Bird Species Identification Using Spectrogram Based on Multi-Channel Fusion of DCNNs

Feiyu Zhang †, Luyang Zhang †, Hongxiang Chen and Jiangjian Xie *

School of Technology, Beijing Forestry University, Beijing 100083, China; zfy15810038216@163.com (F.Z.); zly12730172@163.com (L.Z.); millerchx@gmail.com (H.C.)
* Correspondence: shyneforce@bjfu.edu.cn
† These authors contributed to the work equally and should be regarded as co-first authors.

Abstract: Deep convolutional neural networks (DCNNs) have achieved breakthrough performance on bird species identification using a spectrogram of bird vocalization. Aiming at the imbalance of the bird vocalization dataset, a single feature identification model (SFIM) with residual blocks and modified, weighted, cross-entropy function was proposed. To further improve the identification accuracy, two multi-channel fusion methods were built with three SFIMs. One of these fused the outputs of the feature extraction parts of three SFIMs (feature fusion mode), the other fused the outputs of the classifiers of three SFIMs (result fusion mode). The SFIMs were trained with three different kinds of spectrograms, which were calculated through short-time Fourier transform, mel-frequency cepstrum transform and chirplet transform, respectively. To overcome the shortage of the huge number of trainable model parameters, transfer learning was used in the multi-channel models. Using our own vocalization dataset as a sample set, it is found that the result fusion mode model outperforms the other proposed models, the best mean average precision (MAP) reaches 0.914. Choosing three durations of spectrograms, 100 ms, 300 ms and 500 ms for comparison, the results reveal that the 300 ms duration is the best for our own dataset. The duration is suggested to be determined based on the duration distribution of bird syllables. As for the performance with the training dataset of BirdCLEF2019, the highest classification mean average precision (cmAP) reached 0.135, which means the proposed model has certain generalization ability.

Keywords: bird vocalization; spectrogram feature; multi-channel; deep convolutional neural

1. Introduction

Birds have been widely regarded as important indicators of biodiversity [1], thus, it is significant to monitor bird species. At the species level, bird vocalizations are relatively stable, which can be used for species identification [2,3]. Furthermore, it is possible to monitor bird species by using autonomous recording units [4,5]. Considering the large amount of audio data from long-term recording programs, an efficient, automated identification method of bird species should be induced to shorten the analysis time and decrease the workload.

The spectrogram of bird sounds represents the intensity of the sound signal by different colors or gray values, which contains the time–frequency feature of bird sounds. In the spectrogram, bird vocalization can be seen as a kind of special object. Then, we can identify bird species through the image features of the special object. The key point of automated bird species identification is the extraction of identifiable features of bird vocalizations. Deep learning has a strong ability of self-learning and feature extraction, which can automatically acquire characteristic information from inputs [6]. Koops et al. [7] trained eight deep neural networks with the inputs of mel-frequency cepstral coefficients (MFCCs) of bird audio segments. The results showed that the best network classified 73% correctly. Piczak [8] studied three different DCNNs and a simple ensemble model to complete the LifeCLEF 2016 bird identification task. The highest mean average precision (MAP)

of 52.9% was achieved for foreground species. Toth and Czeba [9] fed the spectrograms into a convolutional neural network (CNN) to realize the classification of bird species. The solution reached the MAP score of 40% for main species. When there were background species, the MAP score of 33% was achieved. Sprengel et al. [10] processed the background noise by image process methods before feeding into the CNN. The MAP score of 0.686 was achieved when identifying the main species. When background species were considered as additional prediction targets, the MAP score decreased to 0.555. Cakir et al. [11] proposed the convolutional recurrent neural networks to realize automated bird audio detection and achieved an 88.5% Area Under ROC Curve (AUC) score on the unseen evaluation data. Ágnes et al. [12] presented a CNN system to classify bird sounds with a spectrogram as input; when using an RGB spectrogram, the highest accuracy was about 82%. Xie et al. [13] proposed a bird call classification model based on VGG16 with three types of spectrogram; the MAP reached 0.9871 when classifying 18 bird species. Xie et al. [14] utilized an autoencoder improved by self-attention as the acoustic classifier of *Nipponia Nippon* individuals: the highest accuracy was 0.971. Kahl et al. [15] designed the BirdNET with a series of residual stacks and classification blocks, aiming to identify 984 bird species by the Fast Fourier Transform (FFT) spectrograms of bird vocalizations. The MAP of 0.791 was achieved for single-species recordings. Turker et al. [16] introduced the discrete wavelet transform (DWT) to calculate the spectrogram, which was used to classify the bird sound. The accuracy of 96.67% was achieved on the 18 classes bird sound dataset. All of the above studies show that the bird species identification methods based on deep learning are effective. Through designing a reasonable network architecture and selecting appropriate input features, the identification accuracy of deep learning can surpass other classification methods.

