Advancing Bird Species Classification: A Fusion of Audio and Image Data

Jie Xie[1], Xueyan Dong[2], Zhe Wu[1], Zheng Lang[1], Yuji Wang[1], Chunrong He[4], Jin Pei Song[4], Zhuobin Zhang[4], Guiqing Yu[4], and Jia Tang[4]

[1] School of Computer and Electronic Information /School of Artificial Intelligence, Nanjing Normal University, Nanjing, China
xiej8734@gmail.com
[2] Beijing Union University, 100101, Beijing, China
[3] Hunan Hupingshan National Nature Reserve Administration Bureau, Changde 415300, China

Abstract. Automated classification of bird species is crucial for large-scale environmental monitoring, providing valuable insights into temporal and spatial changes in ecosystems. Previous studies have primarily focused on using either acoustic or visual data for bird species recognition. However, few studies have explored the simultaneous use of both acoustic and visual data to improve classification performance. In this study, we propose a dual branch network based on pre-trained models to enhance bird species classification by integrating acoustic and visual information. Specifically, ResNet50 is used for visual data, while CNN14 is employed for acoustic data. The extracted feature embeddings are then fused, and attention mechanisms are applied to further improve classification performance. Experimental results demonstrate that our proposed model achieves significantly higher accuracy compared to using audio or image data alone. The best-performing model achieved an accuracy of 96.44%, precision of 96.62%, recall of 94.30%, and F1-score of 95.01%. This study highlights the potential of combining acoustic and visual data for bird species classification and suggests that attention mechanisms can further enhance model performance.

Keywords: Bird species classification · Fine-grained classification · Fusion of information · Transfer learning

1 Introduction

Although birds are widely recognized as excellent indicators of biodiversity due to their crucial ecosystem services, a global decline in bird populations has been observed [3]. To enhance protection policies and boost bird populations, the first step is to understand the current status of birds, which can be achieved through regular monitoring. Furthermore, large-scale monitoring of bird diversity is crucial for gaining a better understanding of bird populations. Traditional bird monitoring methods, which require ecologists to conduct censuses, are both time-consuming and expensive. Therefore, the monitoring scale is limited in both

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_10

temporal and spatial scales. Recently, the declining prices of sensor technology, coupled with the simultaneous development of artificial intelligence (AI) technology, have made it possible to monitor bird diversity in larger spatial and temporal scales. Increasingly, efforts are relying on sensors to collect data, which is then analyzed using AI techniques. Based on the analysis output, knowledge of bird diversity can be obtained.

Previous studies have proposed various approaches for bird species classification, which are either based on acoustic or visual information. For the recent work on bioacoustic signals, lots of deep learning architectures have been proposed including deep neural network [18], convolutional neural network [24, 23, 15, 4, 8, 25], recurrent neural networks [17], and transformer [27, 22, 19]. In addition, several techniques have been investigated to further improve classification performance, such as self-supervised learning [26, 6, 16] and data augmentation [21, 12, 10]. Similar to acoustic classification of bird species, deep learning methods are widely explored in classifying bird species using images [11, 2, 13, 1].

Classifying bird species by their calls or images alone is challenging due to factors like similarity between species, background noise, varying acquisition conditions, and variations in background, lighting, and capture angles in images [14]. Thus, combining acoustic and visual information may enhance classification performance. In this study, we propose a dual-branch network based on pre-trained models for bird sound classification. Specifically, ResNet50 is used as the image encoder, while CNN14 is used as the audio encoder. Then, extracted feature embeddings are fused to improve the classification performance. In addition, the attention mechanism is used to improve the classification performance. The main contribution is to deal with the optimal attention mechanism selection and combination of all aforementioned steps to improve bird species classification performance.

