

Article

A Multi-Scale Feature Fusion Hybrid Convolution Attention Model for Birdsong Recognition

Lianglian Gu ¹, Guangzhi Di ^{2,*}, Danju Lv ^{1,*}, Yan Zhang ³, Yueyun Yu ¹, Wei Li ¹ and Ziqian Wang ¹

¹ School of Big Data and Intelligent Engineering, Southwest Forestry University, Kunming 650224, China; gulianglian@swfu.edu.cn (L.G.); yuyueyun103@swfu.edu.cn (Y.Y.); liwei8152@swfu.edu.cn (W.L.); wangziqian@swfu.edu.cn (Z.W.)

² Southwest Forestry University, Kunming 650224, China

³ School of Science, Southwest Forestry University, Kunming 650224, China; zhangyan@swfu.edu.cn

* Correspondence: swfcdgz@swfu.edu.cn (G.D.); lvdanjv@swfu.edu.cn (D.L.)

Abstract: Birdsong is a valuable indicator of rich biodiversity and ecological significance. Although feature extraction has demonstrated satisfactory performance in classification, single-scale feature extraction methods may not fully capture the complexity of birdsong, potentially leading to suboptimal classification outcomes. The integration of multi-scale feature extraction and fusion enables the model to better handle scale variations, thereby enhancing its adaptability across different scales. To address this issue, we propose a multi-scale hybrid convolutional attention mechanism model (MUSCA). This method combines depthwise separable convolution and traditional convolution for feature extraction and incorporates self-attention and spatial attention mechanisms to refine spatial and channel features, thereby improving the effectiveness of multi-scale feature extraction. To further enhance multi-scale feature fusion, a layer-by-layer alignment feature fusion method is developed to establish a deeper correlation, thereby improving classification accuracy and robustness. Using the above method, we identified 20 bird species on three spectrograms, wavelet spectrogram, log-Mel spectrogram and log-spectrogram, with recognition rates of 93.79%, 96.97% and 95.44%, respectively. Compared with the resnet18 model, it increased by 3.26%, 1.88% and 3.09%, respectively. The results indicate that the MUSCA method proposed in this paper is competitive compared to recent and state-of-the-art methods.

Keywords: birdsong recognition; hybrid convolution attention mechanism; multi-scale feature extraction; deep learning



Academic Editor: Christos Bouras

Received: 12 March 2025

Revised: 17 April 2025

Accepted: 17 April 2025

Published: 21 April 2025

Citation: Gu, L.; Di, G.; Lv, D.; Zhang, Y.; Yu, Y.; Li, W.; Wang, Z. A Multi-Scale Feature Fusion Hybrid Convolution Attention Model for Birdsong Recognition. *Appl. Sci.* **2025**, *15*, 4595. <https://doi.org/10.3390/app15084595>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Birds play an important role in ecosystems as key biological indicators and are closely linked to humans [1]. By analyzing bird calls, researchers can gain insights into bird behavior, population dynamics and ecosystem health, which are crucial for the conservation and management of ecosystems [2]. Image recognition technology has made significant progress in bird research, such as Anusha and ManiSai [3], who used deep convolutional neural networks (DCNNs) to classify bird images based on high-dimensional representations. However, collecting such data in a complex forest environment is still quite challenging. In contrast, bird vocalization-based recognition methods provide solutions to these challenges with their penetrability, efficiency, non-invasiveness and wide applicability [4]. This approach has become indispensable in ecological research and biodiversity monitoring, significantly increasing the efficiency and coverage of data collection.

Developing methods for birdsong recognition usually involves four steps [3]: sample collection, signal preprocessing, feature extraction and fusion, model recognition result output. Among these, feature extraction plays a pivotal role in determining the model's performance. Recent studies emphasize the need for selecting appropriate feature representations tailored to birdsong characteristics in order to achieve reliable classification results [5]. A common approach involves converting raw audio into image-like spectrogram representations, which are then fed into neural networks to capture temporal and spectral textures. Widely used spectral features include log-Mel spectrograms, Mel-Frequency Cepstral Coefficients (MFCCs), Gammatone Cepstral Coefficients (GTCCs), wavelet transforms (WTs), Hilbert–Huang Transforms (HHTs), Short-Time Fourier Transforms (STFTs) and Linear Predictive Cepstral Coefficients (LPCCs) [6].

Early research mainly relied on traditional feature extraction and classification methods. Manual feature extraction, such as Mel-Frequency Cepstral Coefficient (MFCC) and Linear Predictive Coding (LPC), is then combined with classical machine learning classifiers, such as Support Vector Machine (SVM) and Hidden Markov Model (HMM), for birdsong recognition and detection [7]. However, these approaches often fail to capture complex patterns and contextual information, limiting model performance and generalization. With the advent of deep learning, automated high-level feature extraction has significantly enhanced recognition performance while reducing the need for manual preprocessing. For instance, Koh et al. applied an Inception-v3 architecture to the BirdCLEF2019 dataset to classify 695 bird species, achieving a modest accuracy of 16% [5]. They pointed out that due to the shift-invariance and parameter-sharing characteristics of CNN, it was difficult for CNN to distinguish the spectrograms that occur at different frequencies but have similar shapes.

To address these limitations, recent studies have introduced multi-scale and attention-based architectures. Liu et al. proposed the Ensemble Multi-Scale CNN (EMSCNN), which achieved 91.49% accuracy on 30 bird species by concatenating features from convolutional kernels of varying sizes [8]. However, the method had many model parameters, which made training and inference time-consuming. In addition, the multi-scale features were directly cascaded, and the possible correlations and dependencies between these features were not deeply explored and utilized. Noumida and Rajan [9] discussed the classification of 10 bird species on ResNet50, CNN and VGG-16 networks, with classification accuracies of 96.3%, 93.7% and 91.9%, respectively. Experiments show that these methods were significantly better than the MFCC-DNN method based on sound signals. Hu et al. proposed MFF-ScSEnet, which fuses Mel spectrogram features with Sinc features to address the poor performance of single features for classification [10]. It reached 96.66% on 20 bird species and 96.28% on 15 self-built datasets. Despite these promising results, many models still lack effective multi-scale fusion strategies and rely heavily on single-type features.

In light of the above challenges, we propose a multi-scale feature fusion residual attention model (MUSCA), which extends ResNet18 by integrating both depthwise separable and traditional convolutional blocks for enhanced spectral texture extraction [11]. To address the limitation of CNNs in distinguishing frequency-variant but shape-similar spectrograms, MUSCA incorporates channel and spatial attention mechanisms to emphasize relevant features [12]. Furthermore, we introduce a multi-scale feature alignment and fusion strategy to model cross-scale dependencies and improve classification robustness.

The main contributions of this paper are as follows:

- (1) Birdsong signals contain multidimensional information, such as frequency, timbre, loudness and duration. Extracting effective sound features from them is a challenging task. To address this problem, a multi-scale feature fusion residual attention model (MUSCA) based on spectrogram birdsong recognition is constructed.

