birds

MDPI

Article

Acoustic Classification of Bird Species Using an Early Fusion of Deep Features

Jie Xie [1,*] and Mingying Zhu [2,3]

[1] School of Computer and Electronic Information, School of Artificial Intelligence, Nanjing Normal University, Nanjing 210024, China
[2] School of Economics, Nanjing University, Nanjing 210093, China
[3] Johns Hopkins University-Hopkins-Nanjing Center for Chinese and American Studies, Nanjing University, Nanjing 210093, China

Simple Summary: Bird populations have been experiencing a rapid decline. Bird monitoring plays an essential role in bird protection because it can help evaluate the conservation status of species. For bird monitoring, recent work often uses acoustic sensors for bird sound collection and an automated bird classification system for recognizing bird species. To build an automated bird classification system, traditional machine learning needs to manually extract features and train the model to recognize bird sounds. Recent advances in deep learning models make it possible to automatically extract features. However, few studies have explored it for bird species classification. This study investigated various transfer learning models for extracting deep cascaded features with multi-view spectrograms. We found that deep cascade features being extracted from VGG16 and MobileNetV2 can achieve the best performance using repeat-based spectrogram.

Abstract: Bird sound classification plays an important role in large-scale temporal and spatial environmental monitoring. In this paper, we investigate both transfer learning and training from scratch for bird sound classification, where pre-trained models are used as feature extractors. Specifically, deep cascade features are extracted from various layers of different pre-trained models, which are then fused to classify bird sounds. A multi-view spectrogram is constructed to characterize bird sounds by simply repeating the spectrogram to make it suitable for pre-trained models. Furthermore, both mixup and pitch shift are applied for augmenting bird sounds to improve the classification performance. Experimental classification on 43 bird species using linear SVM indicates that deep cascade features can achieve the highest balanced accuracy of 90.94% $\pm$ 1.53%. To further improve the classification performance, an early fusion method is used by combining deep cascaded features extracted from different pre-trained models. The final best classification balanced accuracy is 94.89% $\pm$ 1.35%.

Keywords: bioacoustics; transfer learning; feature selection

Citation: Xie, J.; Zhu, M. Acoustic Classification of Bird Species Using an Early Fusion of Deep Features. *Birds* **2023**, *4*, 138–147. https://doi.org/10.3390/birds4010011

Academic Editor: Jukka Jokimäki

Received: 16 November 2022
Revised: 23 February 2023
Accepted: 27 February 2023
Published: 1 March 2023

1. Introduction

With the development of sensor techniques, acoustic sensors are now widely deployed in the wild for collecting bird sounds [1]. Each day, several gigabytes of compressed data can be generated by an acoustic sensor, which makes it difficult to manually analyze the data [2]. Automated analysis of bird sounds is becoming very important [3–5].

Using the collected bird sounds, there are still several research questions that need to be answered. For instance, (1) how many bird species are there in the place where bird sounds are collected? (2) What is the number of each interesting bird species? (3) What is the relationship between a bird's vocalization and its behavior? Those questions can be partly answered by analyzing the collected sounds [6,7]. However, thus far most previous studies focused on the estimation of bird species in short recordings. Compared to the

identification of an individual bird and detecting the start and stop time of bird song, this is relatively simple [8]. For bird species classification in acoustic data, most earlier work directly used and further modified speech/speaker recognition models to classify bird species [9–11]. Recently, deep learning has been applied in the acoustic classification of bird species [12,13]. However, there are still some challenges in achieving satisfactory classification results for wild recordings, such as the small amount of annotated data and the imbalance among bird species [14]. To deal with the small amount of annotated bird data, there are two popular strategies: transfer learning and training from scratch using a self-defined lightweight model.

Transfer learning is used to solve the basic problem of insufficient annotated training data [15]. For the acoustic classification of bird species, various studies have used transfer learning. Sevilla and Glotin presented an adaptation of the deep convolutional network Inception-v4 tailored to solving bioacoustic classification problems [16]. Ntalampiras proposed a Transfer Learning framework exploiting the probability density distributions of ten different music genres for acquiring a degree of affinity between the bird species and each music genre [17]. Zhong et al. evaluated three different models including convolutional neural network (CNN) using VGG16 architecture, transfer learning (ResNet50) and fine-tuning with pre-trained CNN model, and transfer learning (ResNet50) with custom loss function and pseudo labeling [18,19]. Kumar et al. verified multiple deep transfer learning models such as ResNet50, DenseNet201, InceptionV3, Xception, and EfficientNet for classifying 22 bird species [20]. Dufourq et al. compared 12 transfer learning models with or without fine-tuning for recognizing four bird species [21].

