applied sciences

MDPI

Article

Automatic Bird Species Recognition from Images with Feature Enhancement and Contrastive Learning

Feng Yang [1,2,*], Na Shen [1,2] and Fu Xu [1,2]

[1] School of Information Science and Technology, Beijing Forestry University, Beijing 100083, China; shenna@bjfu.edu.cn (N.S.); xufu@bjfu.edu.cn (F.X.)
[2] Engineering Research Center for Forestry-Oriented Intelligent Information Processing of National Forestry and Grassland Administration, Beijing 100083, China
* Correspondence: fengyang@buaa.edu.cn

Featured Application: The research introduced in the paper develops an automatic bird species recognition system powered by feature enhancement and contrast learning, aimed at advancing ecological conservation and biological research. This system bolsters the precision of identifying bird species, aiding in the protection of endangered birds and automating the monitoring of bird populations and their migratory behaviors. It also supports in-depth behavioral and ecological research and evaluates the impact of human activities on avian life. The technology's potential applications are vast, including its use in citizen science initiatives, environmental impact assessments, educational programs, and the tourism sector, where it can provide real-time species identification, thereby enriching eco-tourism experiences and raising biodiversity awareness.

Abstract: Accurate bird species recognition is crucial for ecological conservation, wildlife monitoring, and biological research, yet it poses significant challenges due to the high variability within species and the subtle similarities between different species. This paper introduces an automatic bird species recognition method from images that leverages feature enhancement and contrast learning to address these challenges. Our method incorporates a multi-scale feature fusion module to comprehensively capture information from bird images across diverse scales and perspectives. Additionally, an attention feature enhancement module is integrated to address noise and occlusion within images, thus enhancing the model's robustness. Furthermore, employing a siamese network architecture allows effective learning of common features within instances of the same class and distinctions between different bird species. Evaluated on the CUB200-2011 dataset, our proposed method achieves state-of-the-art performance, surpassing existing methods with an accuracy of 91.3% and F1 score of 90.6%. Moreover, our approach showcases a notable advantage in scenarios with limited training data. When utilizing only 5% of the training data, our model still achieves a recognition accuracy of 65.2%, which is significantly higher than existing methods under similar data constraints. Notably, our model exhibits faster execution times compared to existing methods, rendering it suitable for real-time applications.

Keywords: bird species recognition; feature enhancement; contrastive learning; multi-scale feature

check for updates

Citation: Yang, F.; Shen, N.; Xu, F. Automatic Bird Species Recognition from Images with Feature Enhancement and Contrastive Learning. *Appl. Sci.* **2024**, *14*, 4278. https://doi.org/10.3390/app14104278

Academic Editor: Antonio Fernández-Caballero

Received: 11 April 2024
Revised: 9 May 2024
Accepted: 13 May 2024
Published: 17 May 2024

1. Introduction

Birds serve as vital indicators of ecosystem health, their populations and diversity reflecting the environmental state [1,2]. Accurate and efficient bird species recognition is crucial for conservation efforts. Traditional methods, such as expert identification [3], radar [4,5], and sound recognition [6,7], face inherent limitations. Expert identification is both time-consuming and expensive, while radar accuracy is low. Sound recognition struggles with ambient noise interference [8,9]. Advancements in sensor technology enable image-based recognition, offering advantages like reduced environmental impact, remote capture, and leveraging existing ecological data [10,11].

Deep learning revolutionizes bird species recognition by enabling machines to automatically extract and learn informative features from images, surpassing the limitations of traditional methods [12,13]. Within the realm of deep learning, two primary approaches have emerged: strongly supervised and weakly supervised classification [14,15]. Strongly supervised methods, while achieving high accuracy, demand a significant investment of time and effort in meticulously annotating training data [16,17]. These annotations typically encompass not only the bird species label but also detailed information such as bounding boxes around the bird and annotations for various body parts [18]. This labor-intensive process can become a bottleneck, especially when dealing with large datasets or rare bird species [19].

In contrast, weakly supervised methods rely solely on image-level category labels, significantly reducing annotation effort [20,21]. As a result, weakly supervised methods have become a major focus of research in fine-grained bird species recognition. Sermanet et al. [22] modeled the dynamic way humans observe objects to improve fine-grained classification performance. Zhao et al. [23] introduced a diverse visual attention network that reduces the reliance on strongly supervised information during network training. Xiao et al. [24] utilized an attention mechanism to predict separate features for both the bird and its individual parts before combining them for classification.

While image-level labels simplify data collection, their limited information necessitates both large labeled datasets and complex models for high accuracy in weakly supervised bird recognition [21]. This leads to three challenges: High computing power is a barrier for real-time and resource-constrained applications due to the complexity of models required to learn from simple labels; Large, labeled datasets for weakly supervised bird recognition are expensive and time-consuming to create, especially for complex images [19]; Visually similar birds, especially rare species, challenge weakly supervised recognition due to limitations of image-level labels. Addressing these challenges requires developing methods that effectively balance label simplicity and accuracy, making weakly supervised bird species recognition more practical and scalable [25].

We address the limitations of existing methods by proposing a novel architecture that incorporates two key modules: a multi-scale feature fusion module and a feature enhancement module. The multi-scale feature fusion module integrates features extracted at different scales within the image. By combining coarse global information with finer local details, the model gains a more comprehensive understanding of the bird's appearance. The feature enhancement module leverages an attention mechanism to dynamically adjust feature response values. This allows the model to selectively amplify informative feature channels by exploiting the relationships between them. Furthermore, the module utilizes correlations within the feature space to enhance features in the target bird region while suppressing irrelevant noise.