In the wild, the sample size of rare bird vocalizations may be limited. Moreover, there are regional differences among the vocalizations of birds in different places. Therefore, we should not simply download their vocalizations from websites as the training data. In this situation, the sample size of bird vocalizations is relatively small compared to the demands of a deep neural network model, which tends to cause the overfitting problem when training the neural network model. Transfer learning extracts features from a pretrained model, which decreases the number of trainable parameters significantly, then reduces the demand for the number of samples [17,18]. Thus, it can avoid overfitting. Atoine [19] proposed an efficient bird sound classification method: Soundception, which was realized by the transfer learning of Inception-V4. Soundception reached the MAP of 0.714 in the task of classifying of 1500 bird species.

In this paper, we proposed a DCNN model to overcome the imbalance of dataset and studied the performance of two multiple model fusion schemes in bird vocalization classification, finding the best fusion strategy. The main contributions of this paper are as follows:

(1) Considering the imbalance of the bird vocalization dataset, a single feature identification model (SFIM) was built with residual blocks and modified, weighted, cross-entropy function. Three SFIMs were trained with three kinds of spectrograms calculated by short-time Fourier transform, mel-frequency cepstrum transform and chirplet transform, respectively.

(2) To achieve better performance, two multi-channel fusion models using three different SFIMs were studied. Furthermore, transfer learning was introduced to decrease the number of trainable parameters of fusion models. The resulting fusion mode model outperforms the feature fusion mode model and SFIMs, the best mean average precision (MAP) reaches 0.914.

(3) Through the comparative experiments with different durations of spectrograms, the results revealed that the duration is suggested to be determined based on the duration distribution of bird syllables.

The rest of this paper is as follows. Section 2 describes the dataset and the proposed bird vocalization identification models. Section 3 presents and discusses the experimental results. Finally, Section 4 gives a brief summary of this paper.

2. Materials and Methods

2.1. Datasets

2.1.1. Vocalization Signals

In the breeding season, we recorded the vocalization of birds at Beijing Song-Shan National Nature Reserve (east longitude 115°43′44″–115°50′22″, north latitude 40°29′9″–40°33′35″) with digital solid-state recorder Marantz PMD-671 (MARANTZ, Japan) and directional microphone Sennheiser MKH416-P48 (SENNHEISER ELECTRONIC, German) for many years. The vocalization signals are in 16-bit linear WAV format with 44.1 kHz sampling rate. In this paper, we selected the vocalization signals of eighteen bird species, which have been clearly identified by ornithologists. Each signal only contains the vocalization of one species, and there is no overlap between vocalizations. Table 1 lists the detailed information of eighteen bird species. The column of time means the cumulative time of the vocalization signals.

Table 1. Vocalization signal details of 18 kinds of bird.

Order	Family	Species	Time (s)
Galliformes	Phasianidae	*Phasianus colchicus*	12
		C. micropterus	13
Cuculiformes	Cuculidae	*C. saturatus*	52
		Cuculus sparverioides	34
	Corvidae	*Corvus macrorhynchos*	27
		Urocissa erythroryncha	96
	Turdidae	*Phoenicurus auroreus*	37
		Ficedula zanthopygia	61
	Muscicapidae	*F. narcissina*	82
		F. elisae	49
Passeriformes		*P. major*	54
	Paridae	*Parus palustris*	33
		P. montanus	38
		P. venustulus	26
	Sittidae	*S. villosa*	29
		Sitta europaea	36
	Emberizidae	*Emberiza godlewskii*	23
		E. elegans	71

2.1.2. Signal Pre-Processing

Bird vocalization signal is a kind of non-stationary signal. Before the time–frequency transform, pre-processing is needed. Pre-emphasis filter is used to compensate for the high frequency attenuation of vocalization signal at first. The pre-emphasis coefficient was set to 0.95. After that, the vocalization signal was segmented into frames and windowed using the Hamming window function. We chose the frame length of 50 ms to make sure that at least one fundamental frequency peak was included, and 30% overlap was chosen to divide the vocalization signal into windowed frames. The primary element of bird vocalization is 'notes' that can be combined into syllables, which, in turn, constitute song types. Acoustic classification of bird species mainly focuses on the classification of individual syllables [20]. Segmenting vocalization into distinct syllables is a crucial step. We performed the segmentation operation in the time domain based on energy, the frames with high energy (higher than a half of the maximum energy) are considered to be syllables, otherwise there is silence.

2.1.3. Spectrogram Calculation

Acoustic signals are usually transformed to spectrograms, which can be used to characterize the time–frequency characteristics of bird vocalization. Bird vocalization can be regarded as the special object in spectrogram, where the characteristic of special object represents the time–frequency characteristic of the bird vocalization.