2 Related work

Researchers have explored the integration of acoustic and visual information for the classification of bird species. In a study by [14], the Scale-Invariant Feature Transform was employed to detect local features in bird images, which were then used to train a support vector machine classifier. When instances were not classified with sufficient certainty, they were rejected and reclassified using Mel frequency cepstral coefficients extracted from bird songs when available. Another approach introduced by [20] involved automatically identifying bird species from video recordings. This method applied image and audio processing and classification techniques, constructing models using pre-trained neural networks like ResNet50V2 and EfficientNetB0. A novel method proposed by [5] integrated visual and auditory data for species identification, improving accuracy and robustness. A deep CNN was used to extract features from bird images, while an LSTM network analyzed bird calls. By integrating these modalities early in the classification process rather than relying solely on either data type or performing late fusion, they achieved a significant performance improvement. However, the

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_10

aforementioned studies have some limitations. Firstly, relying on handcrafted features may hinder performance. Secondly, using pre-trained models originally intended for ImageNet may not be ideal for processing audio data. Lastly, all methods assume an equal number of training and testing data for each species, which is often not the case in real-world scenarios.

3 The Proposed Method

A dual branch network based on pre-trained models is proposed for bird species classification (Fig. 1). First, the issue of mismatched quantities of sound and image training data within the same category is processed to having the same quantities. Then, audio data is converted to a log-Mel spectrogram by a short-time Fourier transform. Next, two pre-trained models based on AudioSet and ImageNet are applied to acoustic and visual data separately for extracting embeddings. Finally, five different attention mechanisms are applied to audio and image embedding separately to improve the final classification performance.

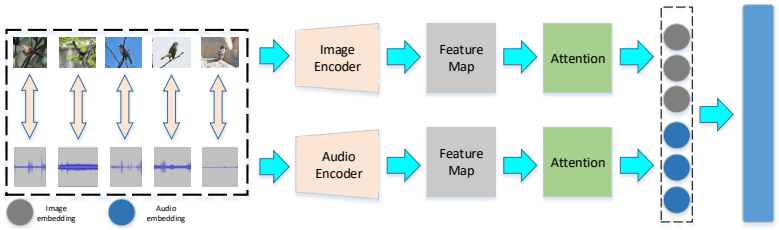

Fig. 1. Flowchart of our proposed dual branch network based on pre-trained models for bird sound classification

3.1 Datasets Description

In this study, we select 20 common bird species in Huping Moutain, Hunan Province, China as the experimental target. The audio data is collected from Xeno-canto website[4], while the images are obtained from eBird website[5]. For the image, a maximum cap of 300 images has been established for each bird species to maintain uniformity within our image dataset. Consequently, the upper limit for the number of images in any avian category is restricted to 300. For the audio, we first download the data from Xeno-canto, which is then segmented into 10 seconds with a fixed sampling rate of 16 kHz. Finally, the number of audio and images for each bird species is described in Table 1. Since we process the audio and image data simultaneously, the number of audio and image samples should be identical. For each bird species, the number of audio and image samples is first compared, and the modality having more samples will be decreased.

[4] https://xeno-canto.org/
[5] https://ebird.org/home

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_10

Table 1. Number of image and audio samples for all bird species

Category	Audio Count	Image Count	Category	Audio Count	Image Count
Black-streakedScimitarBabbler	166	248	Large-billedCrow	314	300
BrownDipper	47	300	Light-ventedBulbul	447	300
ChestnutBulbul	98	300	ManipurFulvetta	37	300
CollaredFinchbill	79	300	PygmyCupwing	350	300
CrestedKingfisher	13	300	Radde'sWarbler	318	300
Elliot'sLaughingthrush	129	300	Red-breastedFlycatcher	551	300
Fork-tailedSunbird	112	300	Rufous-facedWarbler	239	300
GreatTit	565	300	Streak-breastedScimitarBabbler	241	300
Green-backedTit	258	300	WarblingWhite-eye	326	300
Grey-headedWoodpecker	611	300	Yellow-browedBulbul	112	300

3.2 Pre-trained models

For audio classification tasks, large-scale pre-trained audio neural networks (PANNs) have been employed to extract feature embeddings [9]. PANNs represent a diverse range of convolutional neural networks designed to classify 527 distinct sound classes. Specifically, a 14-layer CNN was transferred and fine-tuned for several audio pattern recognition tasks. This CNN, pre-trained on the AudioSet dataset, has demonstrated robust generalization across numerous audio pattern recognition tasks [9]. In this study, we utilized the CNN14 architecture from [9], which comprises five blocks of 3×3 convolutional filters, followed by batch normalization and ReLU activation [9].

To extract feature embeddings from bird images, we utilized the ResNet50 model, which was pre-trained on the ImageNet dataset [7]. This pre-trained model capitalizes on its deep learning architecture to effectively capture the nuanced visual features of bird images, thereby providing a robust foundation for subsequent analysis and classification tasks.