Finally, we propose a Siamese network [26] that employs contrastive learning to exploit the limited image-level labels for effective feature learning. This network utilizes triplet loss to minimize the distance between feature vectors belonging to the same bird species, while maximizing the distance between feature vectors of different species. This approach encourages the network to focus on extracting features that are common within a bird class while simultaneously learning to differentiate between fine-grained variations that distinguish different bird species. Through these complementary strategies, our proposed architecture aims to achieve high accuracy in weakly supervised bird species recognition.

By integrating a multi-scale feature fusion module, a feature enhancement module, and a Siamese network, we present an automated method for bird species recognition. Our approach offers several key advantages over previous works:

(1) We propose a feature fusion network that integrates features at different scales, enabling the classifier to better distinguish between similar classes.

(2) We design a feature enhancement module that selectively enhances effective feature channels and suppresses irrelevant noisy regions.

(3) We introduce a siamese network that mines the feature differences between bird species while learning common features among instances of the same class.

(4) We validate our method on the CUB200-2011 [27] bird recognition dataset and demonstrate the superior performance of our proposed model through extensive experiments.

The structure of this paper is organized as follows. In Section 2, we provide an overview of our proposed method and present a detailed description of each module. Section 3 describes the dataset used in our experiments and reports our experimental results. Extensive experiments and in-depth analyses are presented in Section 4. Finally, in Section 5, we provide our conclusions and discuss future work.

2. Materials and Methods

2.1. Dataset

To benchmark our method's effectiveness, we utilized the CUB-200-2011 (Caltech-UCSD Birds-200-2011) dataset. This widely used resource for fine-grained image classification, particularly in bird species recognition, features 11,788 images across 200 North American bird species (at least 30 images per species). Split into training (5994 images) and testing (5794 images) sets, CUB-200-2011 offers valuable annotations beyond image labels, including bounding boxes, keypoint regions, and bird attributes. This comprehensive dataset provides all the necessary attributes for fine-grained bird classification research.

2.2. The Overview of Our Proposed Method

An overview of our proposed method is presented in Figure 1, which consists of consists of three key components: feature extraction, feature enhancement, and softmax function. The feature extraction component utilizes a multi-feature fusion module to extract informative features from input bird images. This allows the model to distinguish between various bird species based on these extracted features. The feature enhancement component improves the performance of the bird classifier by enhancing the robustness and accuracy of the extracted features, thereby facilitating better differentiation between different bird species. The softmax function is used to classify bird features and output the probability of each bird species. By working together, these three modules form a comprehensive bird classification system for automated image identification.

We further propose a novel approach using contrastive learning and Siamese networks to enhance bird species classification performance. A Siamese network consists of three identical sub-networks with shared weights and architecture that process input data in parallel. Given three input images, the network generates embedding vectors for each image. We then leverage the triplet loss function [28] to optimize the Siamese network. This ensures that images from the same class have embedding vectors closer together in the embedding space, while those from different classes are farther apart. This process helps our model learn more distinctive feature representations, ultimately leading to improved classification accuracy.

To evaluate our proposed approach, we input the test dataset into our classifier model. The model then classifies the images and assigns class labels to each image. Finally, we employ various performance metrics to assess the accuracy and effectiveness of our method in classifying bird species.

2.3. Multi-Scale Feature Fusion

Bird images often exhibit significant variations in size. Traditional single-scale feature extraction approaches might miss crucial details, leading to inaccurate classification. To address this challenge, we propose a novel multi-scale feature fusion module (Figure 2) that captures both local fine-grained details and global structural information from the bird image. This comprehensive feature representation ultimately improves bird image classification performance.