Here, three most frequently used time–frequency transform methods in the audio signal processing, short-time Fourier transform (STFT), mel-frequency cepstral transform (MFCT) and chirplet transform (CT) were utilized to calculate the spectrograms. STFT is one of the earliest time–frequency analysis methods, which presents the energy distribution across linear range of the frequencies. MFCT was proposed to approximately represent the logarithmic frequency sensitivity of human hearing. We calculated 32-dimensional MFCCs through MFCT, the last 31 dimensions were composed to form the Mel spectrograms. CT is a kind of linear time–frequency representation, which refers to the time–frequency representations of each atom on the modulated time–frequency plane. It is a broad class of filters, which include wavelets and Fourier bases as particular cases, and there is an obvious advantage in the representation of short-time stationary signal [21]. We carried out the CT on each frame with fast chirplet decomposition algorithm [22], then, the calculated wavelet coefficients were used to compose the chirplet spectrogram.

2.1.4. Create Sample Sets

We utilized the above three kinds of time–frequency transforms to calculate the spectrograms. Figure 1 represents the signal and its spectrograms of *Phoenicurus auroreus*.
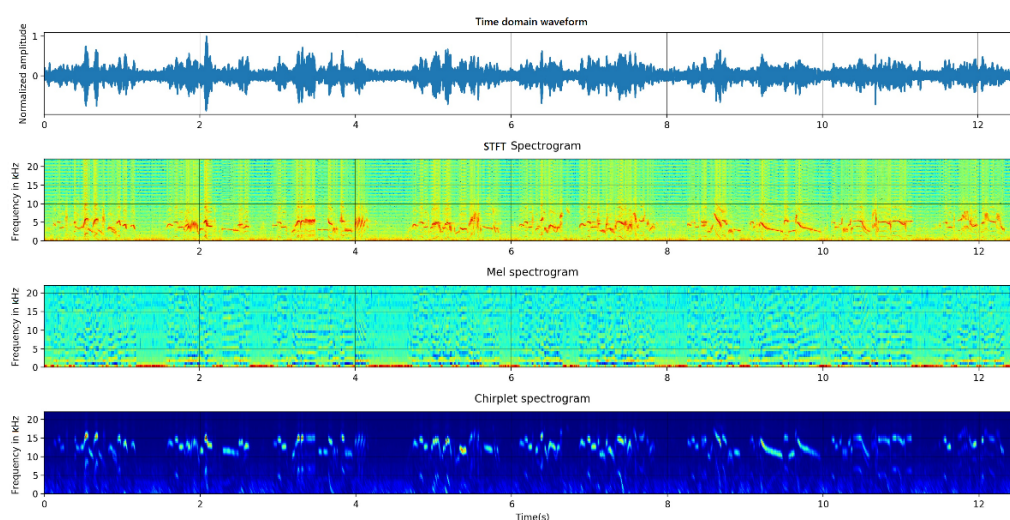

Figure 1. Spectrogram of *Phoenicurus auroreus* (from up to down, they are time domain waveform, STFT spectrogram, mel spectrogram and chirplet spectrogram.).

Bird species identification is always regarded as the classification of individual syllable types [20]. Hence, the spectrograms of a certain duration were saved as the 224×224 RGB color images, instead of the spectrograms of the whole bird vocalization signal. These spectrograms formed the sample set, which will be fed into the identification model. Later, we will discuss the influence of different durations on the performance of the identification models. With three kinds of time–frequency transform methods, three different sample sets with the same size can be built.

2.2. Identification Models

2.2.1. Single Feature Identification Model (SFIM) Based on DCNN

With the spectrogram of bird vocalization as the input, bird species identification can be thought as an image classification problem. DCNNs can self-learn the image features via some convolutional and pooling layers, then classify the features by some fully connected

layers to realize the classification of images [23,24]. The deep residual network (ResNet) is a kind of DCNN model, which yields high performance in the imagenet large scale visual recognition challenge (ILSVRC) of 2015, the top-5 error rate of which was 3.57% [25]. This has been widely used in the field of image recognition [26–28]. Compared with the common DCNN, the main innovation of ResNet is the identity shortcut connection, which was adopted to address the degeneration problem of deeper networks. The simple modification can greatly increase the training speed of the model and improve the training performance without raising extra parameters. We constructed single feature identification model with several residual blocks, its framework is shown in Figure 2, and its configuration is shown in Table 2. N is the number of bird species to be classified.
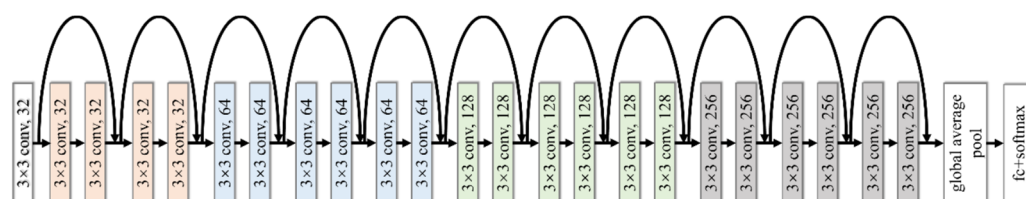

Figure 2. Framework of the single feature identification model.

Table 2. Configuration of the single feature identification model.