3.3 Attention mechanism

To further enhance feature representation in deep neural networks, Convolutional Block Attention Module (CBAM), Squeeze-and-Excitation (SE), Criss-Cross Attention (CCA), Efficient Channel Attention (ECA), and Shuffle Attention (SA) mechanisms are inserted into pre-trained models for improving bird sound classification performance. In this study, the attention mechanism is inserted into the generated feature maps of both image and audio encoders separately, and then fused together for bird sound classification.

4 Experiments

4.1 Comparison of single modality and double modalities

Fig. 2 shows the performance of bird sound classification using single- and double-mode methods. From the table, we can observe that the performance of double modalities is better than both acoustic and visual modality, where the highest accuracy, precision, recall, and F1-score are 94.91%, 94.57%, 90.94%,

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_10

92.06%. For image classification, ResNet50 achieves better performance than EfficientB0. In contrast, the use of pre-trained AudioSet models (CNN14) is better than pre-trained ImageNet models (EfficientNetB0), which is often the case in previous studies [20]. One reason is that the domain mismatch between AudioSet and bird sounds is smaller than between ImageNet and bird sounds.

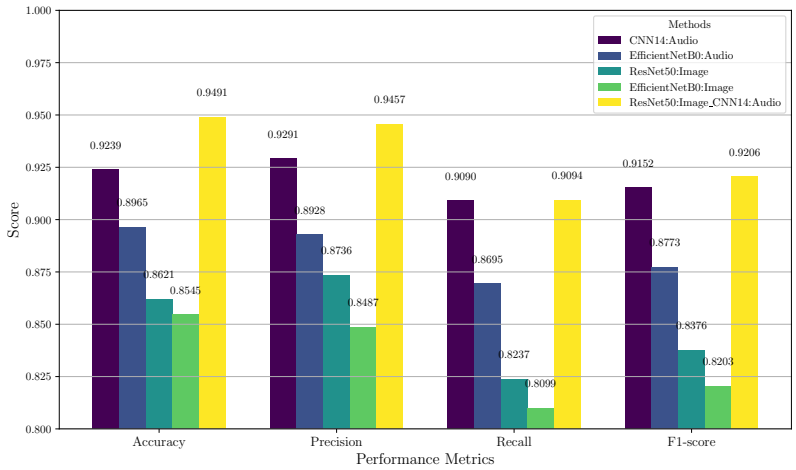

Fig. 2. Classification performance of different single modality and one double modality.

4.2 Comparison of various attention mechanisms

Since ResNet50 and CNN14 achieve the best performance for both image and audio data, they are selected and combined for subsequent analysis. To further improve the classification performance, several attention mechanisms are investigated. Table 2 shows the performance comparison of attention mechanisms. All attention mechanisms can improve classification performance, indicating the necessity of using the attention mechanism. Among these attention mechanisms, ECA achieves the best performance, where precision, precision, recall, and F1 score are 96.44%, 96.62%, 94.30%, and 95.01%, respectively. Given that we apply attention to the deep-level feature map with a small spatial size, simply employing channel-wise attention yields better performance.

Furthermore, we plot the confusion matrix of the best-performing model (Fig. 3). The highest confusion is between *Grey-headed woodpecker* and *Light-vented bulbul*. Previous studies have explored the use of audio and image data for bird sound classification, which often used a pre-trained ImageNet-based model [20]. However, the difference between nature image and spectrogram makes the pre-trained AudioSet-based model a better fit for processing audio. Compared to [20], the classification accuracy is improved by 0.38% without attention mech-

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_10

Table 2. Comparison of various attention mechanisms for our proposed dual branch network

Method	Accuracy	Precision	Recall	F1-score
ResNet50+CNN14	0.9491	0.9457	0.9094	0.9206
ResNet50+CNN14+CBAM	0.9517	0.9569	0.9272	0.9384
ResNet50+CNN14+CCA	0.9499	0.9555	0.9310	0.9393
ResNet50+CNN14+SE3	0.9550	0.9540	0.9274	0.9355
ResNet50+CNN14+ECA	**0.9644**	**0.9662**	**0.9430**	**0.9501**
ResNet50+CNN14+SU	0.9580	0.9595	0.9309	0.9411

anisms. Using ECA, the accuracy can be improved by 2.39% which verifies the effectiveness of information fusion and attention mechanisms.