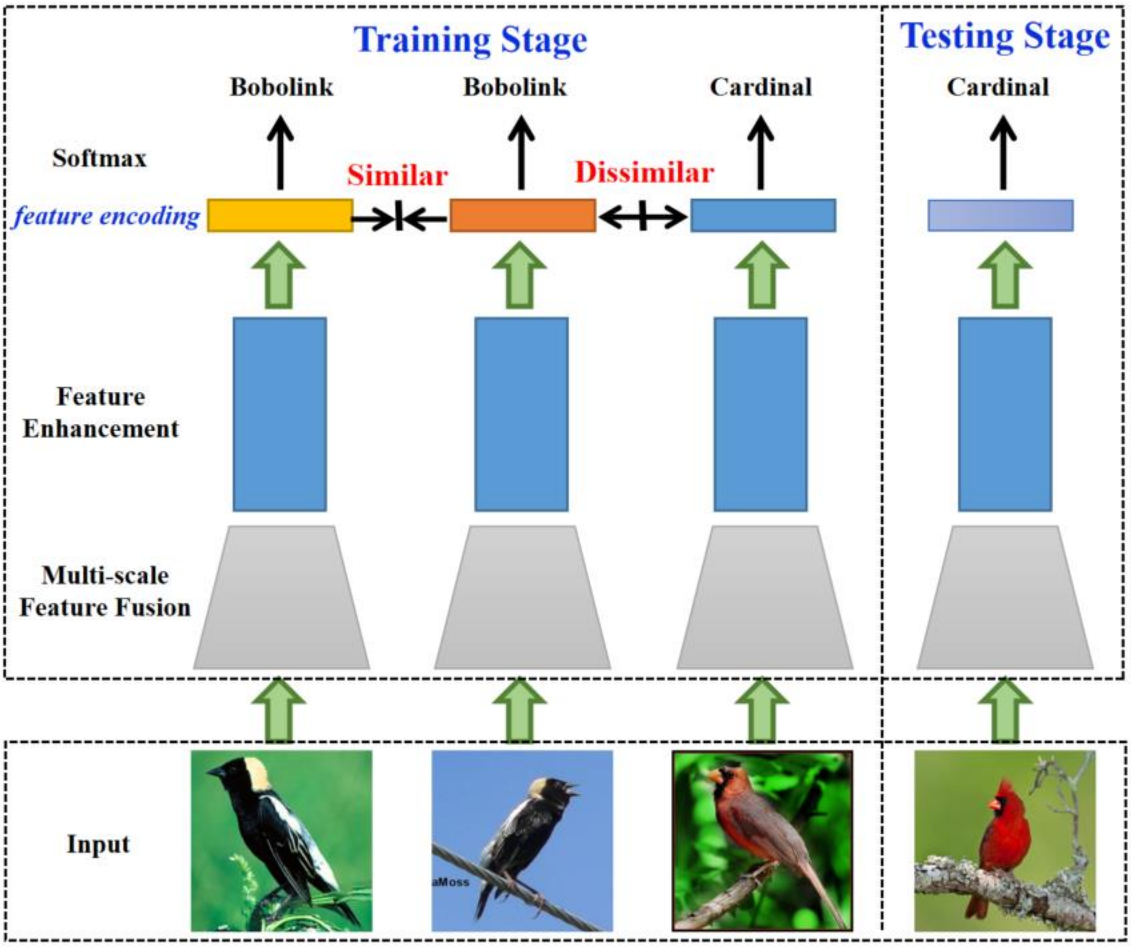

Figure 1. The overall framework of our method. It leverages a Siamese network architecture with three shared classifiers equipped with multi-scale feature extraction and feature enhancement. During training, the network learns from triplets (similar/dissimilar images) using Triplet and Softmax losses. In testing, a new bird image fed into each classifier generates a rich feature vector for classification.

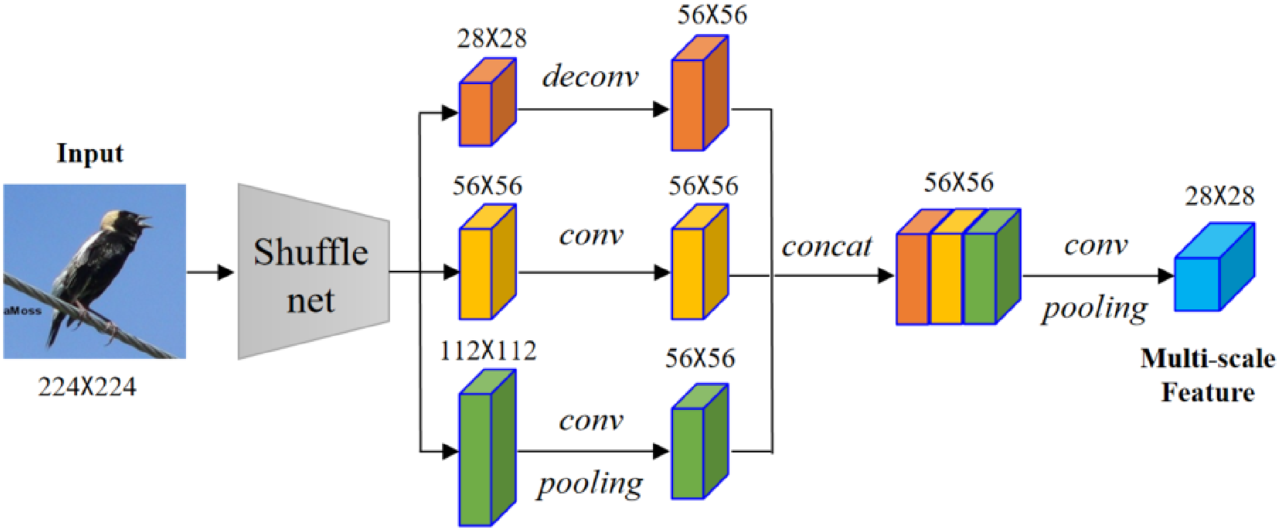

Figure 2. The framework of multi-scale feature fusion module. The module combines features from different levels of feature maps to capture both local and global information.

We leverage a pre-trained ShuffleNetV2 [29,30] model for its efficient architecture, strong feature extraction capabilities, and transfer learning potential. This efficient model excels at balancing accuracy and computational cost, making it suitable for our task. As shown in Figure 2, the ShuffleNetV2 model takes an input image of size 224 × 224 and processes it through multiple convolutional layers. The lower layer captures fine-grained details like feather textures and eye shapes. To maintain this spatial resolution for detailed information, we employ max pooling to downsample the feature map from its original size 112 × 112 to 56 × 56. The intermediate layer captures a balance between local and global information. The feature map from this layer already has a resolution of 56 × 56, so no further scaling is necessary. The higher layer captures global structural information about the bird's pose and overall body shape. To increase the resolution and incorporate more contextual information, we use a deconvolutional operation to upsample the feature map to 56 × 56.

All the processed feature maps from the three scales (each with a resolution of 56 × 56) are concatenated into a single high-dimensional feature representation, which we refer to as the "output cube". This concatenation process allows us to comprehensively capture the semantic information extracted at different resolutions (refer to Figure 2). The output cube encodes both the fine-grained details from the lower layer, local and global information from the intermediate layer, and global structural information from the higher layer, providing a richer and more informative representation for bird image classification [31].