- (2) Aiming at the problem that CNN cannot distinguish spectrograms with different frequencies but similar shapes due to translation invariance and parameter sharing, channel attention and spatial attention are added to weight the spatial information of features.
- (3) Generalization ability and robustness are key indicators for measuring model performance. In the experimental design, we validated the performance of the proposed birdsong recognition model using the publicly available bird dataset.
- (4) The processing methods of birdsong signals lead to different scales of feature maps, causing the same model to show significant differences in recognition performance on spectrogram, wavelet spectrogram and Mel spectrogram features. This paper verifies the classification performance of the proposed method on different features, including spectrogram, wavelet spectrogram and Mel spectrogram.

2. Research Method

This study mainly includes two parts. The first part introduces the construction method of time–frequency spectrum features, that is, three spectral representations are extracted from the original audio signal: logarithmic power spectrum (log-power), logarithmic Mel spectrum (log-Mel) and logarithmic wavelet spectrum (log-wavelet). These features can describe the frequency distribution and time variation characteristics of birdsong from different dimensions. The second part describes, in detail, the network structure of the proposed MUSCA model (multi-scale feature fusion hybrid convolutional attention model). The model combines a hybrid convolution module with channel attention and spatial attention mechanisms and improves the effectiveness and classification performance of feature extraction through a multi-scale feature fusion strategy.

2.1. Construction of Time–Frequency Spectrum Features

Feature extraction is a crucial step in speech signal and image classification. The performance of classification largely depends on the extracted features. In this article, the main features used are the logarithmic spectrum, logarithmic Mel spectrum and logarithmic wavelet spectrum for birdsong classification. A brief description of each feature extraction method is as follows.

2.1.1. Log-Power Spectrogram Features

Log-power spectrum is a commonly used representation method for processing audio signals, which is based on the logarithmic transformation of the spectrum. The spectrum is the result of the audio signal through the short-time Fourier transform (STFT) to convert the frequency components and amplitudes of the signal from the time domain to the frequency domain [13]. When the logarithm of the spectrum is taken, the result is the logarithmic spectrum, which is widely used in many audio analysis and processing tasks [14]. The formula is as follows (1), where y is the speech signal. The extracted birdsong power spectrum features are shown in Figure 1a. The relevant description is in Equation (1).

$$\text{Log - Spectrum } (t, f) = \log \left(|STFT(y)|^2 + \epsilon \right) \quad (1)$$

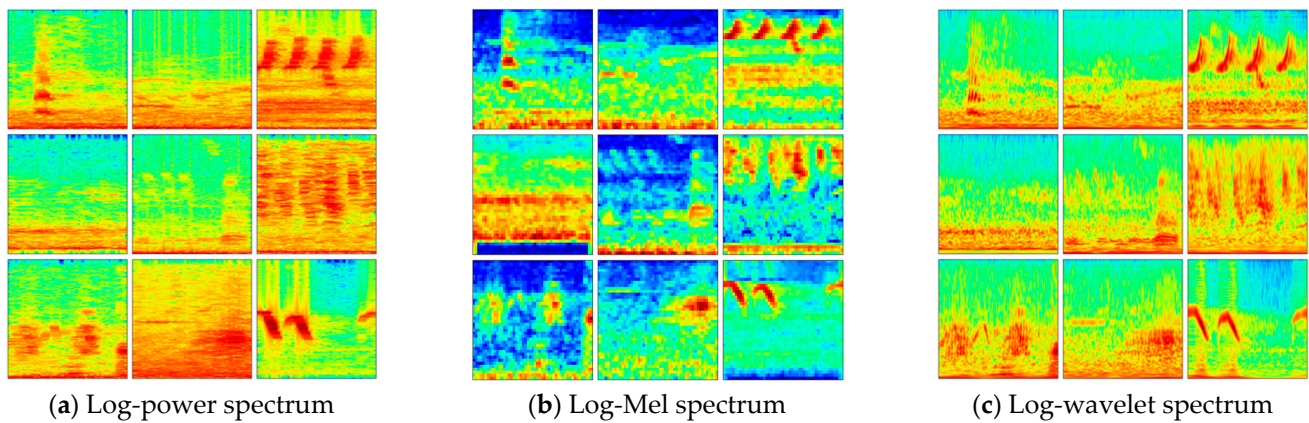


Figure 1. The spectral feature maps used in this article.

2.1.2. Log-Mel Spectrogram Features

Log spectrogram is one of the most widely used input features in many speech signal processing applications [15]. The Mel spectrogram represents the energy distribution of a signal on the Mel frequency axis, which is obtained by converting linear spectral data to the Mel scale using a filter bank. Compared with linear spectrograms, the log-Mel spectrogram captures the frequency components of birdsong in a way that emphasizes perceptually and structurally important frequency bands, making it more effective for distinguishing subtle acoustic patterns.

The Mel frequency scale is a nonlinear frequency scale that approximates how frequency components are spaced in real-world sounds. It allocates more resolution to lower frequencies and compresses higher frequencies, which helps emphasize the parts of the signal that carry more discriminative information.

In this study, first, we apply the Short Time Fourier Transform (STFT), as described by Griffin and Lim, to obtain spectrograms [13]. Then, we use a Mel filter bank with 40 frequency bands to project the spectrogram onto the Mel scale, producing the log-Mel spectrogram. The extracted birdsongs are shown in Figure 1b. The formula is shown in (2), where *Mel bank* represents the Mel filter bank. The relevant description is in Equation (2).

$$\text{Log-MelSpectrum}(t, f) = \log \left(\text{Mel bank} \left(|STFT(y)|^2 \right) + \epsilon \right) \quad (2)$$

2.1.3. Log-Wavelet Spectrogram Features

Continuous wavelet transform (CWT) employs the concept of multi-resolution analysis to perform non-uniform partitioning in the time–frequency domain, achieving appropriate time–frequency resolutions for different time signals. This characteristic makes wavelets particularly suitable for speech signal processing. For speech signals, the high-frequency components require better time resolution to detect transient parts that change rapidly, while the low-frequency components need higher frequency resolution to accurately track features such as slowly varying formats over time.

The formula is shown in Equation (3), where *Log-CwtSpectrum*(*t*, *f*) represents the log-wavelet spectrum at time *t* and frequency *f*. The term *cwtmatr*(*t*, *ω*) denotes the wavelet coefficient matrix obtained from the continuous wavelet transform, representing the wavelet coefficients at time *t* and frequency *ω*. An example of the extracted birdsong audio is illustrated in Figure 1c.

$$\text{Log-CwtSpectrum}(t, f) = \log(|\text{cwtmatr}(t, \omega)| + \epsilon) \quad (3)$$

2.2. Multi-Scale Feature Fusion Hybrid Convolutional Attention Model(MUSCA)

MUSCA is a residual mixed convolutional attention model based on multi-scale multi-layer perceptron fusion. The structure of MUSCA proposed in this article for the bird singing classification task is shown in Figure 2.