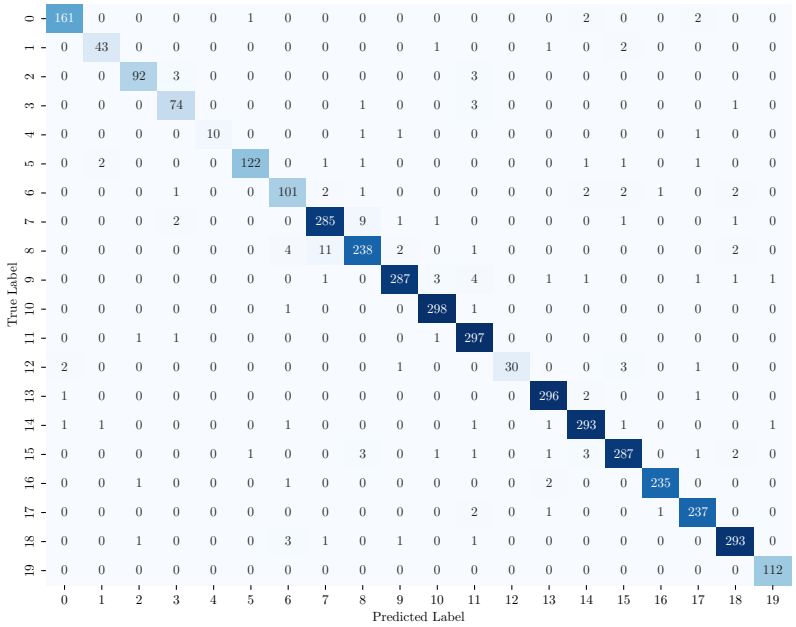

Fig. 3. Sum of confusion matrix of the best-performing model. Here, the index from 0 to 19 corresponds to the bird species from *Black streaked Scimitar Babbler* to *Yellow-browed bulbul*

5 Conclusion and Future Work

In this study, we proposed a dual branch network based on pre-trained models to improve bird species classification by integrating acoustic and visual data. The proposed method uses ResNet50 for image data and CNN14 for audio data, with

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_10

feature fusion and attention mechanisms to improve classification performance. Experimental results in the HPS dataset demonstrated that the combined use of acoustic and visual data significantly outperformed using either modality alone. The best performing model achieved an accuracy of 96. 44%, a precision of 96. 62%, a recall of 94. 30%, and a F1 score of 95. 01%. 95. 01%. Future work includes exploring advanced models, sophisticated data augmentation, generalizing to other species, and incorporating other modalities such as text or environmental data for bird species classification.

6 Acknowledgment

This work is supported by National Natural Science Foundation of China (Grant No: 32371556, 61902154 and 72004092). This work is also supported by the 2019 Science and Technology Plan of Beijing Municipal Education Commission (Grant No. KM201911417005)

References

1. Chen, T., Li, Y., Qiao, Q.: Fine-grained bird image classification based on counterfactual method of vision transformer model. The Journal of Supercomputing **80**(5), 6221–6239 (2024)
2. Chen, X., Zhang, H., Song, J., Guan, J., Li, J., He, Z.: Micro-motion classification of flying bird and rotor drones via data augmentation and modified multi-scale cnn. Remote Sensing **14**(5), 1107 (2022)
3. Fraixedas, S., Lindén, A., Piha, M., Cabeza, M., Gregory, R., Lehikoinen, A.: A state-of-the-art review on birds as indicators of biodiversity: Advances, challenges, and future directions. Ecological Indicators **118**, 106728 (2020)
4. García-Ordás, M.T., Rubio-Martín, S., Benítez-Andrades, J.A., Alaiz-Moretón, H., García-Rodríguez, I.: Multispecies bird sound recognition using a fully convolutional neural network. Applied Intelligence **53**(20), 23287–23300 (2023)
5. Gavali, P., Banu, J.S.: Visual-acoustic fusion techniques for accurate identification of indian bird species. International Journal of Computing and Digital Systems **17**(1), 1–22 (2024)
6. Hagiwara, M.: Aves: Animal vocalization encoder based on self-supervision. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
8. Hu, S., Chu, Y., Wen, Z., Zhou, G., Sun, Y., Chen, A.: Deep learning bird song recognition based on mff-scsenet. Ecological Indicators **154**, 110844 (2023)
9. Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., Plumbley, M.D.: Panns: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing **28**, 2880–2894 (2020)
10. Kumar, A.S., Schlosser, T., Kahl, S., Kowerko, D.: Improving learning-based birdsong classification by utilizing combined audio augmentation strategies. Ecological Informatics **82**, 102699 (2024)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_10