To further reduce the computational complexity and memory footprint of the model, we apply a combination of convolution and pooling operations to the output cube, compressing its spatial dimensions from 56 × 56 to 28 × 28. The resulting compressed output, now with a size of 28 × 28, serves as the multi-scale feature representation for the bird image classification task. This compressed representation retains the essential information extracted from different scales while reducing the computational burden, making the model more efficient for both training and inference [32].

2.4. Feature Enhancement

The feature enhancement module is designed to enhance the object cues and weaken the background noise, as shown in Figure 3. The enhancement module first divides the given feature map into groups along the channel dimension, infers channel and spatial attention maps for each sub-feature in parallel, and then generates refined sub-feature by multiplying the attention map with the input feature. Finally, all the refined sub-features are concatenated as the new feature map for fine-grained classification.

The feature enhancement module first divides the feature map F into m groups along the channel dimension, where each sub-feature F_k captures a specific semantic response during training. Each sub-feature F_k is then split into two branches, as shown in Figure 3. One branch is used to create a channel attention map by leveraging the inter-relationship of channels, while the other branch generates a spatial attention map by utilizing the inter-spatial relationship of features.

By applying global average pooling operation [33] to the feature map F_{k1} for feature compression, a feature vector containing global information is obtained. Then, the vector is passed through a fully connected layer and a sigmoid activation function to calculate the channel weight coefficients. Finally, the channel weight coefficients are multiplied with the feature map F_{k1} to generate a feature map A_{k1} embedded with channel attention mechanism [34].

$$A_{k1} = \sigma(W_1 \cdot GP(F_{k1}) + b_1) \cdot F_{k1} \tag{1}$$

where W_1, b_1 represent the parameters to be learned, GP represents global pooling operation, and $\sigma(\cdot)$ represents the sigmoid activation function.

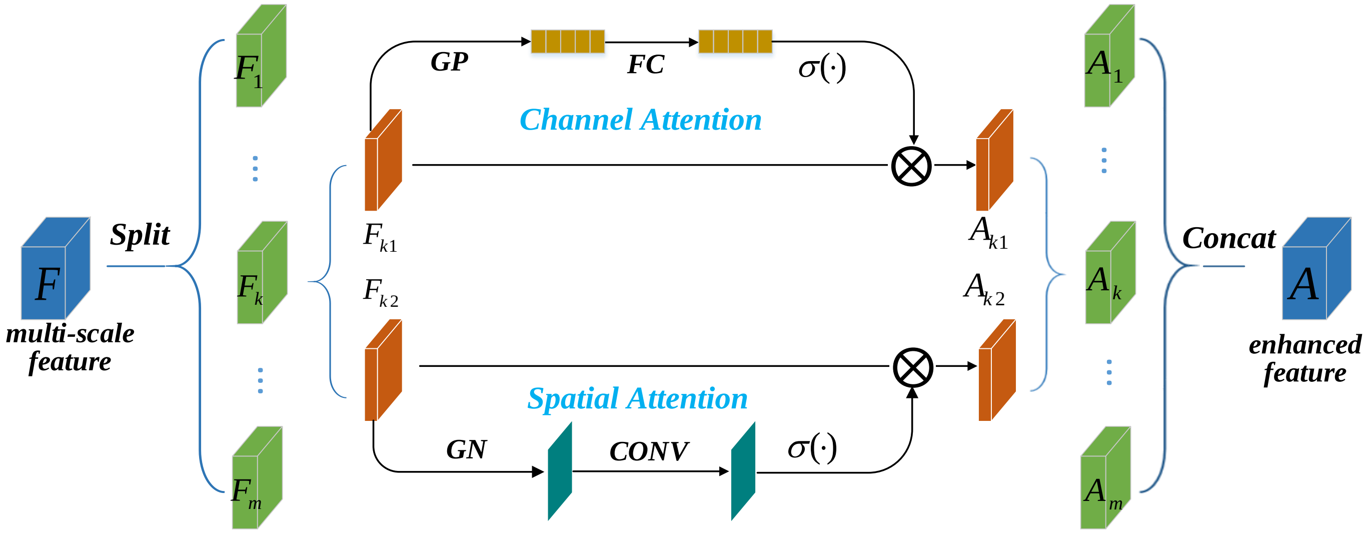

Figure 3. The framework of feature enhancement module. It splits channels and processes sub-features concurrently. Two branches extract informative features: channel attention focuses on important channels using GAP, while spatial attention captures spatial dependencies with group normalization. Both branches are concatenated, followed by sub-feature aggregation and a channel shuffle operation to promote information exchange. This enhances the overall feature representation for tasks like fine-grained bird recognition.

First, the feature map F_{k2} is subjected to group normalization operation. Then, the calculated result is passed through a convolutional layer and a sigmoid function to calculate the spatial weight coefficients. Finally, the spatial weight coefficients are multiplied with the feature map F_{k2} to generate the feature map A_{k2} with embedded spatial attention mechanism [35].

$$A_{k2} = \sigma(W_2 \cdot GN(F_{k2}) + b_2) \cdot F_{k2} \tag{2}$$

where W_2, b_2 represent the parameters to be learned, and $GN(\cdot)$ represents group normalization operation.

Merge the outputs A_{k1}, A_{k2} of two branches, concatenate all grouped features A_k together to generate a feature map A embedded with attention mechanism. Compared to input features, the output features are enhanced through the attention mechanism.

2.5. Siamese Network

Siamese network is proposed to learn similarity or distance between inputs that consists of three identical subnetworks (sharing weights) which process three inputs in parallel. Each subnetwork employs the same architecture and parameters, and processes the three inputs separately during forward propagation to obtain corresponding feature vectors [36,37]. These feature vectors can then be passed into a metric learning layer to compute the similarity or distance between the three inputs.