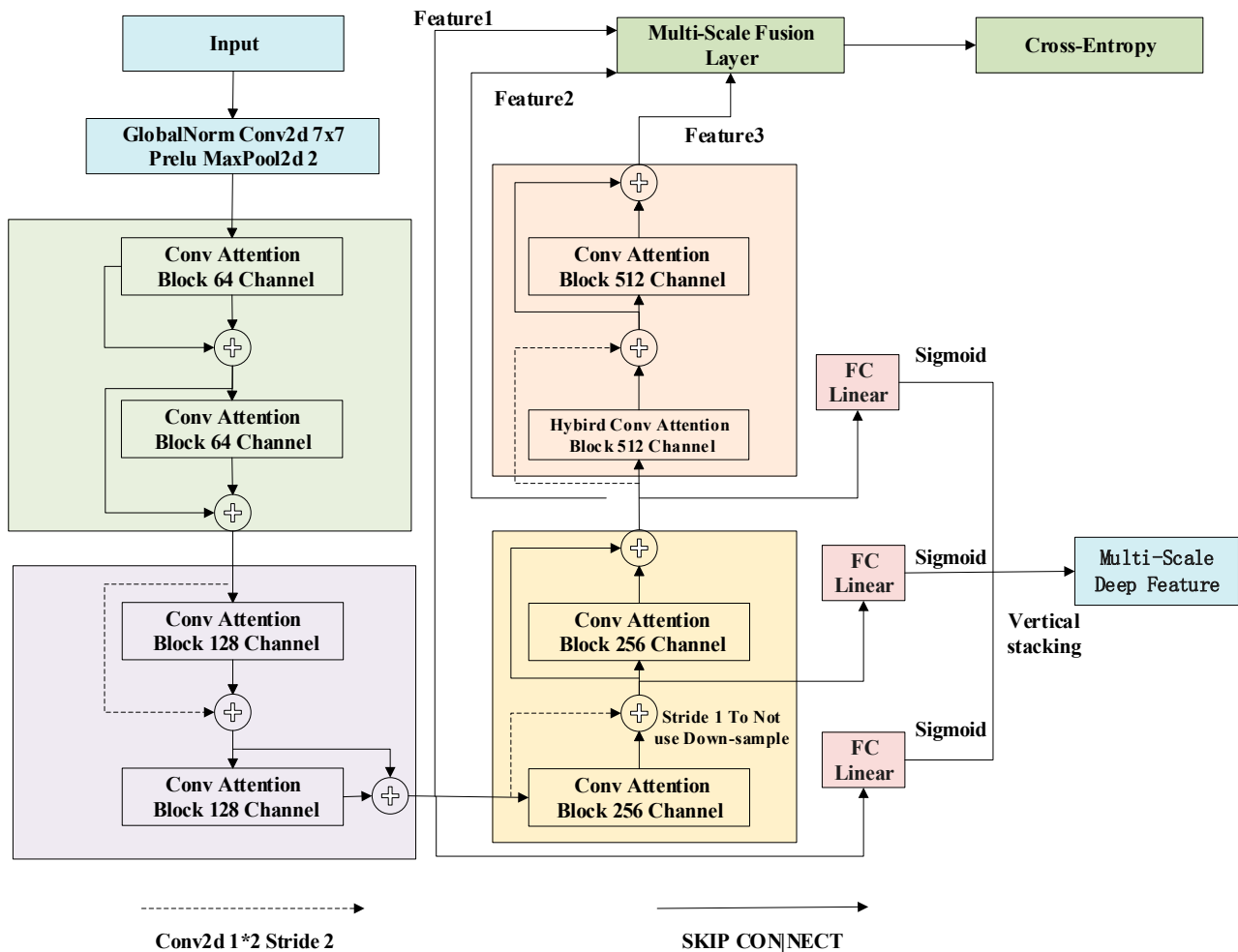


Figure 2. Structure of MUSCA. Conv2d 1*2 Stride 2 denotes a convolution kernel of length 1 and width 2, Stride is 2.

The MUSCA begins with a 7×7 convolutional layer to down-sample the input features and generate a feature map with 64 channels, which is then fed into the hybrid convolutional attention block. Within this block, some input feature maps undergo down-sampling with a stride of 2, reducing the feature map size to decrease computational load and filter out irrelevant features. The model then utilizes two hybrid convolutional attention blocks with skip connections to form a residual block, allowing the network to learn residual mappings instead of direct mappings [11]. This design facilitates easier gradient flow through the network, addressing issues of vanishing and exploding gradients. Through multiple convolutional down-sampling and residual operations, the model generates feature maps at large, medium and small scales. These multi-scale feature maps undergo global average pooling, and the resulting features are passed through fully connected layers to obtain classification probabilities for multiple categories. These probabilities are normalized to the 0–1 range using the sigmoid activation function. The multi-scale predictions are concatenated, and the combined output is compared to the one-hot encoded true labels using binary cross-entropy loss. This step reduces inter-class distance and weakens irrelevant features. The design idea refers to the backbone network of YOLOv3; by down-sampling

the image and combining deconvolution up-sampling, the multi-scale features are spatially aligned to achieve multi-scale fusion [16]. Similar to this, our method also uses multi-scale feature maps, and the global average pooling features from large, medium and small scales are sent to the multi-scale feature fusion layer. Global average pooling is used to reduce the spatial dimension of the feature map while retaining the channel information. It also helps to avoid overfitting by reducing the number of parameters. The layer integrates, weighs and scales features from different scales and then maps the aggregated features to the final category probability through the fully connected layer.

2.2.1. Hybrid Residual Convolutional Attention Block

Traditional convolution blocks involve convolution operations across channels and spatial dimensions, which limits the capture ability of single-channel features and hinders the extraction of spatial dimension details. To address this issue, this paper proposes a hybrid convolutional attention block. This structure combines depthwise separable convolutions with traditional convolutional blocks to form a hybrid convolutional block. It incorporates channel attention and spatial attention mechanisms to create a hybrid convolutional attention block aimed at extracting finer-grained features. The hybrid convolutional attention block is illustrated in Figure 3, and its specific structure is described in Equation (4).