In addition to transfer learning, some previous studies used self-defined CNN models for bird sound classification. Xie et al. used a VGGish model to classify 43 bird species [22]. Sinha et al. used braided convolutional neural networks to classify bird species [23]. Ruff et al. proposed a CNN model including four convolutional layers followed by two fully connected layers [24]. Permana et al. presented a bird sounds classification study using a self-defined CNN model having four convolutional layers followed by three fully connected layers [25].

In this study, we investigated both transfer learning and training from scratch for bird sound classification using five transfer learning models: VGG16, ResNet50, EfficientB0, MobileNetV2, and Xception. For transfer learning, the VGG16 model was used to compare transfer learning and training from scratch. Specifically, we extracted deep cascade features from selected layers. Then, extracted deep features from different transfer learning models were fused and put into linear support vector machines (SVM) for classifying bird species. As for training from scratch, we unfroze all layers of those transfer learning models and built an end-to-end bird species classification framework.

2. Materials and Methods

This section presents the framework of acoustic classification of bird species, as shown in Figure 1. Here, we first preprocess bird acoustic data which are then characterized using multi-view spectrograms. Next, five pre-trained models are investigated to find the best-performing model. Finally, the deep cascade feature extracted from the best-performing model is fused with other models to achieve the best performance using linear SVM.

2.1. Bird Recordings

A publicly available dataset, named CLO-43DS, is used to evaluate our proposed method [26]. This dataset includes flight calls of 43 different North American wood-warblers. The number of instances for all bird species is shown in Supplementary Figure S1. All audio clips are recorded in different conditions using different recording devices including highly directional microphones and omnidirectional microphones. In addition, the signal-to-noise ratio varies among different recordings. The common names, scientific names, and abbreviations of those 43 wood-warbler species involved in this dataset can be found in Supplementary Table S1.

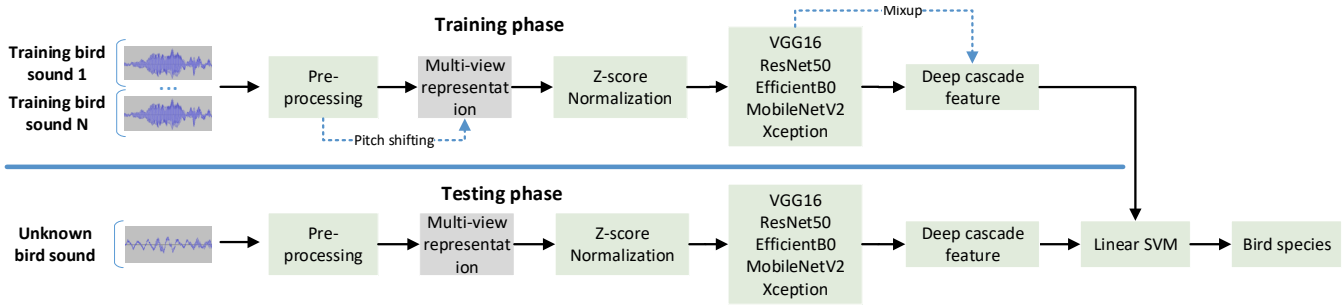


Figure 1. The flow diagram of our proposed approach. Here, data augmentation is only applied in the training phase.

2.2. Preprocessing and Multi-View Representation

For signal preprocessing, the duration of audio files in the CLO-43DS data is different, and they cannot be directly used as the input to CNN. Therefore, the multi-variate varying length audio data is repeated from the beginning to force the fixed duration of 2 s, which has been used in [22,27].

In the CLO-43DS data, each audio clip is processed and clipped to only contain a single flight call. Those audio clips are sampled at 22.05 kHz. The provided Mel-spectrograms by the authors are obtained using 11.6 ms frame size with an overlap of 1.45 ms and 120 Mel-bands. It must be noted that 11.6 ms frame size is optimal for analysis of flight calls as suggested by [28]. Finally, the Mel-spectrogram is converted to a log-magnitude representation for further analysis.

In this study, the transfer learning model is used as a deep cascade feature extraction program to characterize bird calls. Since the input of the transfer learning model should be a three-channel feature representation, here we simply repeat Mel-spectrogram three times to construct a multi-view spectrogram. In addition to repeated multi-view spectrograms, we included HPSS (harmonic–percussive source separation)-based and delta-based multi-view spectrograms for comparison.