Output	Identification Model
$224 \times 224 \times 32$	conv, 3×3, 32, stride 1
$224 \times 224 \times 32$	$\begin{bmatrix} \text{conv}, 3 \times 3, 32, \text{stride} = 1 \\ \text{conv}, 3 \times 3, 32, \text{stride} = 1 \end{bmatrix} \times 2$
$112 \times 112 \times 64$	$\begin{bmatrix} \text{conv}, 3 \times 3, 64, \text{stride} = 2 \\ \text{conv}, 3 \times 3, 64, \text{stride} = 1 \end{bmatrix}$
$112 \times 112 \times 64$	$\begin{bmatrix} \text{conv}, 3 \times 3, 64, \text{stride} = 1 \\ \text{conv}, 3 \times 3, 64, \text{stride} = 1 \end{bmatrix} \times 2$
$56 \times 56 \times 128$	$\begin{bmatrix} \text{conv}, 3 \times 3, 128, \text{stride} = 2 \\ \text{conv}, 3 \times 3, 128, \text{stride} = 1 \end{bmatrix}$
$56 \times 56 \times 128$	$\begin{bmatrix} \text{conv}, 3 \times 3, 128, \text{stride} = 1 \\ \text{conv}, 3 \times 3, 128, \text{stride} = 1 \end{bmatrix} \times 2$
$28 \times 28 \times 256$	$\begin{bmatrix} \text{conv}, 3 \times 3, 256, \text{stride} = 2 \\ \text{conv}, 3 \times 3, 256, \text{stride} = 1 \end{bmatrix}$
$28 \times 28 \times 256$	$\begin{bmatrix} \text{conv}, 3 \times 3, 256, \text{stride} = 1 \\ \text{conv}, 3 \times 3, 256, \text{stride} = 1 \end{bmatrix} \times 2$
$1 \times N$	global average pool, full connect (fc), softmax

2.2.2. Multi-Channel Identification Models

With three kinds of spectrograms as inputs, three different single feature identification models (SFIMs) can be achieved. We further fused three SFIMs together to improve the efficiency and accuracy of the identification. Here, each SFIM is separated to two parts: the feature extraction part and the classifier part. The classifier part only contains the full connect layers and softmax layer. Two fusion modes were designed, one is feature fusion mode, which fuses directly the feature outputs of three SFIMs, its structure is shown in Figure 3.
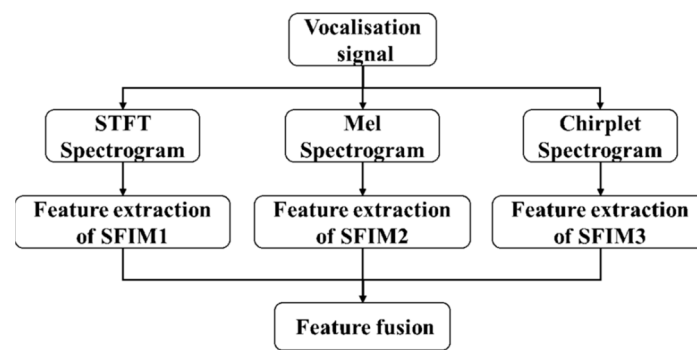
Figure 3. The structure of feature fusion model.

The other is result fusion mode, which fuses the classifier outputs of three SFIMs, its structure is shown in Figure 4.
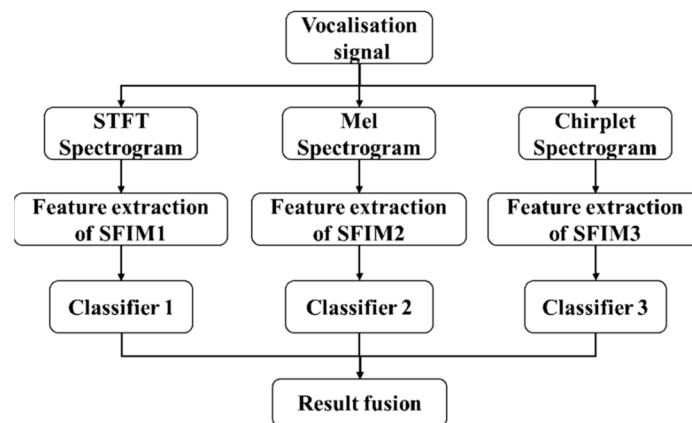

Figure 4. The structure of multi-channel identification models.

Adaptive linear weighted method is utilized to fulfil the fusion operation, which can ensure that the dimensions of 1 output and input are the same without adding model parameters. The fusion operations of both fusion modes are the same. The fused feature *F* is given by

$$F = \sum_{n=1}^{3} \omega^n y^n \qquad (1)$$

where ω^n and y^n are the weight and the vector to be fused of feature n. Different weights indicate the contribution of different vectors in the identification, and $\sum_{n=1}^{3} \omega^n = 1$ should be satisfied. The weights are updated in the training until the optimal values are achieved.