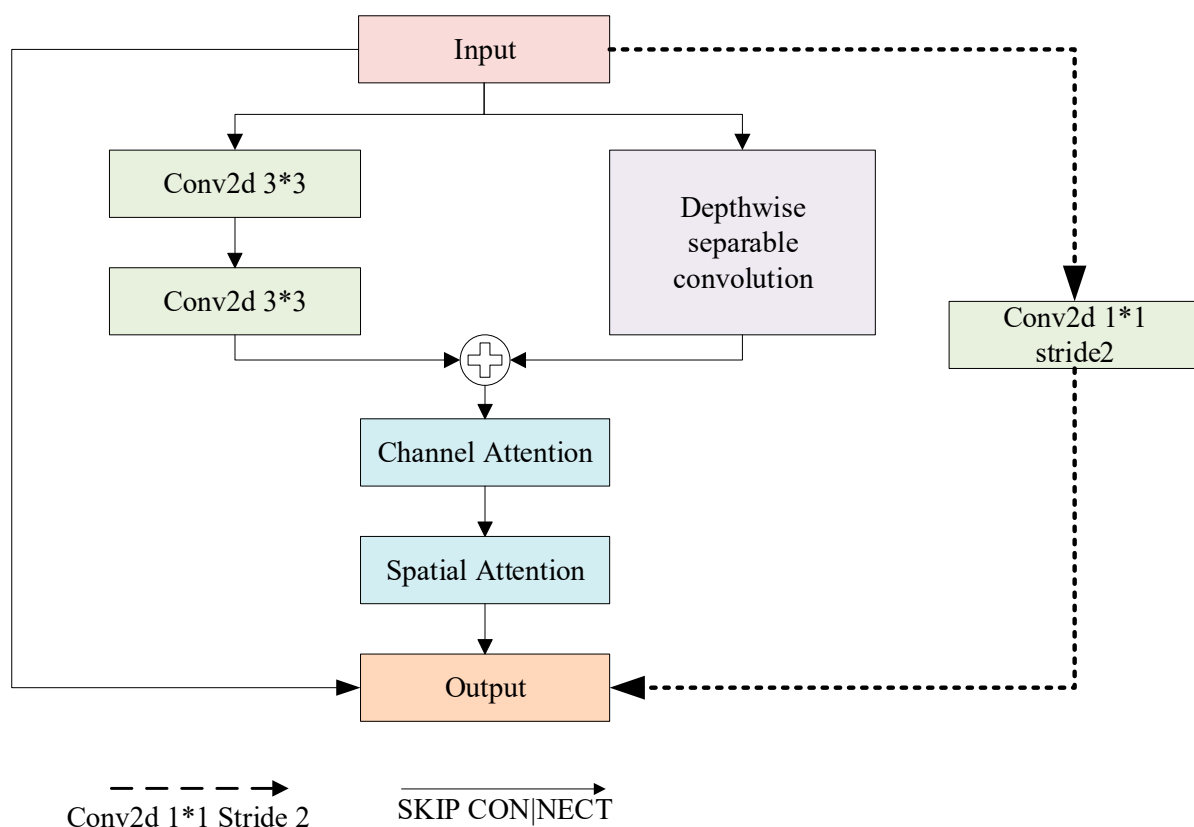


Figure 3. Hybrid convolutional attention block structure diagram. Conv2d 1*1 Stride 2 denotes a convolution kernel of length 1 and width 1, Stride is 2.

This hybrid convolutional attention block is primarily used for feature extraction and down-sampling operations. If the stride is set to 1, down-sampling is not performed, and the input and output are directly added pointwise. On the contrary, the down-sampling is performed, and the convolution block scales the feature map before adding the feature map to realize the residual connection. The residual operation allows the model to retain more original feature information in the subsequent layers, preventing the gradient vanishing

problem caused by excessive model depth and enhancing the model's learning efficiency and convergence speed.

$$\begin{aligned}
 Z &= \text{Conv}2d_{3 \times 3}(\text{Conv}2d_{3 \times 3}(X)) + \text{DepthwiseSeparableConv}(X) \\
 O &= \text{SpatialAttention}(\text{Channel Attention}(Z)) \\
 \text{ResidualOutput} &= \begin{cases} X + O, & \text{if } X \text{ is not down-sampled} \\ \text{conv}_{1 \times 1, \text{stride}=2}(X) + O, & \text{if } X \text{ is down-sampled} \end{cases} \quad (4)
 \end{aligned}$$

Among them, X is the input feature, O is the output of spatial attention and self-attention, and Z is the mixed convolution result.

Hybrid Convolution: The mixed convolution block, as depicted in Figure 4, amalgamates depthwise separable convolution with traditional convolution blocks. Depthwise separable convolution autonomously executes spatial convolution operations on each input channel, enabling it to capture feature information from each channel individually. Meanwhile, the convolutional block comprising two traditional convolutions conducts inter-channel and spatial convolution operations on the input to capture feature information across channels. The features derived from depthwise separable convolutions and ordinary convolutional blocks are fused through element-wise addition.

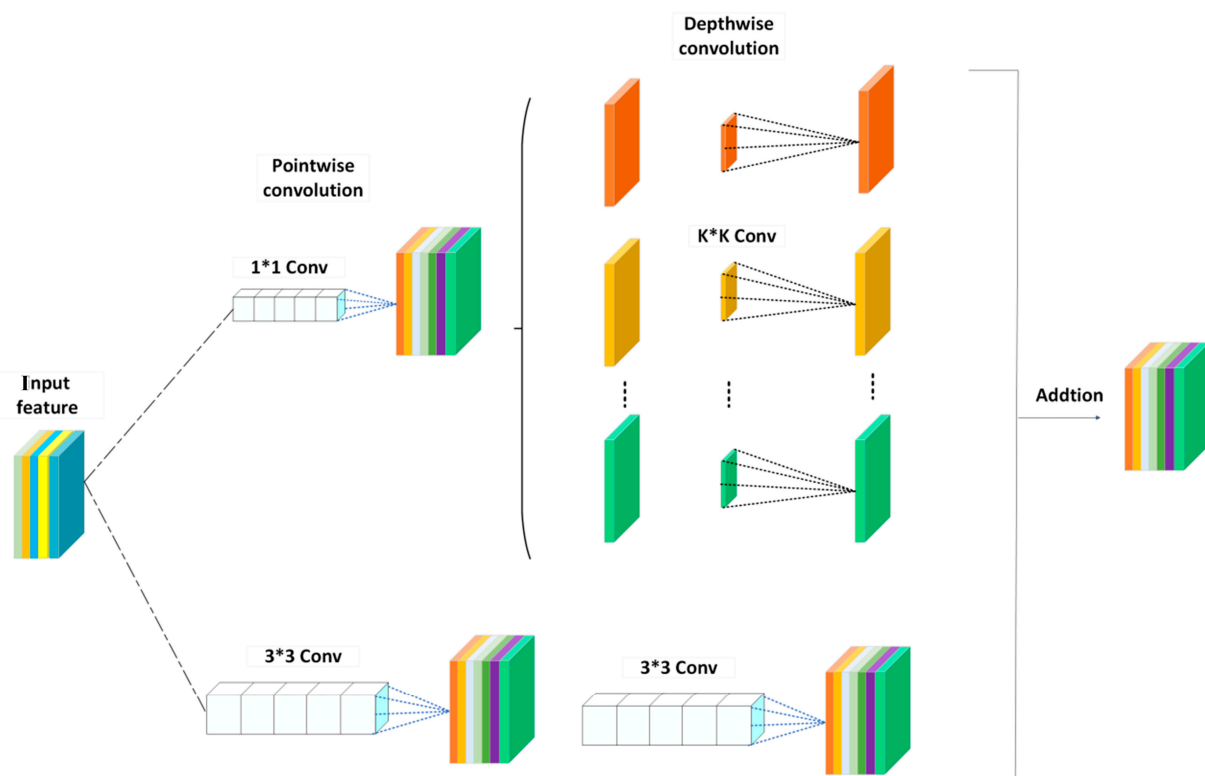


Figure 4. Hybrid residual convolution block structure diagram. The front of all * in the figure represents the length of the convolution kernel, and the back of * represents the width of the convolution kernel.

Channel Attention Mechanism: Figure 5 introduces the structure of channel attention mechanism, which enables the model to learn to assign different weights according to the importance of each channel. The channel attention mechanism performs global average pooling on input features, fully connects and activates the global average pooling, and uses Sigmoid to obtain weighted weights for each channel. The input features are then weighted according to the input channel dimension to achieve channel attention.