The weights of selected transfer learning models are originally trained using natural images (ImageNet [29]), which are different from Mel-spectrograms. Therefore, we assume that training from scratch using pre-trained models can achieve better performance than fine-tuning, which will be verified in the Experimental section.

2.3. Deep Cascade Feature

In this study, transfer learning is selected to obtain deep cascade features. Here, we investigate five pre-trained models that have been trained on ImageNet: VGG16, ResNet50, EfficientB0, MobileNetV2, and Xception. Specifically, the weights of those pre-trained models are applied by selectively freezing layers of the model while training the model. Take VGG16 for instance; it is shown in Supplementary Figure S2 whose original input size is $224 \times 224 \times 3$. However, the size of generated multi-view representation is $120 \times 198 \times 3$. Therefore, we ignore the top layer when loading VGG16. As for the feature construction, we apply a global averaging pooling layer to four Conv layers which are further fused to obtain the final deep cascade feature. Here, the dimension of the final feature set is 1408. After the concatenated layer, we add a dense layer (512) and a dropout layer whose rate is set to 0.5, which is used to address over-fitting. Finally, a dense layer with the number of bird species (43) is added. For other pre-trained models, the layers that are selected and concatenated are shown in Table 1.

Table 1. Layers to concatenate deep features for five different transfer learning models. For all transfer learning models, four individual layers are selected for extracting features except Xception, where three layers are used for feature extraction.

TL Model	Selected Layers for Generating Concatenated Deep Features			
VGG16	block2_conv2	block3_conv3	block4_conv3	block5_conv3
ResNet50	conv2_block3_out	conv3_block4_out	conv4_block6_out	conv5_block3_out
MobileNetV2	block_13_project_BN	block_14_add	block_15_add	global_average_pooling2d
EfficientNetB0	block4c_add	block5c_add	block6d_add	avg_pool
Xception	block4_sepconv1	block5_sepconv1	block14_sepconv1	—

All deep models are trained using an Adam optimizer with an initial learning rate of 10^{-4}. The categorical loss entropy and focal loss are utilized as the loss function. The batch size is 32 samples, and the network is trained with 500 epochs. In addition, early stopping is applied for addressing over-fitting.

2.4. Linear SVM

For the classification, we use the linear SVM based on LIBLINEAR library [30]. Here, an SVM model is considered as a representation of the data or a map of these data to points in space; the data of different categories are divided by the hyperplane. New data are then mapped into that same space, and a prediction is made regarding the category they belong to, based on the side of the hyperplane they fall into. As for the hyper-parameters, the default setting is used. To verify the effectiveness of using linear SVM for the classification, we also report the result with the softmax layer.

2.5. Offline Data Augmentation: Pitch Shifting

For offline data augmentation, pitch shifting aims to modify the original pitch of a sound. This technique has been demonstrated to be an effective method in previous audio data analysis [31,32]. Here, pitch shifting is implemented using *librosa.effects* with two factors 2 and -2. The examples of *HEWA* and *GWWA* using pitch shifting are shown in Supplementary Figure S3.

2.6. Online Data Augmentation: Mixup

Mixup has been demonstrated to be an effective method to augment training data, where pairs of data samples are mixed [33]. Let us define two different audio data x_1 and x_2, and their respective one-hot encoded labels y_1 and y_2. Then, we obtain the mixed sample and its label by a convex combination as follows:

$$\begin{aligned} x_{min} &= \lambda x_1 + (1-\lambda)x_2 \\ y_{min} &= \lambda y_1 + (1-\lambda)y_2 \end{aligned} \tag{1}$$

where λ is a scalar sampled from a symmetric Beta distribution at each mini-batch generation:

$$\lambda \sim Beat(\lambda, \lambda) \tag{2}$$

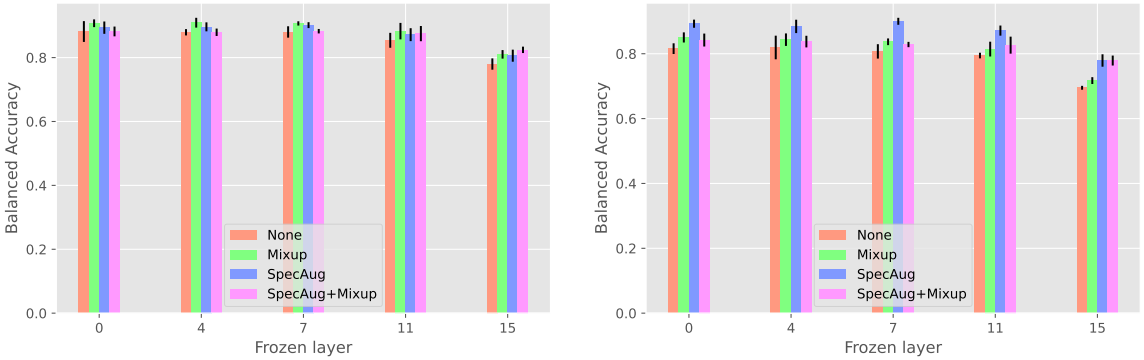
where λ is a real-valued hyper-parameter to tune. In this study, the value of λ is set to 0.2.

3. Results

3.1. Comparison of Data Augmentation Methods

Figure 2 shows the classification results using pitch shifting or mixup. Note that the error bars are also shown in Figure 2 and the following figures where each error bar is a distance of one standard deviation above or below the average classification accuracy. We observe that the use of online or offline augmentation can improve performance. This result verifies the effectiveness of augmentation techniques. For end-to-end classification, pitch shifting (SpecAug) achieves the best performance, and the highest balanced accuracy

is 90.00% with seven frozen layers. Here, frozen layers denote that the weights of those layers in the trained model do not change when reused on a subsequent downstream mission. Using the deep cascade feature with linear SVM, the highest balanced accuracy is 90.92% with four frozen layers. However, the difference between zero frozen layers (train from scratch) and four frozen layers is only significant when using SpecAug + mixup as the data augmentation method, with a significance level of 1% (t = 2.43, df = 4084, $p = 0.0076$). For other augmentation methods, they do not show a statistical difference ($p > 0.1$).



(**a**) (**b**)

Figure 2. (**a**) Classification performance using deep cascade features with linear SVM, (**b**) End-to-end classification using softmax. For deep cascade features using linear support vector machines (SVM), mixup is slightly better than other methods except for having 15 frozen layers. For end-to-end classification, spectral augmentation (SpecAug) achieves the best performance.

3.2. Comparison of Different Multi-View Representations

Simply repeated spectrogram achieves the highest balanced (90.94% $\pm$ 1.53%), which is obtained by four frozen layers (Figure 3). In contrast, when all layers are frozen, the performance is the worst. One reason is that the ImageNet weights are obtained by natural images. However, the spectrogram is different from natural images. Thus, training from scratch (0) or fine-tuning with a few frozen layers (4, 7) can obtain better performance.

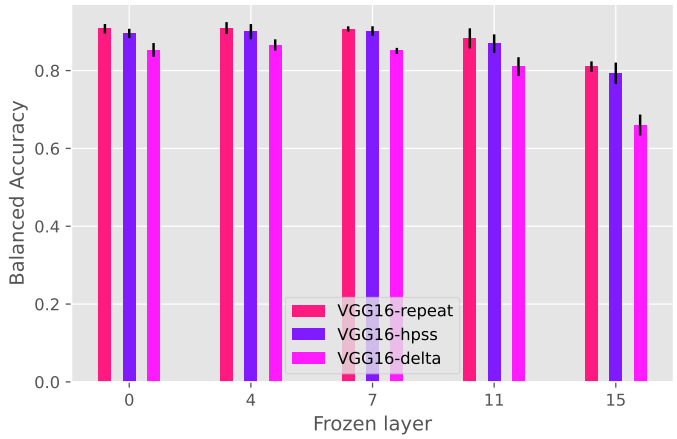


Figure 3. Comparison of three types of multi-view representation. Here, HPSS denotes harmonic–percussive source separation. Among the three representations, repeat-based spectrogram achieves the best performance.

3.3. Deep Cascade Features Using Different Pre-Trained Models

In addition to VGG16, five other pre-trained models are investigated. Based on the classification result of VGG16, rather than fine-tuning the pre-trained CNN models,

we simply choose to train on datasets from scratch without considering selecting the optimal layers for extracting deep cascade features. VGG16 performs the best in terms of balanced accuracy, with ResNet50 being the second best, and EfficientB0 being the third best (Figure 4).