11. Kumar, M., Yadav, A.K., Kumar, M., Yadav, D.: Bird species classification from images using deep learning. In: International Conference on Computer Vision and Image Processing. pp. 388–401. Springer (2022)
12. Lauha, P., Somervuo, P., Lehikoinen, P., Geres, L., Richter, T., Seibold, S., Ovaskainen, O.: Domain-specific neural networks improve automated bird sound recognition already with small amount of local data. Methods in Ecology and Evolution **13**(12), 2799–2810 (2022)
13. Liu, H., Zhang, C., Deng, Y., Xie, B., Liu, T., Li, Y.F.: Transifc: Invariant cues-aware feature concentration learning for efficient fine-grained bird image classification. IEEE Transactions on Multimedia (2023)
14. Marini, A., Turatti, A.J., Britto, A., Koerich, A.L.: Visual and acoustic identification of bird species. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2309–2313. IEEE (2015)
15. Morales, G., Vargas, V., Espejo, D., Poblete, V., Tomasevic, J.A., Otondo, F., Navedo, J.G.: Method for passive acoustic monitoring of bird communities using umap and a deep neural network. Ecological Informatics **72**, 101909 (2022)
16. Moummad, I., Farrugia, N., Serizel, R.: Self-supervised learning for few-shot bird sound classification. In: 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW). pp. 600–604. IEEE (2024)
17. Noumida, A., Rajan, R.: Multi-label bird species classification from audio recordings using attention framework. Applied Acoustics **197**, 108901 (2022)
18. Rajan, R., Johnson, J., Abdul Kareem, N.: Bird call classification using dnn-based acoustic modelling. Circuits, Systems, and Signal Processing **41**(5), 2669–2680 (2022)
19. Rauch, L., Schwinger, R., Wirth, M., Sick, B., Tomforde, S., Scholz, C.: Active bird2vec: Towards end-to-end bird sound monitoring with transformers. arXiv preprint arXiv:2308.07121 (2023)
20. Sharma, N., Vijayeendra, A., Gopakumar, V., Patni, P., Bhat, A.: Automatic identification of bird species using audio/video processing. In: 2022 international conference for advancement in technology (ICONAT). pp. 1–6. IEEE (2022)
21. Sun, Y., Maeda, T.M., Solís-Lemus, C., Pimentel-Alarcón, D., Buřivalová, Z.: Classification of animal sounds in a hyperdiverse rainforest using convolutional neural networks with data augmentation. Ecological Indicators **145**, 109621 (2022)
22. Tang, Q., Xu, L., Zheng, B., He, C.: Transound: Hyper-head attention transformer for birds sound recognition. Ecological Informatics **75**, 102001 (2023)
23. Xiao, H., Liu, D., Chen, K., Zhu, M.: Amresnet: An automatic recognition model of bird sounds in real environment. Applied Acoustics **201**, 109121 (2022)
24. Xie, J., Zhu, M.: Sliding-window based scale-frequency map for bird sound classification using 2d-and 3d-cnn. Expert Systems with Applications **207**, 118054 (2022)
25. Xie, S., Xie, J., Zhang, J., Zhang, Y., Wang, L., Hu, H.: Mdf-net: A multi-view dual-attention fusion network for efficient bird sound classification. Applied Acoustics **225**, 110138 (2024)
26. Zhang, C., Li, Q., Zhan, H., Li, Y., Gao, X.: One-step progressive representation transfer learning for bird sound classification. Applied Acoustics **212**, 109614 (2023)
27. Zhang, S., Gao, Y., Cai, J., Yang, H., Zhao, Q., Pan, F.: A novel bird sound recognition method based on multifeature fusion and a transformer encoder. Sensors **23**(19), 8099 (2023)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_10