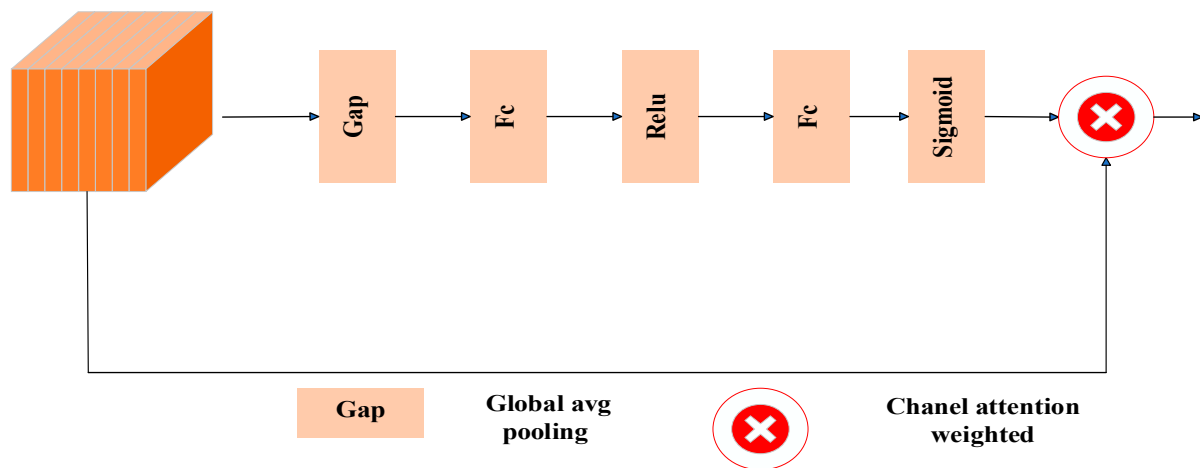


Figure 5. Structure diagram of channel attention mechanism.

Spatial Attention Mechanism: Integrating spatial attention into the convolution block aims to enhance the model's attention to various spatial locations in the input feature map and improve its ability to capture relevant spatial features, thereby improving its overall performance and generalization ability, as shown in Figure 6. Initially, the spatial attention mechanism computes the maximum and average values of the input features along the channel dimension, resulting in a feature map with identical dimensions as the input but with a single channel. Subsequently, these maximum and average values are concatenated along the channel dimension to form a feature map with two channels. This feature map then undergoes 2D convolution to fuse the bilinear features, following which the sigmoid activation function is applied to normalize the values between 0 and 1, yielding spatial attention weights. These weights are then multiplied elementwise with the input features, and the resulting product is added to the input to mitigate potential errors in subsequent layers due to deviations in spatial attention weights. The detailed structure of the spatial attention mechanism is elaborated in the diagram.

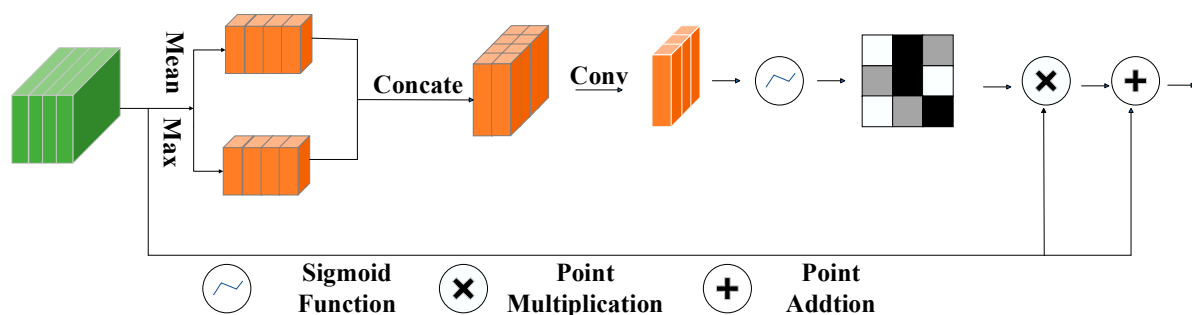


Figure 6. Structure diagram of spatial attention mechanism.

2.2.2. Multi-Scale Feature Fusion

The three features in Figure 7 are derived from the feature maps after the last three down-sampling operations in the model and the results obtained after global average pooling. Specifically, feature 1 is the pooling result of the current layer, and feature 2 is the pooling result of the previous scale. Through the above operations, the fusion and guidance between different scale features are realized. Firstly, feature 1 is mapped to a high-dimensional space through the fully connected layer, and then the result is normalized to the range of 0–1 using the sigmoid function and multiplied by feature 2. This stacking of multi-layer features is performed in this way. In addition, a skip connection is used every three layers to avoid overfitting due to the excessive number of fully connected layers.

These operations facilitate the effective fusion and combination of features in the feature space and enhance the representation ability and overall performance of the model.