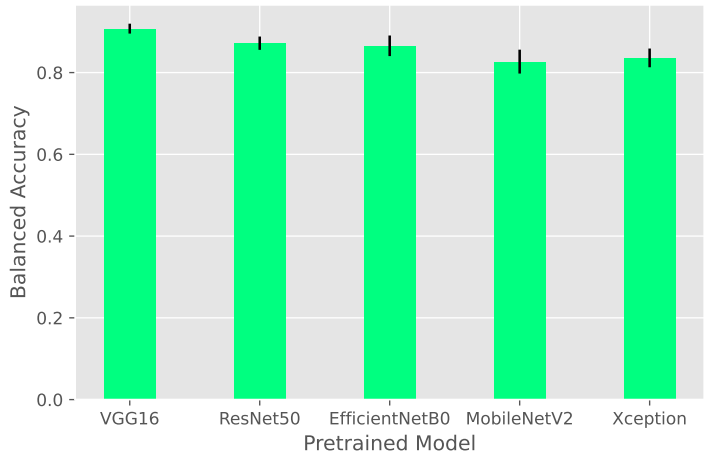


Figure 4. Comparison of different pre-trained models using the same input: repeat-based spectrogram. Here, VGG16 performs the best in terms of balanced accuracy.

3.4. Feature Fusion

To further improve the classification performance, we investigate feature fusion, where the features extracted from different layers using the selected two models are fused. Specifically, VGG16 is selected as the base model and combined with another model to better characterize bird sounds. The fusion of deep cascade features extracted from VGG16 + MobileNetV2 achieves the best performance (94.89% ± 1.35%), with VGG16 + EfficientB0 being the second best and VGG16 + ResNet50 being the third best (Figure 5). A previous study [34] has shown that VGG16 and MobileNetV2 have very different architectures and feature selection processes. Therefore, deep cascade features being extracted from VGG16 and MobileNetV2 are complementary, which explains the better performance of VGG16 + MobileNetV2.

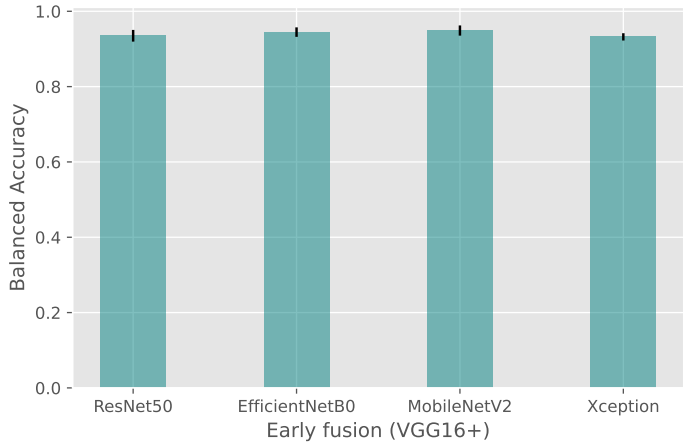


Figure 5. Comparison of four early fusion-based models. Here, VGG16 + MobileNetV2 achieves the best performance (94.89% ± 1.35%).

Finally, we visualize the confusion matrix of the best-performing model (VGG16 + MobileNetV2) (Supplementary Figure S4). We can observe that the highest confusion is between BLPW and MAWA, the second highest confusion is BTYW and PAWA, and BWWA

and MAWA is the third highest. Here, the highest confusion is different from our previous studies, which indicates the potential of improving the final performance by fusing different models. However, the computational complexity of the final framework will be higher.

4. Discussion

For bird sound classification, we use scientometrics to analyze publications and research trends. The bibliographic data are derived from the WoS core collection with all indices, which is one of the largest and most comprehensive bibliographic databases covering multidisciplinary areas. The search strategy was defined as follows: *TS=((acoustic* OR audio OR sound* OR call* OR vocalization*) AND (monitoring OR classification OR analysis OR recognition OR detection) AND (bird* OR avian*) AND ("Deep learning" OR CNN OR LSTM OR RNN OR transformer))*. In total, we obtained 111 publications, and our keyword network is shown in Supplement Figure S5 in VOSviewer [35]. The size of the circle denotes the number of publications and the color represents the time. We can obtain that most publications are related to "deep learning". Of the 111 publications, 30 papers were published in 2021, but there were only two in 2015, which indicates that this is a rapidly developing research topic. In addition to the predefined search strategy, we added "AND (transfer learning)". Only 12 publications were obtained, which further verifies the need of investigating transfer learning for bird sound classification.