Training a siamese network involves using a triplet loss and softmax loss. Triplet loss is a loss function used for learning similarity or distance metrics, which uses triplets to define the relationship between samples. Each triplet consists of an anchor sample, a positive sample, and a negative sample, where the positive sample is similar to the anchor sample and the negative sample is dissimilar to the anchor sample.

The goal of triplet loss is to minimize the distance between the anchor sample and the positive sample, while maximizing the distance between the anchor sample and the negative sample, such that the distance between the anchor sample and the positive sample is less than the distance between the anchor sample and the negative sample. This can encourage the model to learn an embedding space where similar samples are closer together and dissimilar samples are farther apart.

$$L_{tp} = max(0, d(f(a), f(p)) - d(f(a), f(n)) + margin) \tag{3}$$

where $f(a)$, $f(p)$, and $f(n)$ represent the feature vectors of the anchor sample, positive sample, and negative sample in the embedding space, respectively, $d(u,v)$ represents the distance between two feature vectors, and margin is a hyperparameter used to control the difference between the distances of the positive and negative samples.

Softmax loss is used to measure the difference between the predicted probability distribution and the true probability distribution of the classes. The softmax loss function penalizes the model when the predicted probability for the correct class is low and rewards the model when the predicted probability for the correct class is high. The softmax loss function can then be defined as:

$$L_{st} = -\frac{1}{N}\sum_{j=1}^{K} y_{i,j} logP(y_{i,j} = 1) \tag{4}$$

where $y_{i,j}$ is a binary indicator (0 or 1) whether the j-th class is the correct classification for the i-th training sample, and $P\left(y_{i,j} = 1\right)$ is the predicted probability of the j-th class for the i-th training sample.

The total loss is a weighted sum of the triplet and softmax losses, where the weights are determined by the relative importance of each component to the overall task.

$$L = L_{st} + \lambda L_{tp} \tag{5}$$

During training, a batch of randomly selected triplets are used, and their triplet and softmax losses is computed. The goal is to minimize the loss function by adjusting the weights and biases of the neural network using backpropagation. This process updates the parameters of the network until the loss function is minimized, and the network can accurately classify the input data.

3. Results

3.1. Implementation Details

To accelerate convergence, we initialized our model with weights pre-trained on the ImageNet dataset. During training, we employed the Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9 and weight decay of 0.0001. SGD is a widely used and efficient optimizer, while momentum (0.9) aids in smoother convergence by dampening fluctuations in gradient updates. Weight decay (0.0001) introduces an L2 penalty to prevent overfitting. Additionally, a learning rate scheduler that reduces the learning rate by a factor of 10 at specific epochs (e.g., 60th and 100th) was employed. The entire training process ran for 120 epochs using the PyTorch framework on a Tesla T4 GPU.

3.2. Evaluation Metrics

We employed two key metrics to quantitatively assess our model's performance: accuracy and F1 score. This combination provides a comprehensive analysis of the model's effectiveness.

Accuracy reflects the proportion of images correctly classified by the model out of the total number of images in the dataset. The F1 score considers both precision (proportion of true positives among predicted positives) and recall (proportion of true positives among actual positives) to provide a more balanced view of the model's performance.

Additionally, we consider running time as a metric to assess speed. Running time refers to the average time taken by the classification model to process individual input images and generate classification outcomes. It's a critical performance metric, especially for real-time applications where efficiency is crucial.

3.3. Ablation Experiments

To assess the contribution of each core module in our bird species recognition model, we performed ablation experiments. This involved removing each module individually and retraining the modified model. Table 1 summarizes the results.

Table 1. Ablation experiments. We evaluated four model variants: (1) baseline, which includes all three modules; (2) without multi-scale feature fusion; (3) without feature enhancement; and (4) without siamese network training. For each variant, we report the accuracy and F1 score on the test set.

Models	Accuracy (%)	F1 Score
Proposed Model (All Modules)	91.3	90.6
w/o multi-scale feature fusion	87.4	87.1
w/o feature enhancement	85.2	84.8
w/o siamese network training	86.9	86.2

We first conducted an experiment to assess the impact of multi-scale feature fusion by removing this module from the proposed model and training the modified model from scratch. The results showed that the accuracy decreased from 91.3% to 89.6% without this module, indicating the crucial role of capturing multi-scale information in bird images. Additionally, we observed a decline in F1 score from 0.906 to 0.871, further highlighting the importance of multi-scale feature fusion in the proposed model.

Next, we evaluated the effectiveness of attention feature enhancement by removing this module and training the modified model from scratch. As shown in Table 1, the removal of attention feature enhancement resulted in a decrease in both accuracy, from 91.3% to 85.2%, and F1 score, from 0.906 to 0.848. These results indicate that the feature enhancement module enhances important features for bird classification.

Finally, we conducted an experiment to evaluate the importance of siamese network training by training a modified model from scratch without this module. The results in Table 1 indicate a significant decrease in both accuracy, from 91.3% to 86.9%, and F1 score, from 0.906 to 0.862, upon removing the siamese network training module. These findings underscore the importance of exploring the interaction between features and optimizing the model during training through the use of the siamese network.

3.4. State-of-the-Art Comparison

To validate the efficacy of our proposed model, we compared it with established fine-grained classification approaches, encompassing both strongly supervised (HSnet [38], Mask-CNN [39]) and weakly supervised (RA-CNN [40], ADKD [41]) techniques.