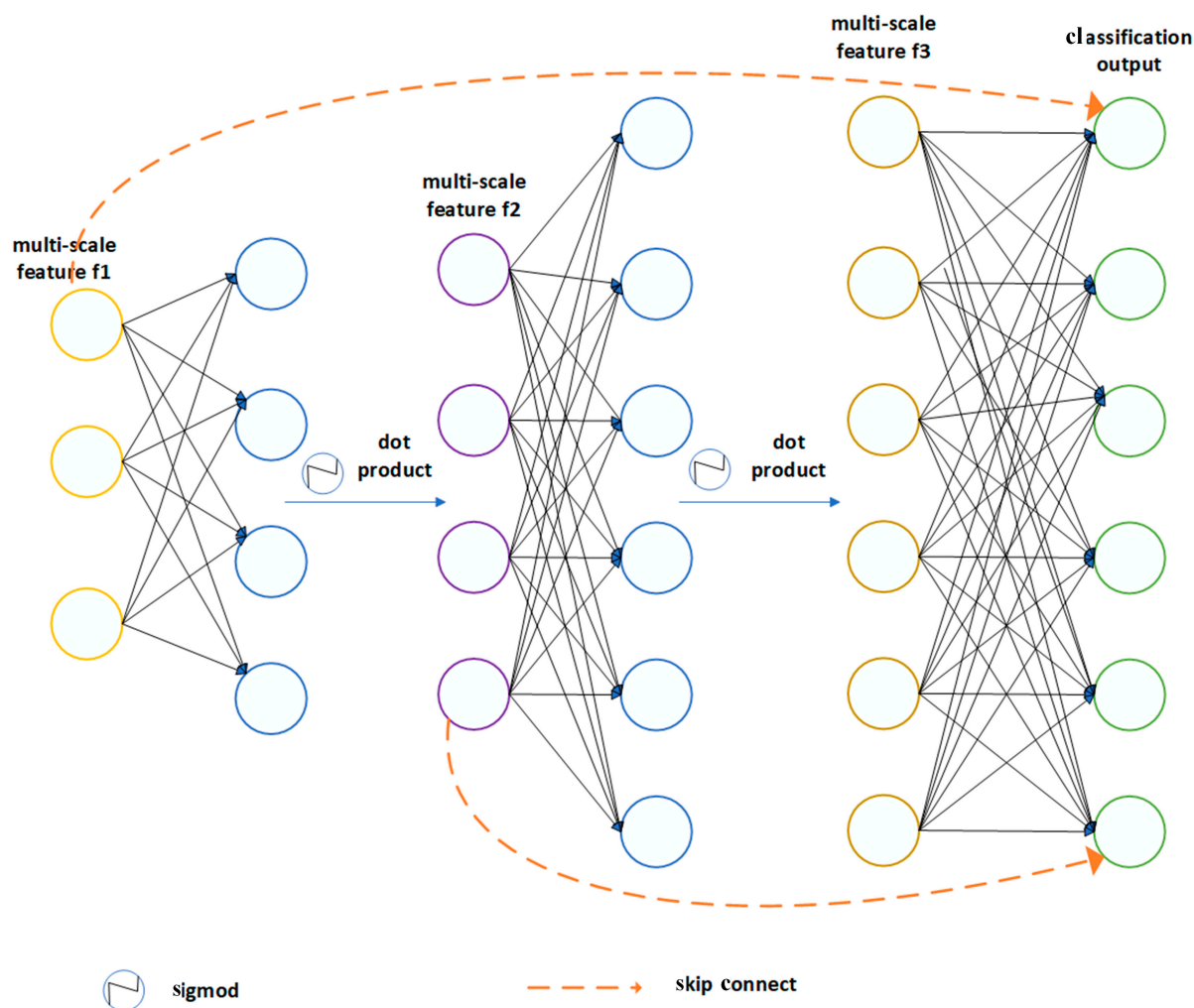


Figure 7. Structure diagram of multi-layer feature fusion and multi-layer perceptron.

The above model structure is described in Equation (5):

$$\begin{aligned} h_{i+1} &= \sigma(w_i h_i + b_i) \odot x_{i+1}; h_0 = x_1; y = \\ h_n &+ h_n \odot \sigma(w_1 x_1 + b_1) + h_n \odot \sigma(w_2 x_2 + b_2) \end{aligned} \quad (5)$$

Among them σ is the sigmoid activation function, $\odot$ is the dot product, which represents each layer of MLP, and x is the multi-scale feature.

2.2.3. Multi-Scale Category Weight Loss Function

Cross-entropy loss used for multi-classification tasks, outputting a probability vector for each sample, represents the model's predicted probability distribution for each category. However, in the case of imbalanced categories, cross-entropy loss may cause bias in accuracy and other indicators, as the model may perform well in most categories and perform poorly in a few categories [17,18]. This article introduces category weight cross-entropy. The calculation formula for its loss function in Equation (6):

$$L(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C w_c \cdot y_{ic} \cdot \log(\hat{y}_{ic}) \quad (6)$$

Among them, N is the number of samples, c is the number of categories, y_{ic} is the true label (0 or 1) of category c corresponding to sample i , and $\hat{y}_{ic}$ is the probability predicted by the model that sample i corresponds to category w_c , the sample weight for the c category.

In deep learning, binary cross-entropy is used as a loss function in binary classification problems. It measures the difference between the predicted label and the real label of each sample in the binary classification task through the model. The relevant description is in Equation (7).

$$L(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (7)$$

where N is the number of samples, y_i is the true label of the i sample, where the value is 0 or 1, and $\hat{y}_i$ is the predicted output of the model for the i sample, which represents the probability that the sample belongs to the first positive class.

The loss function used in this article is described in Equation (8):

$$\begin{aligned} L(y, \hat{y}) &= \frac{1}{N} \sum_{i=1}^N \sum_{m=1}^M [-y_{mi} \log(\hat{y}_{mi}) - (1 - y_{mi}) \log(1 - \hat{y}_{mi})] + \\ L(y, \hat{y}) &= \frac{1}{N} \sum_{i=1}^N \sum_{m=1}^M [-y_{mi} \log(\hat{y}_{mi}) - (1 - y_{mi}) \log(1 - \hat{y}_{mi})] + \\ &\quad \left(-\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C w_c \cdot y_{ic} \cdot \log(\hat{y}_{ic}) \right) \end{aligned} \quad (8)$$

In this context, M represents the loss across m scales, with the cumulative binary cross-entropy loss across multiple scales primarily addressing multi-scale task scenarios. This loss function facilitates the model's enhanced adaptability to input data of varying scales, consequently augmenting its generalization capability for multi-scale tasks. Weighted cross-entropy serves the purpose of balancing the sample sizes across diverse categories.

3. Experiment and Result Analysis

This chapter mainly introduces the design and performance evaluation of model experiments. Firstly, the dataset and preprocessing process are briefly described. Then, the hardware platform and parameter settings of the experiment are described. Then, the performance is compared with a variety of classical models under different spectral characteristics. In order to further verify the effectiveness of this method, a horizontal comparison with the current advanced methods is also made. Finally, the specific contribution of the multi-scale feature fusion module is analyzed by ablation experiments.

3.1. Dataset

The data used in this study are from the publicly available dataset provided by the Beijing Academy of Artificial Intelligence (BAAI), which is from <https://www.birdsdata.com/> (accessed on 20 July 2024). The dataset comprises the sounds of 20 common bird species in China, spanning eight orders, 12 families, 18 genera and 20 species, with a total of 14,311 natural audio clips, each clip lasting 2 s. In this study, effective data preprocessing is performed on the original audio before resampling to prevent obvious frequency aliasing after the sampling rate is reduced. Subsequently, the audio is resampled to 16 kHz, which is sufficient to cover the basic frequency information of most bird calls and can reduce the computational cost of model training. In addition, there are often local differences in the performance of bird sounds in the time dimension and spectrum. The audio is divided into frames of 400 ms. Through segmentation processing, the representative acoustic features in the local time range can be captured more accurately, thereby enhancing the model's perception of short-term audio modes and improving the overall classification performance. In addition, three spectral features of log-Mel, log-power and log-wavelet are used to

extract more comprehensive frequency domain and time-varying features, which enhances the recognition ability of the model under the limitation of sampling rate. Subsequently, the dataset was divided into the training set (50,088 clips), validation set (14,382 clips) and test set (7085 clips) at a ratio of 7:2:1. The distribution of each category in the segmented dataset is shown in Figure 8.