After the fusion operation, two fully-connected layers and a softmax layer are selected as the classification part to realize the classification. To decrease the number of trainable parameters of fusion models, the parameter-based transfer learning is used here. The parameters of each SFIM were frozen, only the parameters of fusion and classification part were trained.

3. Results and Discussion

3.1. Experimental Setup

The experiments were conducted on an Ubuntu16.04 Linux workstation with 32 G memory, one E5-2620CPU (6×2.1 GHz) and two GTX1080ti GPUs (11 GB memory). The models were programmed based on the deep learning framework Tensorflow1.9.

Because the number of spectrograms in each sample set is not large, only a training and test set were built. Each sample set was randomly split into a training and test set with

a ratio of 7:3. Based on these samples, the identification models were trained and verified. The training set was divided into several batches to speed up the training process. The detailed training settings are listed in Table 3. Early stopping was used to avoid overfitting.

Table 3. Training settings.

Items	Value or Method
Batch size	50
Parameter initialization	Random initialization
Optimization algorithm	Adam [29]
Learning rate	0.001
Epochs	100

As Table 1 shows, the cumulative time of different bird vocalization signals are different; also, the number of spectrograms of different bird species in the sample set is quite different. In other words, sample sets are unbalanced, which is not beneficial to the training of DCNN models [30]. We proposed a kind of weighted cross-entropy as the loss function. The loss function can increase the weights of the bird species that have few samples, so that the problem of unbalanced data can be solved. For multi-class classification, the improved cross-entropy loss of the ith class is

$$\text{WCE}_i = -\eta_i y_i \log \widehat{y}_i - (1 - y_i) \log(1 - \widehat{y}_i) \tag{2}$$

where y_i represents whether the sample belongs to the ith class, its value is 1 when the sample belongs to the ith class, otherwise it is 0. $\widehat{y}_i$ denotes the prediction probability that the sample belongs to the ith class. η_i is the weight of the ith class, which is determined by the following equation

$$\eta_i = \frac{1 - \beta_i}{\beta_i} \tag{3}$$

where β_i indicates the ratio of the sample size of the ith class to the whole sample size.

The improved cost function is presented as follows

$$\text{Cost} = \frac{1}{N_B} \sum_{j=1}^{N_B} \beta_{ij} \cdot \text{WCE}_{ij} \tag{4}$$

where N_B is the value of the batch size, the subscript j is the index of samples in a batch, the subscript i denotes the corresponding value of the sample belonging to the ith class. i can be any class index of the sample in the current batch.

3.2. Different Models with the Same Duration of Spectrogram

In the experiments, there are five models, including three SFIMs, the multi-channel model with result fusion (Re-fuse) and the multi-channel model with feature fusion (Fe-fuse). The spectrogram sample sets of 300 ms duration were used to train all the models. Firstly, three spectrogram sample sets, including Ch, Mel and Spe, were fed to three SFIMs, respectively. After that, we froze the parameters of three SFIMs and trained the Re-fuse and Fe-fuse models with the Ch, Mel and Spe spectrograms at the same time.

Mean average precision (MAP) is commonly used to evaluate the performance of an identification model, which is defined as

$$\text{MAP} = \frac{\sum_{q=1}^{N} Ave\text{P}(q)}{N} \tag{5}$$

where q is the class index and $Ave\text{P}(q)$ is the average identification accuracy of the qth class samples.

Table 4 shows the test MAPs of all the above models. Figure 5 shows the test MAP variations at different steps.

Table 4. MAPs of all the identification models.

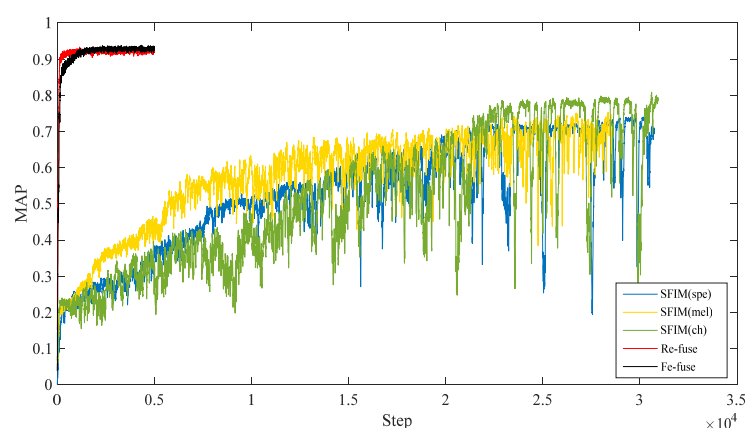| Model	MAP
SFIM (Spe)	0.742
SFIM (Mel)	0.754
SFIM (Ch)	0.808
Re-fuse	0.914
Fe-fuse	0.908

Figure 5. Test MAPs of the proposed models.