We evaluated the models' performance on a benchmark dataset. As shown in Table 2, our model achieves a superior accuracy of 91.3%, surpassing HSnet, Mask-CNN, RA-CNN, and ADKD_T by 3.8%, 4.0%, 5.9%, and 1.8%, respectively. This demonstrates the effectiveness of our multi-scale feature fusion, attentional feature enhancement, and contrastive learning strategies in bird classification.

Table 2. Comparison of the accuracy results of different models on the CUB-200-2011 dataset.

Models	Backbone Network	Accuracy (%)
HSnet [31]	GoogLeNet	87.5
Mask-CNN [36]	VGG-16	87.3
RA-CNN [40]	VGG-19	85.4
ADKD_T [41]	Densenet121	89.5
FS-NET(Proposed Model)	ShuffleNetV2	91.3

Beyond classification accuracy, we compared model size and execution time. Our model utilizes a ShuffleNetV2 backbone, resulting in a compact size of approximately 193 MB (Table 3). This is significantly smaller than the compared models: 20% of HSnet, 11% of Mask-CNN, 1.8% of RA-CNN, and 74% of ADKD_T. Furthermore, our model exhibits faster execution times compared to all benchmarks. It achieves speedups of 7×, 14×, 8×, and 4.8× against HSnet, Mask-CNN, RA-CNN, and ADKD_T, respectively. This efficiency makes our model suitable for real-time applications.

Table 3. The comparison of the model size and running time of the model.

Models	Model Size	Params	Running Time
HSnet [31]	293.85 MB	24.62 MB	1.64 s
Mask-CNN [36]	634.53 MB	48.74 MB	3.47 s
RA-CNN [40]	3.06 GB	265.94 M	1.79 s
ADKD_T [41]	1.32 GB	6.95 M	1.1 s
FS-NET(Proposed Model)	193.58 MB	5.17 M	234 ms

We further investigated the proposed method's performance under conditions with limited training data. The experimental results depicted in Figure 4 demonstrate a clear advantage for our model. When the training data is scarce (only 5.0%), our method achieves a recognition accuracy of 65.2%, which is about 10% higher than the second-best method. This advantage persists even as the training data volume increases. With 25.0% training data, our model still maintains a lead of approximately 5% over the second-best performer.

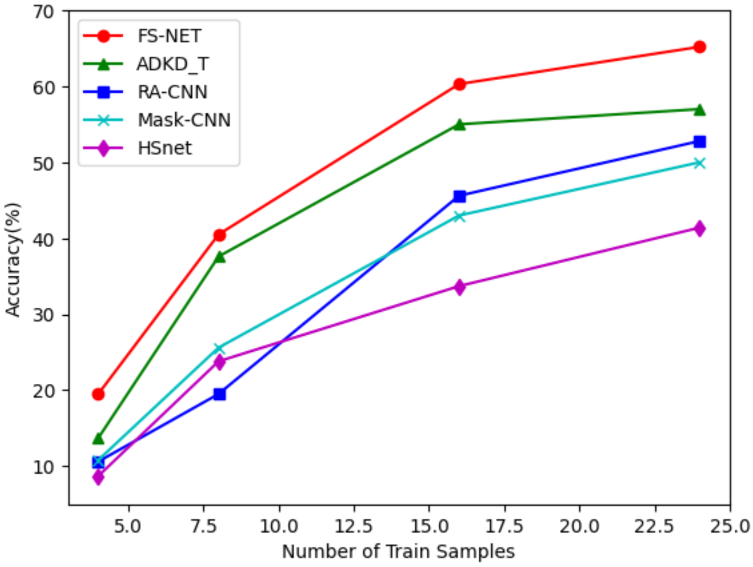

Figure 4. Comparison of bird species recognition accuracy under number of train samples.

4. Discussion

Following our experimentation, the incorporation of the three core modules—multi-scale feature fusion, feature enhancement, and Siamese network training—significantly improves the model's performance. Evaluation results unequivocally demonstrate that our model outperforms the compared methods in both accuracy and F1 score, underscoring the effectiveness of our approach in fine-grained bird classification. Notably, the multi-scale feature fusion module enables the model to capture comprehensive information from bird images across various scales and perspectives, thereby enhancing classification accuracy by providing a more nuanced feature representation. The feature enhancement module plays a crucial role in mitigating the impact of noise and occlusion within images, thereby strengthening the model's overall robustness. Additionally, the Siamese network training module effectively learns feature interactions and optimizes the model during training, resulting in a better balance between precision and recall.

In addition to superior classification accuracy and F1 score, the experimental results also highlight that our proposed model exhibits faster execution times compared to HSnet, Mask-CNN, RA-CNN, and ADKD_T. This efficiency is primarily attributed to the utilization of the ShuffleNetV2 backbone, renowned for its computational efficiency. Moreover, our model achieves faster execution times compared to all benchmarks, rendering it suitable for real-time applications where speed is paramount.

Furthermore, our proposed model achieves significantly better recognition accuracy than other methods, particularly when trained with limited data (as depicted in Figure 4). Even with only 5.0% of the training data, our method achieves a remarkable 10% higher recognition accuracy compared to the second-best performing model. This suggests that our approach is exceptionally efficient in learning effective feature representations even with a scarcity of training examples. Moreover, as the volume of training data increases, our model maintains its performance lead, indicating its superior utilization of available data and consistent performance across varying data volumes. This robustness underscores the scalability and effectiveness of our model in real-world applications.