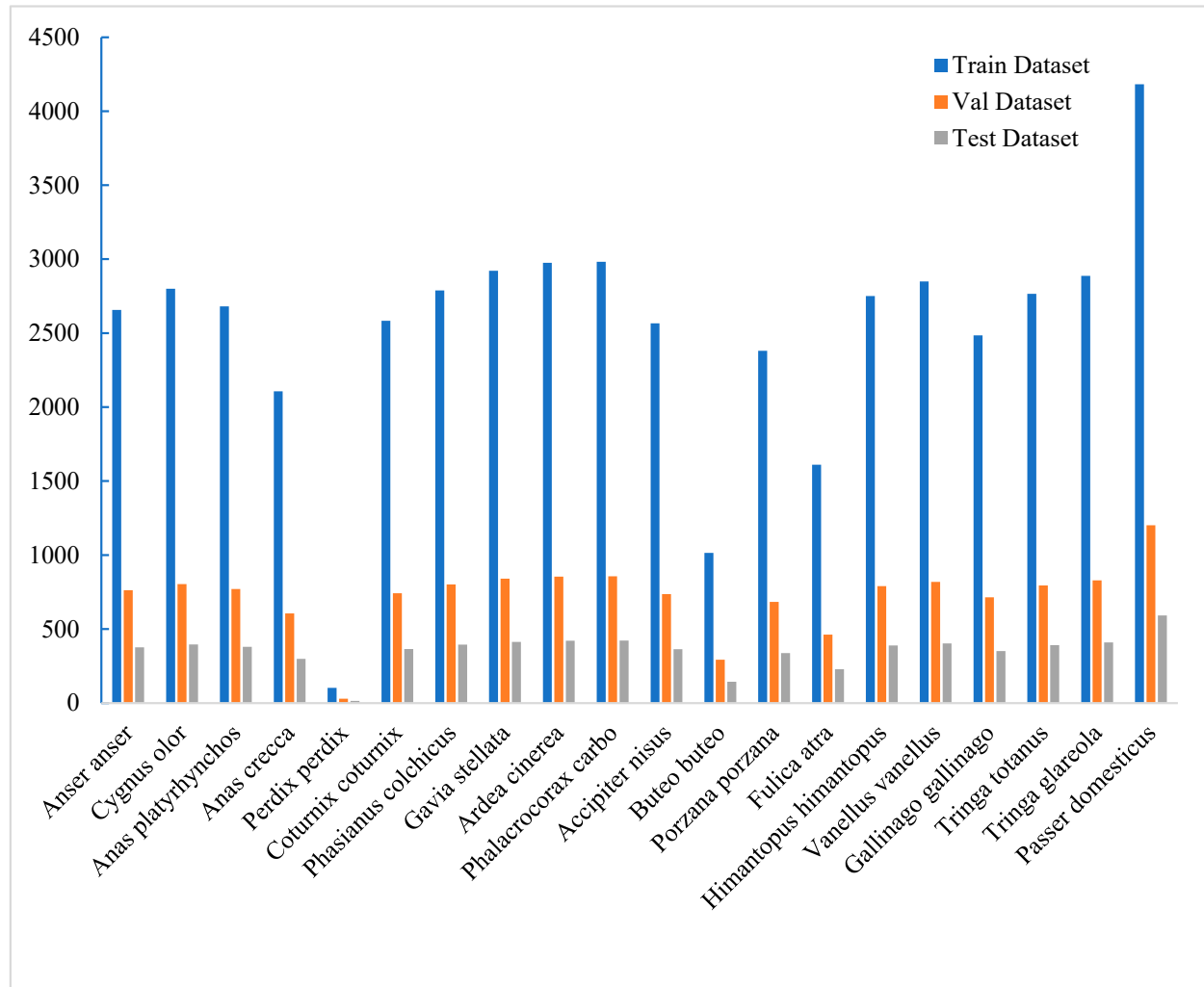


Figure 8. Statistical charts for each dataset.

3.2. Experimental Environment and Performance Evaluation

The experimental setup involved a desktop computer equipped with 128 GB of memory, a 16-core 32-thread CPU operating at a main frequency of 3.40 GHz and a NVIDIA GeForce RTX 4090 24 G GPU. The experiments were conducted on a 64-bit Ubuntu operating system. Anaconda3, Py Charm 2023.3.2 and Python 3.10 are used as deep learning platforms.

In this study, binary cross-entropy loss is introduced on the multi-scale features of each layer to achieve more fine-grained supervision constraints. The final total loss consists of two parts: the first part is the sum of the binary cross-entropy loss calculated between each layer of features after full connection and category multi-label coding, which is not used in inference; the second part is the standard multi-classification cross-entropy loss calculated after these fully connected features are aligned and fused by the multi-scale feature fusion module described in Section 2.2.2. The final total loss is a weighted combination of the above two. The inference stage only relies on the fusion features output by the multi-scale feature fusion module, and the final multi-classification results are generated through the

softmax layer. In order to comprehensively measure the performance of the model in each category, we calculate the evaluation indexes, such as precision and recall, for each category. The design aims to introduce continuous supervision signals for each layer of the feature extraction process so as to guide the model to learn more discriminative semantic representations at different scales. At the same time, the structure helps to enhance the gradient flow in the back propagation process, accelerate the convergence process of the model and improve the stability of training. The birdsong classification task in this study is a multi-classification problem involving 20 species of birds. The precision, recall and F1 values are calculated on the basis of each class and averaged using the macro-averaging method to ensure equal weights in all classes.

The performance of the proposed birdsong classification model is evaluated by accuracy, precision, recall and F1 score [19]. The definition of each evaluation index is shown in Table 1.

Table 1. Explanation of parameters in evaluation metric calculations.

Term	Description	Explanation
TP	True positive	Predicted as positive and actually positive
FP	False positive	Predicted as positive but actually negative
FN	False negative	Predicted as negative but actually positive
TN	True negative	Predicted as negative and actually negative

In multi-classification, when focusing on a specific category A, the sample whose true label is A is called a positive example of the category. On the contrary, the real label is not all the samples of category A, which is called the negative example of the category, that is, the samples of the ‘other categories’.

Accuracy is used to measure the performance of recognition models and indicate the proportion of the number of correct samples predicted by the model to the total number of predicted samples:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (9)$$

Precision is judged based on the prediction results:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (10)$$

Recall is judged based on the actual samples:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (11)$$

F1 score is the harmonic mean of precision and recall used to balance these two metrics. The expression is as follows:

$$\text{F1 - score} = (1 + \beta) \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

TP, TN, FP and FN, respectively, represent true positives, true negatives, false positives and false negatives. The F1 score is a metric used to evaluate the accuracy of a model. Precision indicates the proportion of true positive samples among all samples classified as positive. Recall denotes the proportion of true positive samples among all actual positive samples.

3.3. Experimental Results and Analysis

To validate the effectiveness of the proposed method, 12 sets of comparative experiments are designed. Using log-power spectrograms, log-Mel spectrograms and log-wavelet spectrograms, the proposed model is compared with the classic ResNet18, ResNet50 and CNN models. To ensure a controlled and comparable experimental environment, all models were trained using the same optimizer and learning rate. While model-specific tuning could potentially improve performance, our goal was to evaluate the relative performance of architectures under consistent training conditions. Each model was optimized using the Adam optimizer and trained on the training set for 50 epochs with a learning rate of 1×10^{-3} . During each epoch, the validation set was evaluated, and the best weights were saved. Finally, the evaluation metrics were calculated on the test set for comparison.

For the log-power spectrogram, the parameters of STFT are as follows: 512 frequency points, a window length of 512 samples and a 75% frame overlap. For the log-Mel spectrogram, using the same STFT parameters, the number of Mel filters is 40. The log-wavelet spectrogram was computed with a frame length of 257 samples, 75% overlap between frames and 258 wavelet bases, enabling detailed time–frequency analysis for feature extraction.

3.3.1. Comparison with Traditional Methods

In this study, the proposed method was compared with traditional CNN, Resnet18 and Resnet50 models using multiple spectrogram features. Specifically, S1, S2 and S3 represent log-power spectrograms, log-Mel spectrograms and log-wavelet spectrograms, respectively. The detailed comparative experimental results are presented in Table 2, demonstrating that our proposed method achieves superior performance. As anticipated, multi-scale feature extraction and fusion of different sound spectral features significantly enhance birdsong classification performance. Related Code Address: https://github.com/GLL620811/GLL620811/tree/main/bird_cls/bird_cls (accessed on 5 April 2025)).

Table 2. Comparison of performance of different models in different features.