For transfer learning, we use five models for the experiment: VGG16, ResNet50, EfficientB0, MobileNetV2, and Xception. However, the original weights of those pre-trained models are trained using natural images which are different from log-Mel spectrograms. This difference explains the better performance using fewer frozen layers. For five different pre-trained models, VGG16 achieves the best performance and an early fusion of VGG16 and MobileNetV2 achieves the best performance which verifies the usefulness of complementary models to be fused. We also compare our proposed method with other previous studies and the proposed model achieves higher balance accuracies than 86.31% [22] and 88.7% [36]. In the work of [22], the authors used a fusion of two end-to-end frameworks: Mel-CNN and subnet-CNN. As for [36], the authors explored handcrafted features for classification. Compared to those methods, the use of deep cascaded features achieves the best performance, which verifies their effectiveness in discriminating bird calls.

To further improve the classification performance, data augmentation is used. Specifically, we only use pitch shifting to improve performance rather than stretch and trim. One reason is that this CLO-43DS dataset is manually segmented and temporal information is well documented. Preliminary experiments indicate that the use of temporal scale augmentation such as trim does not help the classification. Meanwhile, the CLO-43DS dataset is highly imbalanced, which might reduce the performance of deep models. However, we do not apply any imbalance learning methods. One reason is that we incorporate data augmentation methods in building deep learning models since previous studies have demonstrated that data augmentation-based framework can handle class imbalance problem [37].

Several interesting questions are still open for future research. Pre-trained models are used for extracting features without optimizing the selection of layers for generating concatenated deep features. In addition, this study used manually segmented bird sounds, which might overestimate the performance when classifying bird species in continuous recordings. In addition, the start and stop time of bird sounds in a continuous recording is important for understanding bird behavior. Since we force all audio data to the fixed duration, it will artificially create the oscillation structure for all bird calls, which explains the lower performance using HPSS for creating feature representation. Bird species classification has been widely explored, but the recognition of bird individuals needs more work. Currently, all used transfer learning models are trained using natural images, which are different from spectrograms. The performance might be improved if we obtain a model which is originally trained using spectrograms. Moreover, with the use of image data of birds, it will be possible to further improve the performance.

5. Conclusions

In this study, we propose a deep cascade feature set for improved bird sound classification. The best-performing deep cascaded feature is constructed by concatenating features extracted from two different pre-trained models in different layers. As for the feature representation, simply repeated spectrogram achieves better performance using VGG16 than HPSS- and delta-based multi-view spectrograms. Comparing two types of data augmentation techniques, mixup is more suitable for extracting deep cascade features which are then classified by linear SVM to realize bird sound classification. Pitch shifting can achieve better performance for building an end-to-end classification framework. Finally, an early fusion of deep cascade features being extracted from VGG16 and MobileNetV2 can obtain the highest balanced accuracy (94.89% $\pm$ 1.35%).

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/birds4010011/s1, Table S1: The common names, scientific names, and abbreviations of 43 bird species; Figure S1. The number of instances for all bird species in the CLO-43DS dataset. Here, the x-axis denotes the abbreviations of common names of those 43 bird species, which can be found in Table S1; Figure S2. Flowchart of bird sound classification system. Here, four main steps are included which are preprocessing, multi-view representation, model training, and classification; Figure S3. The example of Setophaga occidentalis and Vermivora chrysoptera using pitch shift. (a) original Setophaga occidentalis, (b) Setophaga occidentalis using pitch shift with a factor of 2, (c) Setophaga occidentalis using pitch shift with a factor of -2. (d) original Vermivora chrysoptera, (e) Vermivora chrysoptera using pitch shift with a factor of 2, (f) Vermivora chrysoptera using pitch shift with a factor of -2; Figure S4. Confusion matrix (%) of the best result using the fusion of selected pre-trained models. Here, the x and y axes denote the code of each bird species to be classified. The x-axis is the true label and y-axis is the predicted label.

Author Contributions: Conceptualization, J.X. and M.Z.; formal analysis, J.X. and M.Z.; data curation, J.X.; writing, J.X. and M.Z.; visualization, J.X. and M.Z.; funding acquisition, J.X. and M.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Grant No: 61902154). This work was also partially supported by the Natural Science Foundation of Jiangsu Province (Grant No: BK20190579) and the Jiangsu Province Post Doctoral Fund (Grant No: 2020Z430).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study can be obtained at the following link (https://wp.nyu.edu/birdvox/codedata/).