In summary, our proposed model featuring three modules has demonstrated superior performance in terms of accuracy, F1 score, and efficiency, underscoring its efficacy in fine-grained bird classification. The research outcomes presented in this study carry substantial implications and opportunities for advancement across various domains. Accurately classifying bird species holds promise for driving progress in wildlife conservation endeavors, biodiversity research, and ecosystem health assessments. Researchers, ornithologists, and environmentalists can leverage these findings for studying bird behavior, ecology, and evolution. Additionally, industries involved in construction and tourism stand to benefit from informed decision-making enabled by precise species identification.

However, the current study highlights potential concerns, including data bias and class imbalance within the training dataset, which could affect performance, particularly for underrepresented species. Generalizing to unseen scenarios and optimizing default hyperparameter values are also noteworthy considerations. Moreover, enhancing the interpretability of the model's decision-making process remains a significant challenge.

Looking ahead, optimization efforts will involve broader evaluation metrics, improved robustness to environmental variations, sustainable integration into real-world applications, and collaboration with domain experts. These refinements aim to ensure broader applicability, reliability, and a deeper comprehension of the model's classifications.

5. Conclusions

Our work introduces a groundbreaking bird species recognition method, achieving state-of-the-art accuracy of 91.3% and F1 score of 90.6% on the CUB200-2011 dataset. Our approach incorporates multi-scale feature fusion, attention feature enhancement, and a Siamese network architecture to effectively capture comprehensive information from bird images and enhance robustness to noise and occlusion. Additionally, our method demonstrates a notable advantage in scenarios with limited training data, achieving a recognition accuracy of 65.2% with only 5% of the training data. This research highlights the importance of multi-scale features and attention enhancement for fine-grained classification tasks and opens doors for further exploration of efficient deep learning model integration strategies in bird classification and potentially other related fields.

Moving forward, our research trajectory will focus on exploring more efficient and effective strategies to incorporate attention mechanisms and siamese networks into deep learning models tailored for fine-grained classification tasks. This endeavor aims to further advance the capabilities of our proposed method and enhance its applicability across various domains, ultimately contributing to the broader advancement of fine-grained classification techniques.

Author Contributions: F.Y. conducted literature research, method conceptualization and implementation, designed experiments, and wrote the manuscript. N.S. and F.X. revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Key R&D Program of China (2022YFF1302700), The Emergency Open Competition Project of National Forestry and Grassland Administration (202303), Outstanding Youth Team Project of Central Universities (QNTD202308).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study is derived from the following publicly available resources: Caltech-UCSD Birds-200-2011 https://vision.cornell.edu/se3/caltech-ucsd-birds-200/ (accessed on 10 May 2024).

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

References

1. Charmantier, A.; Gienapp, P. Climate change and timing of avian breeding and migration: Evolutionary versus plastic changes. *Evol. Appl.* **2014**, *7*, 15–28. [CrossRef]
2. Acevedo, M.A.; Corrada-Bravo, C.J.; Corrada-Bravo, H.; Villanueva-Rivera, L.J.; Aide, T.M. Automated classification of bird and amphibian calls using machine learning: A comparison of methods. *Ecol. Inform.* **2009**, *4*, 206–214. [CrossRef]
3. Saito, T.; Kanezaki, A.; Harada, T. IBC127: Video dataset for fine-grained bird classification. In Proceedings of the 2016 IEEE International Conference on Multimedia and Expo (ICME), Seattle, WA, USA, 11–15 July 2016; pp. 1–6.
4. Kahl, S.; Clapp, M.; Hopping, W.A.; Goëau, H.; Glotin, H.; Planqué, R.; Vellinga, W.P.; Joly, A. Overview of birdclef 2020: Bird sound recognition in complex acoustic environments. In Proceedings of the CLEF 2020-Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, 22–25 September 2020; Volume 2696.
5. Kahl, S.; Wood, C.M.; Eibl, M.; Klinck, H. BirdNET: A deep learning solution for avian diversity monitoring. *Ecol. Inform.* **2021**, *61*, 101236. [CrossRef]
6. Jasim, H.A.; Ahmed, S.R.; Ibrahim, A.A.; Duru, A.D. Classify Bird Species Audio by Augment Convolutional Neural Network. In Proceedings of the 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications(HORA), Ankara, Turkey, 9–11 June 2022; pp. 1–6.
7. 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]
8. Bardeli, R.; Wolff, D.; Kurth, F.; Koch, M.; Tauchert, K.H.; Frommolt, K.H. Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. *Pattern Recognit. Lett.* **2010**, *31*, 1524–1534. [CrossRef]
9. Varghese, A.; Shyamkrishna, K.; Rajeswari, M. Utilization of deep learning technology in recognizing bird species. *AIP Conf. Proc.* **2022**, *2463*, 020035.
10. Yang, Z.; Luo, T.; Wang, D.; Hu, Z.; Gao, J.; Wang, L. Learning to navigate for fine-grained classification. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 420–435.
11. Theivaprakasham, H.; Sowmya, V.; Ravi, V.; Gopalakrishnan, E.; Soman, K. Hybrid Features-Based Ensembled Residual Convolutional Neural Network for Bird Acoustic Identification. In *Advances in Communication, Devices and Networking: Proceedings of ICCDN 2021*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 437–445.
12. Yang, S.; Bo, L.; Wang, J.; Shapiro, L. Unsupervised template learning for fine-grained object recognition. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1–9.
13. Wang, Y.; Wang, Z. A survey of recent work on fine-grained image classification techniques. *J. Vis. Commun. Image Represent.* **2019**, *59*, 210–214. [CrossRef]
14. Tanzi, L.; Vezzetti, E.; Moreno, R.; Moos, S. X-ray bone fracture classification using deep learning: A baseline for designing a reliable approach. *Appl. Sci.* **2020**, *10*, 1507. [CrossRef]
15. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
16. Xu, X.; Yang, C.C.; Xiao, Y.; Kong, J.L. A fine-grained recognition neural network with high-order feature maps via graph-based embedding for natural bird diversity conservation. *Int. J. Environ. Res. Public Health* **2023**, *20*, 4924. [CrossRef] [PubMed]
17. Ji, R.; Li, J.; Zhang, L. Siamese self-supervised learning for fine-grained visual classification. *Comput. Vis. Image Underst.* **2023**, *229*, 103658. [CrossRef]
18. Cai, Q.; Niu, L.; Shang, X.; Ding, H. A Self-Supervised Tree-Structured Framework for Fine-Grained Classification. *Appl. Sci.* **2023**, *13*, 4453. [CrossRef]
19. Zheng, F.; Cao, J.; Yu, W.; Chen, Z.; Xiao, N.; Lu, Y. Exploring low-resource medical image classification with weakly supervised prompt learning. *Pattern Recognit.* **2024**, *149*, 110250. [CrossRef]
20. Lv, Y.; Zhang, J.; Barnes, N.; Dai, Y. Weakly-supervised contrastive learning for unsupervised object discovery. *IEEE Trans. Image Process.* **2024**, *33*, 2689–2702. [CrossRef]
21. Dai, Y.; Song, W.; Gao, Z.; Fang, L. Global-guided weakly-supervised learning for multi-label image classification. *J. Vis. Commun. Image Represent.* **2023**, *93*, 103823. [CrossRef]
22. Sermanet, P.; Frome, A.; Real, E. Attention for fine-grained categorization. *arXiv* **2014**, arXiv:1412.7054.