As shown in Table 4 and Figure 5, it was found that the SFIM(Ch) achieved the highest MAP of 0.808 in three SFIMs, the second was SFIM(Mel), which is consistent with the results of [22]. This demonstrates that the Ch spectrogram is more suitable for bird species identification when using the image classification method. From three kinds of exemplar spectrograms of *Ficedula zanthopygia* listed in Figure 6, it was found that the differences between the bird vocalization region (the redder region) and the background are most obvious in the Ch spectrogram than the other two kinds of spectrograms, and the bird vocalization region of the Ch spectrogram is more compact. All of the above factors enable the best feature extraction ability using the Ch spectrogram as inputs.
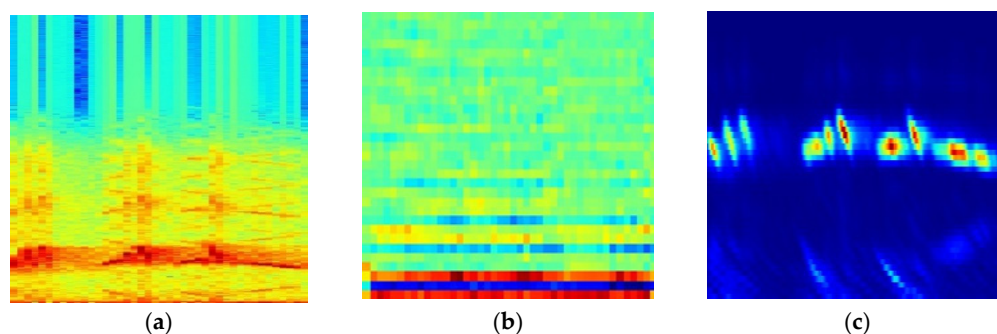

Figure 6. Three kinds of exemplar spectrograms of *Ficedula zanthopygia*. (**a**) Spe spectrogram, (**b**) Mel spectrogram, (**c**) Ch spectrogram.

As for both multi-channel models, the highest MAPs were higher than those of all the SFIMs. The largest MAP difference is between the Re-fuse model and SFIM (Spe), the MAP of the Re-fuse model is 23.2% higher than SFIM (Spe). The smallest MAP difference is between the Fe-fuse model and SFIM (Ch), and the MAP of the Fe-fuse model is 12.4%

higher than SFIM (Ch). The above analysis indicates that using selected multi-features as the input of the network can improve the bird identification accuracy to some extent. Furthermore, because transfer learning was used in the multi-channel models, this shows that the convergences of multi-channel models are much faster than SFIMs, which means that multi-channel models are more efficient. Considering the different fusion modes, the highest MAP of the Re-fuse model is a little higher than that of the Fe-fuse model, and the Re-fuse model arrived at the highest MAP faster. On the other hand, the Re-fuse model has fewer trainable parameters than all of the other models, which is advantageous to realize bird identification when limited samples are available.

3.3. Different Models with Different Durations of Spectrogram

Different durations of spectrogram may affect the performances of the bird identification models. We chose the durations of 100 ms, 300 ms and 500 ms to train the proposed models. A comparison of test MAPs between different models is shown in Figure 7. We found that when the spectrogram sample sets of 300 ms duration were fed in each of the models, the test MAPs were the highest. The worst performances came from the duration of 100 ms.
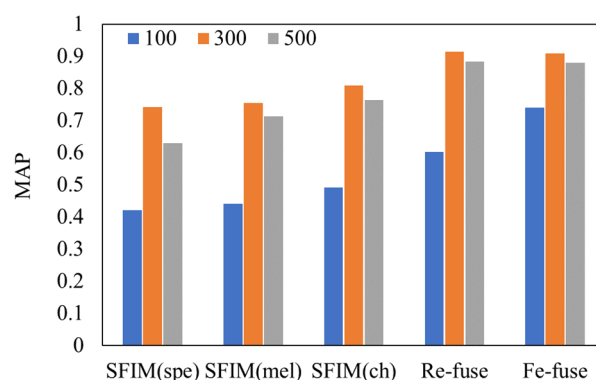

Figure 7. Comparison of test MAP between different models and different durations.

We concluded the syllable durations of the bird species in our dataset. According to the statistics, the syllable durations of the eighteen bird species are between 100 ms and 250 ms. When 100 ms was chosen as the duration, a part of the syllable may be cut off into different spectrograms so that the complete feature cannot be obtained by the classification of spectrograms. As for 500 ms, the number of samples decreased. Hence, compared with the duration of 300 ms, the identification performances deteriorate with the durations of 100 ms and 500 ms, and the impact in the case of the 100 ms duration is more serious. We suggest that the appropriate duration should be selected according to the duration distribution of the identified bird species.