Acknowledgments: The authors thank Kai Hu for his helpful discussions. The authors also thank anonymous reviewers for their careful work and thoughtful suggestions.

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

References

1. Zhang, J.; Huang, K.; Cottman-Fields, M.; Truskinger, A.; Roe, P.; Duan, S.; Dong, X.; Towsey, M.; Wimmer, J. Managing and analysing big audio data for environmental monitoring. In Proceedings of the 2013 IEEE 16th International Conference on Computational Science and Engineering, Sydney, Australia, 3–5 December 2013; pp. 997–1004.
2. Gage, S.H.; Wimmer, J.; Tarrant, T.; Grace, P.R. Acoustic patterns at the Samford Ecological Research Facility in South East Queensland, Australia: The Peri-Urban SuperSite of the Terrestrial Ecosystem Research Network. *Ecol. Inform.* **2017**, *38*, 62–75. [CrossRef]
3. Xie, J.; Hu, K.; Zhu, M.; Guo, Y. Bioacoustic signal classification in continuous recordings: Syllable-segmentation vs sliding-window. *Expert Syst. Appl.* **2020**, *152*, 113390. [CrossRef]
4. Wang, H.; Xu, Y.; Yu, Y.; Lin, Y.; Ran, J. An Efficient Model for a Vast Number of Bird Species Identification Based on Acoustic Features. *Animals* **2022**, *12*, 2434. [CrossRef]
5. Zhang, C.; Chen, Y.; Hao, Z.; Gao, X. An Efficient Time-Domain End-to-End Single-Channel Bird Sound Separation Network. *Animals* **2022**, *12*, 3117. [CrossRef]
6. Dawson, D.K.; Efford, M.G. Bird population density estimated from acoustic signals. *J. Appl. Ecol.* **2009**, *46*, 1201–1209. [CrossRef]

