When the SNR is 50 dB, the accuracy of the model also decreases by only 0.7%. It is certain that in the process of model training, model proposed in this paper has mastered the key features of bird sound data, and even adding noise signals will not interfere with the classification ability of the model. The comparison results are shown in Table 5.

In addition, this paper builds the model on the Jetson TX2 and Jetson Nano platforms. The cost of the former is about $1000 and the latter is about $150. By comparing the effects

TABLE 7. Performance comparative with the proposed method and other bird sound classification methods.

Study	Classes Number	Method	Performance
Kücüktopcu et al. [34]	21	MFCCs, minimum distance classifier	Accuracy: 72%
Zhang et al. [35]	5	k-nearest neighbor, decision tree, and multi-layer perceptron	Accuracy: 85.3%
Albornoz et al. [36]	25	Linear discriminant analysis, decision tree, and SVM	Accuracy: 90%
Jancˇovicˇ et al. [37]	30	Estimation of frequency tracks	Accuracy: 70%
Xie and Zhu[38]	14	CNN	F1-Score: 95.95%
This paper	264	Mel spectrogram with min-max standardization and Lightweight model	Accuracy: 95.12%

of the models on the two platforms, it is found that there are great differences in the reasoning time of the models on the two platforms, but the accuracy of classification is almost the same. As shown in Table 6. This shows that it is feasible to apply the model to the hardware platform, but the low-cost hardware platform still has the problem of long reasoning time. This is the direction we will continue to study in the future.

Finally, this paper compares the proposed method with other bird sound classification methods, as shown in Table 7. As demonstrated in Table 7, the proposed model obtained a high accuracy while classifying more bird sound classes. It can be seen that the scheme proposed in this paper has a great improvement compared with others' schemes. Firstly, there are many birds in the data set of this paper. Secondly, the features selected in this paper are single, and the feature extraction algorithm is simple. Finally, the model designed in this paper is lightweight enough and the classification accuracy is obviously high.

V. CONCLUSION

In this paper, a lightweight bird song recognition algorithm model is proposed. The classification accuracy of this model can reach 95.12%. Compared with other lightweight networks, the model proposed in this paper has a higher recognition rate. Compared with other depth models, the accuracy of the model of this paper is slightly different, and the number of parameters and computations is reduced. From the analysis of ablation experiments, it can be seen that the improvement proposed in this paper can improve the accuracy of model classification and make the model have a good generalization ability.

The future work of this paper includes:

1. Applying the model to embedded devices to realize real-time bird monitoring in nature reserves;
2. Collecting more bird sound data and constructing large bird datasets;
3. Simplifying birds Sound feature extraction, reducing the steps and processes of feature extraction.

REFERENCES

[1] M. D. Dettling, K. E. Dybala, D. L. Humple, and T. Gardali, "Protected areas safeguard landbird populations in central coastal California: Evidence from long-term population trends," *Ornithol. Appl.*, vol. 123, no. 4, Oct. 2021, Art. no. duab035.

[2] G. D. Duckworth and R. Altwegg, "Why a landscape view is important: Nearby urban and agricultural land affects bird abundances in protected areas," *PeerJ*, vol. 9, Jul. 2021, Art. no. e10719.

[3] H. S. Oliveira and L. dos Anjos, "Silent changes in functionally stable bird communities of a large protected tropical forest monitored over 10 years," *Biol. Conservation*, vol. 265, Jan. 2022, Art. no. 109407.

[4] J. Xie, K. Hu, M. Zhu, J. Yu, and Q. Zhu, "Investigation of different CNN-based models for improved bird sound classification," *IEEE Access*, vol. 7, pp. 175353–175361, 2019.

[5] J. Salamon, J. P. Bello, A. Farnsworth, M. Robbins, S. Keen, H. Klinck, and S. Kelling, "Towards the automatic classification of avian flight calls for bioacoustic monitoring," *PLoS ONE*, vol. 11, no. 11, Nov. 2016, Art. no. e0166866.

[6] C.-Z. Sun, L. Zhen, C. Wang, B.-Y. Yan, X.-C. Cao, and R.-Z. Wu, "Impacts of ecological restoration and human activities on habitat of overwintering migratory birds in the wetland of Poyang Lake, Jiangxi Province, China," *J. Mountain Sci.*, vol. 12, no. 5, pp. 1302–1314, Sep. 2015.

[7] A. D. P. Ramirez, J. I. de la Rosa Vargas, R. R. Valdez, and A. Becerra, "A comparative between mel frequency cepstral coefficients (MFCC) and inverse mel frequency cepstral coefficients (IMFCC) features for an automatic bird species recognition system," in *Proc. IEEE Latin Amer. Conf. Comput. Intell. (LA-CCI)*, Nov. 2018, pp. 1–4.

[8] D. R. Lucio, Y. Maldonado, and G. da Costa, "Bird species classification using spectrograms," in *Proc. Latin Amer. Comput. Conf. (CLEI)*, Oct. 2015, pp. 1–11.