3.4. Performance with BirdCLEF2019 Dataset

The BirdCLEF2019 dataset contains about 350 h of soundscapes, which was built for the 2019BirdCLEF challenge [31]. The labels of the test data were not published. We only used the training data to evaluate our fusion model. The training data covered 659 species from South and North America. Because all data were recorded in the wild, compared with our own dataset and the ICML4 B dataset, there were multi-species, high-level ambient noises in the soundscapes. Due to the above differences, the classification mean average precision (cmAP) was proposed as the evaluation metric in the challenge, considering each class c of the ground truth as a query [31]. The cmAP is defined as

$$cmAP = \frac{\sum_{c=1}^{C} AveP(c)}{C} \tag{6}$$

where C is the number of species in the ground truth and $Ave\mathrm{P}(c)$ is the average precision for a given species c computed as

$$Ave\mathrm{P}(c) = \frac{\sum_{k=1}^{n_c} P(k) \times \mathrm{rel}(k)}{n_{rel}(c)} \quad (7)$$

where k is the rank of an item in the list of the predicted segments containing c, n_c is the total number of predicted segments containing c, $P(k)$ is the precision at cut-off k in the list, rel(k) is an indicator function equaling 1 if the segment at rank k is a relevant one and $n_{rel}(c)$ is the total number of relevant segments for class c. The training dataset was randomly divided into ten parts, seven parts were used as training samples, three parts were set as validation samples. During the training step, the duration of 500 ms was selected to compute the spectrograms. The Re-fuse model was trained with the training samples and validated with validation samples.

We performed the experiments five times, the achieved highest cmAP is 0.135, which is a little lower than the results of the ASAS team. The cmAPs of the ASAS team were between 0.140 and 0.160, which made them win the second place [31], and our cmAP is much higher than that of the third place, which is 0.054. The ASAS team used Inception and ResNet architectures and data augmentation to conduct their experiments. They attributed their good result to the sophisticated augmentation strategies. Although the cmAP is a little lower than that of the ASAS team, our model is simpler, the number of trainable parameters is smaller, and we do not need complicated augmentation, which makes our model more efficient. The performance of the Re-fuse model on the BirdCLEF2019 dataset shows that although there are high-level noises, the proposed Re-fuse model also can achieve a relatively high identification accuracy.

4. Conclusions

In this study, we proposed the SFIM based on the residual block, then trained three SFIMs with three kinds of spectrograms, Spe, Mel and Ch spectrogram, respectively. Furthermore, based on these three SFIMs, we built two available multi-channel fusion models to improve the identification accuracy of the bird species. Transfer learning was utilized to reduce the size demand of samples. The experiments reveal that the performance of the Re-fuse model is the best compared with other proposed models, the MAP is highest, and the trainable parameter number is the smallest, which leads to the smallest demand of samples. With different durations of spectrograms as inputs, the performances are various. We recommend that researchers choose suitable durations based on the duration distribution features of bird vocalizations to be identified. Our proposed fusion method achieves relatively high performance in the BirdCLEF2019 dataset, and we are also aware that the best cmAP of BirdCLEF2019 is 0.356, which is much higher than ours. Thus, we will try to add other strategies to the Re-fuse model to improve the bird identification performance in the future.