7. Pérez-Granados, C.; Traba, J. Estimating bird density using passive acoustic monitoring: A review of methods and suggestions for further research. *Ibis* **2021**, *163*, 765–783. [CrossRef]
8. Stowell, D. Computational bioacoustics with deep learning: A review and roadmap. *PeerJ* **2022**, *10*, e13152. [CrossRef]
9. Somervuo, P.; Harma, A.; Fagerlund, S. Parametric Representations of Bird Sounds for Automatic Species Recognition. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14*, 2252–2263. [CrossRef]
10. Zhang, X.; Li, Y. Adaptive energy detection for bird sound detection in complex environments. *Neurocomputing* **2015**, *155*, 108–116. [CrossRef]
11. Tuncer, T.; Akbal, E.; Dogan, S. Multileveled ternary pattern and iterative ReliefF based bird sound classification. *Appl. Acoust.* **2021**, *176*, 107866. [CrossRef]
12. Zhang, X.; Chen, A.; Zhou, G.; Zhang, Z.; Huang, X.; Qiang, X. Spectrogram-frame linear network and continuous frame sequence for bird sound classification. *Ecol. Inform.* **2019**, *54*, 101009. [CrossRef]
13. Xie, J.; Zhu, M. Handcrafted features and late fusion with deep learning for bird sound classification. *Ecol. Inform.* **2019**, *52*, 74–81. [CrossRef]
14. Xie, J.; Hu, K.; Guo, Y.; Zhu, Q.; Yu, J. On loss functions and CNNs for improved bioacoustic signal classification. *Ecol. Inform.* **2021**, *64*, 101331. [CrossRef]
15. Tan, C.; Sun, F.; Kong, T.; Yang, C.; Liu, C. A survey on deep transfer learning. In Proceedings of the Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, 4–7 October 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 270–279.
16. Sevilla, A.; Glotin, H. Audio bird classification with inception-v4 extended with time and time-frequency attention mechanisms. *CLEF (Work. Notes)* **2017**, *1866*, 1–8. [CrossRef]
17. Ntalampiras, S. Bird species identification via transfer learning from music genres. *Ecol. Inform.* **2018**, *44*, 76–81. [CrossRef]
18. Zhong, M.; LeBien, J.; Campos-Cerqueira, M.; Dodhia, R.; Ferres, J.L.; Velev, J.P.; Aide, T.M. Multispecies bioacoustic classification using transfer learning of deep convolutional neural networks with pseudo-labeling. *Appl. Acoust.* **2020**, *166*, 107375. [CrossRef]
19. Zhong, M.; Taylor, R.; Bates, N.; Christey, D.; Basnet, H.; Flippin, J.; Palkovitz, S.; Dodhia, R.; Ferres, J.L. Acoustic detection of regionally rare bird species through deep convolutional neural networks. *Ecol. Inform.* **2021**, *64*, 101333. [CrossRef]
20. Kumar, Y.; Gupta, S.; Singh, W. A novel deep transfer learning models for recognition of birds sounds in different environment. *Soft Comput.* **2022**, *26*, 1003–1023. [CrossRef]
21. Dufourq, E.; Batist, C.; Foquet, R.; Durbach, I. Passive acoustic monitoring of animal populations with transfer learning. *Ecol. Inform.* **2022**, *70*, 101688. [CrossRef]
22. Xie, J.; Hu, K.; Zhu, M.; Yu, J.; Zhu, Q. Investigation of Different CNN-Based Models for Improved Bird Sound Classification. *IEEE Access* **2019**, *7*, 175353–175361. [CrossRef]
23. Sinha, H.; Awasthi, V.; Ajmera, P.K. Audio classification using braided convolutional neural networks. *IET Signal Process.* **2020**, *14*, 448–454. [CrossRef]
24. Ruff, Z.J.; Lesmeister, D.B.; Duchac, L.S.; Padmaraju, B.K.; Sullivan, C.M. Automated identification of avian vocalizations with deep convolutional neural networks. *Remote. Sens. Ecol. Conserv.* **2019**, *6*, 79–92. [CrossRef]
25. Permana, S.D.H.; Saputra, G.; Arifitama, B.; Yaddarabullah; Caesarendra, W.; Rahim, R. Classification of bird sounds as an early warning method of forest fires using Convolutional Neural Network (CNN) algorithm. *J. King Saud Univ.-Comput. Inf. Sci.* **2022**, *34*, 4345–4357. [CrossRef]
26. Salamon, J.; Bello, J.P.; Farnsworth, A.; Robbins, M.; Keen, S.; Klinck, H.; Kelling, S. Towards the Automatic Classification of Avian Flight Calls for Bioacoustic Monitoring. *PLoS ONE* **2016**, *11*, e0166866. [CrossRef]
27. Thakur, A.; Thapar, D.; Rajan, P.; Nigam, A. Deep metric learning for bioacoustic classification: Overcoming training data scarcity using dynamic triplet loss. *J. Acoust. Soc. Am.* **2019**, *146*, 534–547. [CrossRef]
28. Salamon, J.; Bello, J.P.; Farnsworth, A.; Kelling, S. Fusing shallow and deep learning for bioacoustic bird species classification. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 141–145. [CrossRef]
29. Deng, J.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
30. Fan, R.-E.; Chang, K.-W.; Hsieh, C.-J.; Wang, X.-R.; Lin, C.-J. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.* **2008**, *9*, 1871–1874.
31. Schlüter, J.; Grill, T. Exploring data augmentation for improved singing voice detection with neural networks. In Proceedings of the 16th International Society for Music Information Retrieval Conference, Malaga, Spain, 26–30 October 2015; pp.121–126.
32. Terashima, R.; Yamamoto, R.; Song, E.; Shirahata, Y.; Yoon, H.-W.; Kim, J.-M.; Tachibana, K. Cross-Speaker Emotion Transfer for Low-Resource Text-to-Speech Using Non-Parallel Voice Conversion with Pitch-Shift Data Augmentation. *arXiv* **2022**, arXiv:2204.10020.
33. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv* **2017**, arXiv:1710.09412.
34. Sharma, A.; Singh, K.; Koundal, D. A novel fusion based convolutional neural network approach for classification of COVID-19 from chest X-ray images. *Biomed. Signal Process. Control.* **2022**, *77*, 103778. [CrossRef]
35. Van Eck, N.J.; Waltman, L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* **2010**, *84*, 523–538. [CrossRef]

36. Ji, X.; Jiang, K.; Xie, J. LBP-based bird sound classification using improved feature selection algorithm. *Int. J. Speech Technol.* **2021**, *24*, 1033–1045. [CrossRef]
37. Afzal, S.; Maqsood, M.; Nazir, F.; Khan, U.; Aadil, F.; Awan, K.M.; Mehmood, I.; Song, O.-Y. A Data Augmentation-Based Framework to Handle Class Imbalance Problem for Alzheimer's Stage Detection. *IEEE Access* **2019**, *7*, 115528–115539. [CrossRef]