[9] J. Salamon, J. P. Bello, A. Farnsworth, and S. Kelling, "Fusing shallow and deep learning for bioacoustic bird species classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 141–145.

[10] R. Pahuja and A. Kumar, "Sound-spectrogram based automatic bird species recognition using MLP classifier," *Appl. Acoust.*, vol. 180, Sep. 2021, Art. no. 108077.

[11] L. Verde, G. De Pietro, A. Ghoneim, M. Alrashoud, K. N. Al-Mutib, and G. Sannino, "Exploring the use of artificial intelligence techniques to detect the presence of coronavirus COVID-19 through speech and voice analysis," *IEEE Access*, vol. 9, pp. 65750–65757, 2021.

[12] Y. Hu, M. Lu, C. Xie, and X. Lu, "Video-based driver action recognition via hybrid spatial–temporal deep learning framework," *Multimedia Syst.*, vol. 27, no. 3, pp. 483–501, Jan. 2021.

[13] H. Zou and X. Sun, "3D face recognition based on an attention mechanism and sparse loss function," *Electronics*, vol. 10, no. 20, p. 2539, Oct. 2021.

[14] X. Zhang, A. Chen, G. Zhou, Z. Zhang, X. Huang, and X. Qiang, "Spectrogram-frame linear network and continuous frame sequence for bird sound classification," *Ecol. Informat.*, vol. 54, Nov. 2019, Art. no. 101009.

[15] M. Lasseck, "Acoustic bird detection with deep convolutional neural networks," in *Proc. Detection Classification Acoustic Scenes Events Workshop (DCASE)*, 2018, pp. 143–147.

[16] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[17] J. X. Sheng, J. Kun, and X. Jie, "Birdsong recognition algorithm based on multi-dimensional neural network depth feature fusion," *Signal Process.*, vol. 38, no. 4, pp. 844–853, 2022.

[18] Y. Kumar, S. Gupta, and W. Singh, "A novel deep transfer learning models for recognition of birds sounds in different environment," *Soft Comput.*, vol. 26, no. 3, pp. 1003–1023, Feb. 2022.

[19] N. Takahashi, M. Gygli, and L. Van Gool, "AENet: Learning deep audio features for video analysis," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 513–524, Mar. 2017.
[20] J. Xie, J. Yang, Z. Xing, Z. Zhuo, and X. Chen, "Bird species recognition method based on multi-feature fusion," *J. Appl. Acoust.*, vol. 39, no. 2, pp. 199–206, 2020.
[21] X. Dong and J. Jia, "Advances in automatic bird species recognition from environmental audio," *J. Phys., Conf.*, vol. 1544, no. 1, May 2020, Art. no. 012110.
[22] A. Ignatov, A. Romero, H. Kim, and R. Timofte, "Real-time video super-resolution on smartphones with deep learning, mobile AI 2021 challenge: Report," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 2535–2544.
[23] *BirdClef 2019 Challenge*. Accessed: 2019. [Online]. Available: https://www.imageclef.org/BirdCLEF2019
[24] *BirdClef 2020 Challenge*. Accessed: 2020. [Online]. Available: https://www.imageclef.org/BirdCLEF2020
[25] *BirdClef 2021 Challenge*. Accessed: 2021. [Online]. Available: https://www.imageclef.org/BirdCLEF2021
[26] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
[27] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.
[28] M. Tan and Q. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10096–10106.
[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[30] H. Zhang, K. Zu, J. Lu, Y. Zou, and D. Meng, "EPSANet: An efficient pyramid squeeze attention block on convolutional neural network," 2021, *arXiv:2105.14447*.
[31] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[33] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.
[34] O. Küçüktopcu, E. Masazade, C. Ünsalan, and P. K. Varshney, "A real-time bird sound recognition system using a low-cost microcontroller," *Appl. Acoust.*, vol. 148, pp. 194–201, May 2019.
[35] L. Zhang, M. Towsey, J. Xie, J. Zhang, and P. Roe, "Using multi-label classification for acoustic pattern detection and assisting bird species surveys," *Appl. Acoust.*, vol. 110, pp. 91–98, Sep. 2016.
[36] E. M. Albornoz, L. D. Vignolo, J. A. Sarquis, and E. Leon, "Automatic classification of furnariidae species from the paranaense littoral region using speech-related features and machine learning," *Ecol. Informat.*, vol. 38, pp. 39–49, Mar. 2017.
[37] P. Jancovic and M Köküer, "Acoustic recognition of multiple bird species based on penalized maximum likelihood," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1585–1589, Mar. 2015.
[38] J. Xie and M. Zhu, "Handcrafted features and late fusion with deep learning for bird sound classification," *Ecolog. Inform.*, vol. 52, pp. 74–81, Jul. 2019.
[39] J. Hu, L. Shen, and G. Sun, "Squeeze- and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.