23. Zhao, B.; Wu, X.; Feng, J.; Peng, Q.; Yan, S. Diversified visual attention networks for fine-grained object classification. *IEEE Trans. Multimed.* **2017**, *19*, 1245–1256. [CrossRef]
24. Xiao, T.; Xu, Y.; Yang, K.; Zhang, J.; Peng, Y.; Zhang, Z. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 842–850.
25. Ren, Z.; Wang, S.; Zhang, Y. Weakly supervised machine learning. *CAAI Trans. Intell. Technol.* **2023**, *8*, 549–580. [CrossRef]
26. He, A.; Luo, C.; Tian, X.; Zeng, W. A twofold siamese network for real-time object tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4834–4843.
27. Wah, C.; Branson, S.; Welinder, P.; Perona, P.; Belongie, S. *The Caltech-Ucsd Birds-200-2011 Dataset 2011*; California Institute of Technology: Pasadena, CA, USA, 2011.
28. Ghosh, A.; Shanmugalingam, K.; Lin, W.Y. Relation preserving triplet mining for stabilising the triplet loss in re-identification systems. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 4840–4849.
29. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 116–131.
30. Sheng, G.; Min, W.; Yao, T.; Song, J.; Yang, Y.; Wang, L.; Jiang, S. Lightweight Food Image Recognition with Global Shuffle Convolution. *IEEE Trans. AgriFood Electron.* **2024**. *early access*. [CrossRef]
31. Huo, X.; Sun, G.; Tian, S.; Wang, Y.; Yu, L.; Long, J.; Zhang, W.; Li, A. HiFuse: Hierarchical multi-scale feature fusion network for medical image classification. *Biomed. Signal Process. Control* **2024**, *87*, 105534. [CrossRef]
32. Jiang, L.; Yuan, B.; Du, J.; Chen, B.; Xie, H.; Tian, J.; Yuan, Z. MFFSODNet: Multi-Scale Feature Fusion Small Object Detection Network for UAV Aerial Images. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 5015214. [CrossRef]
33. Zhang, Q.L.; Yang, Y.B. Sa-net: Shuffle attention for deep convolutional neural networks. In Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 2235–2239.
34. Wang, Y.; Guo, S.; Guo, J.; Zhang, J.; Zhang, W.; Yan, C.; Zhang, Y. Towards performance-maximizing neural network pruning via global channel attention. *Neural Netw.* **2024**, *171*, 104–113. [CrossRef] [PubMed]
35. Zhang, Y.; Liang, J.; Niu, P.; Xu, W. Center-similarity spectral-spatial attention network for hyperspectral image classification. *J. Appl. Remote Sens.* **2024**, *18*, 016509. [CrossRef]
36. Valero-Mas, J.J.; Gallego, A.J.; Rico-Juan, J.R. An overview of ensemble and feature learning in few-shot image classification using siamese networks. *Multimed. Tools Appl.* **2024**, *83*, 19929–19952. [CrossRef]
37. Fedele, A.; Guidotti, R.; Pedreschi, D. Explaining Siamese networks in few-shot learning. *Mach. Learn.* **2024**, 1–38. [CrossRef]
38. Lam, M.; Mahasseni, B.; Todorovic, S. Fine-grained recognition as hsnet search for informative image parts. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2520–2529.
39. Wei, X.S.; Xie, C.W.; Wu, J.; Shen, C. Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization. *Pattern Recognit.* **2018**, *76*, 704–714. [CrossRef]
40. Fu, J.; Zheng, H.; Mei, T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4438–4446.
41. Wang, K.; Yang, F.; Chen, Z.; Chen, Y.; Zhang, Y. A Fine-Grained Bird Classification Method Based on Attention and Decoupled Knowledge Distillation. *Animals* **2023**, *13*, 264. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.