Author Contributions: Conceptualization and methodology, J.X.; software, H.C.; validation, F.Z., J.X. and L.Z.; writing-original draft preparation, L.Z. and H.C.; writing-review and editing, J.X. and F.Z.; supervision, J.X.; All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data required to reproduce these findings cannot be shared at this time as the data also forms part of an ongoing study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Priyadarshani, N.; Marsland, S.; Castro, I. Automated birdsong recognition in complex acoustic environments: A review. *J. Avian Biol.* **2018**, *49*, 1–52. [CrossRef]
2. Green, S.; Marler, P. The Analysis of Animal Communication. *J. Theor. Biol.* **1961**, *1*, 295–317.
3. Graciarena, M.; Delplanch, M.; Shriberg, E.; Stolcke, A. Bird species recognition combining acoustic and sequence modeling. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Prague, Czech, 22–27 May 2011; pp. 1–4.
4. Kalan, A.K.; Mundry, R.; Wagner, O.J.J.; Heinicke, S.; Boesch, C.; Kühl, H.S. Towards the automated detection and occupancy estimation of primates using passive acoustic monitoring. *Ecol. Indic.* **2015**, *54*, 217–226. [CrossRef]
5. Perez-Granados, C.B.G.; Giralt, D.; Barrero, A.; Julia, G.C.; Bustillo, D.L.A.R.; Traba, J. Vocal activity rate index: A useful method to infer terrestrial bird abundance with acoustic monitoring. *Ibis* **2019**, *161*, 901–907. [CrossRef]
6. Dan, X.; Huang, S.; Xin, Z. Spatial-aware global contrast representation for saliency detection. *Turk. J. Electr. Eng. Comput. Sci.* **2019**, *27*, 2412–2429.
7. Koops, H.V.; Van Balen, J.; Wiering, F.; Multimedia, S. A Deep Neural Network Approach to the LifeCLEF 2014 bird task. *LifeClef Work. Notes* **2014**, *1180*, 634–642.
8. Piczak, K. Recognizing Bird Species in Audio Recordings Using Deep Convolutional Neural Networks. *CEUR Workshop Proc.* **2016**, *1609*, 1–10.
9. Toth, B.P.; Czeba, B. Convolutional Neural Networks for Large-Scale Bird Song Classification in Noisy Environment. In Proceedings of theConference and Labs of the Evaluation Forum, Évora, Portugal, 5–8 September 2016; pp. 1–9.
10. Sprengel, E.; Jaggi, M.; Kilcher, Y.; Hofmann, T. Audio Based Bird Species Identification using Deep Learning Techniques. In Proceedings of the CEUR Workshop, Evora, Portugal, 5–8 September 2016; pp. 547–559.
11. Cakir, E.; Adavanne, S.; Parascandolo, G.; Drossos, K.; Virtanen, T. Convolutional recurrent neural networks for bird audio detection. In Proceedings of the 2017 25th European Signal Processing Conference (EUSIPCO), Kos, Greek Island, 28 August–2 September 2017; pp. 1794–1798.
12. Agnes, I.; Henrietta-Bernadett, J.; Zoltan, S.; Attila, F.; Csaba, S. Bird sound recognition using a convolutional neural network. In Proceedings of the 2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY), Subotica, Serbia, 13–15 September 2018; pp. 000295–000300.
13. Xie, J.; Li, W.; Zhang, J.; Ding, C. Bird species recognition method based on Chirplet spectrogram feature and deep learning. *J. Beijing For. Univ.* **2018**, *40*, 122–127.
14. Xie, J.; Yang, J.; Ding, C.; Li, W. High accuracy individual identification model of crested ibis (Nipponia Nippon) based on autoencoder with self-attention. *IEEE Access* **2020**, *8*, 41062–41070. [CrossRef]
15. Kahl, S.; Wood, C.M.; Eibl, M.; Klinck, H. BirdNET: A deep learning solution for avian diversity monitoring. *Ecol. Inform.* **2021**, *61*, 101236. [CrossRef]
16. Turker, T.; Erhan, A.; Sengul, D. Multileveled ternary pattern and iterative ReliefF based bird sound classification. *Appl. Acoust.* **2021**, *176*, 107866.
17. Zhuang, F.; Ping, L.; Qing, H.; Shi, Z. Survey on transfer learning research. *J. Softw.* **2015**, *26*, 26–39.
18. Zebhi, S.; Almodarresi, S.M.T.; Vahid, A. Human activity recognition by using MHIs of frame sequences. *Turk. J. Electr. Eng. Comput. Sci.* **2020**, *28*, 1716–1730. [CrossRef]
19. Antoine, S.; Herve, G. Audio Bird Classification with Inception-v4 extended with Time and Time-Frequency Attention Mechanisms. *LifeClef Work. Notes* **2017**, *1866*, 1–8.
20. Potamitis, I.; Ntalampiras, S.; Jahn, O.; Riede, K. Automatic bird sound detection in long real-field recordings: Applications and tools. *Appl. Acoust.* **2014**, *80*, 1–9. [CrossRef]
21. Bultan, A. A four-parameter atomic decomposition of chirplets. *IEEE Trans. Signal Process.* **2002**, *47*, 731–745. [CrossRef]
22. Glotin, H.; Ricard, J.; Balestriero, R. Fast Chirplet Transform to Enhance CNN Machine Listening—Validation on Animal calls and Speech. *arXiv* **2016**, arXiv:1611.08749.
23. Rizvi, M.; Deb, K.; Khan, M.; Kowsar, M.; Khanam, T. A comparative study on handwritten Bangla character recognition. *Turk. J. Electr. Eng. Comput. Sci.* **2019**, *27*, 3195–3207. [CrossRef]
24. Akram, A.; Debnath, R. An automated eye disease recognition system from visual content of facial images using machine learning techniques. *Turk. J. Electr. Eng. Comput. Sci.* **2020**, *28*, 917–932. [CrossRef]
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
26. Tekeli, U.; Bastanlar, Y. Elimination of useless images from raw camera-trap data. *Turk. J. Electr. Eng. Comput. Sci.* **2019**, *27*, 2395–2411. [CrossRef]
27. Xie, J.; Li, A.; Zhang, J.; Cheng, Z. An Integrated Wildlife Recognition Model Based on Multi-Branch Aggregation and Squeeze-And-Excitation Network. *Appl. Sci.* **2019**, *9*, 2794. [CrossRef]
28. Liu, S.; Tian, G.; Xu, Y. A novel scene classification model combining ResNet based transfer learning and data augmentation with a filter. *Neurocomputing* **2019**, *338*, 191–206. [CrossRef]
29. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015; pp. 1–15.

30. Cai, Z.; Fan, Q.; Feris, R.S.; Vasconcelos, N. A Unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection. In Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 354–370.

31. Kahl, S.; Stoter, F.R.; Goeau, H.; Glotin, H.; Planque, R.; Vellinga, W.P.; Joly, A. Overview of BirdCLEF 2019: Large-Scale Bird Recognition in Soundscapes. Technical Report for 2019BirdCLEF Challenge. Available online: https://hal.umontpellier.fr/hal-02345644/document (accessed on 4 November 2019).