Feature	Model	Accuracy	Precision	Recall	F1 Score
S1	MUSCA(Ours)	0.9544	0.9451	0.9523	0.9483
S1	Resnet18	0.9235	0.9162	0.8944	0.9013
S1	Resnet50	0.9207	0.9236	0.8962	0.9063
S1	CNN	0.8765	0.8797	0.8524	0.8626
S2	MUSCA(Ours)	0.9697	0.9632	0.9636	0.9634
S2	Resnet18	0.9509	0.9483	0.9474	0.9476
S2	Resnet50	0.9339	0.9336	0.9274	0.9297
S2	CNN	0.9139	0.9131	0.8962	0.9032
S3	MUSCA(Ours)	0.9379	0.9318	0.9353	0.9330
S3	Resnet18	0.9005	0.9020	0.8857	0.8921
S3	Resnet50	0.9053	0.9101	0.8978	0.9029
S3	CNN	0.8399	0.8349	0.8178	0.8240

3.3.2. Comparison with the Most Advanced Approaches

In this study, we also compared the performance of the most advanced models for multi-species birdsong recognition, analyzing their accuracy. The detailed comparison results of each model, as cited from research by [10], are presented below. To ensure a fair comparison, the bird datasets were used. As shown in Table 3, our proposed method achieved relatively higher accuracy among all the models, indicating its advantage in the task of birdsong recognition. The experimental data for SDFIE-NET were obtained from Zhang, Hu et al. [20].

Table 3. Classification accuracy on the bird datasets.

Model	Input Type	Accuracy
(2DCNN + 3DCNN) + GRU + GRU	Image, continuous frame-sequence	0.9588
SFLN(GRU)	Continuous frame sequence	0.9592
EfficientnetV1_b0	Mel spectrogram	0.9502
EfficientnetV2_s	Mel spectrogram	0.9530
MFF-ScSEnet	Mel-sinc spectrogram	0.9666
SDFIE-NET	Unknown	0.9410
MUSCA(Ours)	Mel spectrogram	0.9697

3.4. Ablation Study

In this paper, the ablation experiment is carried out by the proposed multi-scale feature alignment and fusion method. The multi-scale features are directly cascaded, and the classification probability is output using the fully connected layer while keeping other parameters consistent. The experimental results, as shown in Table 4, indicate a varying degree of decline in accuracy, precision, recall and F1 score, which aligns with our expectations. This shows that the proposed multi-scale feature alignment and fusion method effectively utilizes the information between each other.

Table 4. Removal of multi-scale feature alignment fusion.

Feature	Model	Accuracy	Precision	Recall	F1 Score
S1	MUSCA (Ours)	0.9490	0.9430	0.9501	0.9459
S2	MUSCA (Ours)	0.9553	0.9534	0.9556	0.9543
S3	MUSCA (Ours)	0.9294	0.9230	0.9293	0.9254

4. Discussion

In this study, the log-Mel spectrogram, log-power spectrogram and log-wavelet spectrogram were used to extract the frequency domain information of birdsong. Multi-scale features are extracted by using a hybrid convolutional layer, and then feature alignment and fusion are performed. In this way, more comprehensive features are obtained, thus improving the recognition performance of the network. In addition, considering that convolutional neural networks cannot distinguish spectral features with different frequency domains but the same shape, this study introduces channel attention and spatial attention mechanisms to weigh the feature maps to improve the recognition accuracy of the network.

During the research process, it was found that using weighted cross-entropy loss as the optimization target significantly improved the recognition of small sample classes. As shown in Figure 9, the model accurately identified the birdsong labeled as category 10. Furthermore, there were differences in classification accuracy among the log-power spectrogram, log-wavelet spectrogram and log-Mel spectrogram, with the log-Mel spectrogram performing the best. This is because the Mel filter makes it easier to capture the effective information of the sound, resulting in a more concentrated spatial distribution of the feature texture, which is easier to identify by the model.

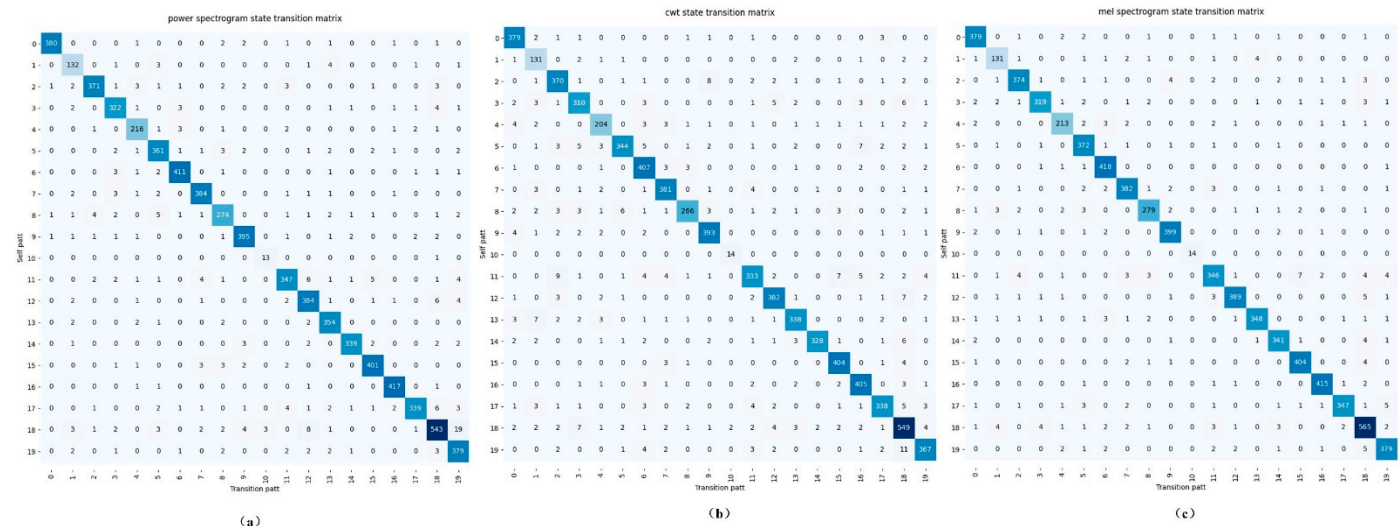


Figure 9. The classification accuracy of the test set in each spectrogram. (a) Confusion matrix of log-power spectrogram; (b) confusion matrix of log-wavelet spectrogram; (c) confusion matrix of log-Mel spectrogram.

5. Conclusions

This study presents a novel multi-scale feature fusion residual attention model (MUSCA) for birdsong recognition based on audio spectrogram features. Leveraging a shallow residual network as the backbone and integrating both channel and spatial attention mechanisms, MUSCA effectively addresses the challenge of extracting meaningful information from complex, multi-dimensional birdsong signals.

By fusing log-power spectrograms, log-Mel spectrograms and log-wavelet spectrograms, the proposed model significantly mitigates the information loss typically associated with single-scale feature extraction. The introduction of attention mechanisms further enhances the model's capacity to distinguish between spectrograms with similar shapes but different frequency distributions, overcoming key limitations of conventional CNNs such as translation invariance and parameter sharing.

Extensive experiments conducted on a publicly available birdsong dataset validate the robustness and generalization ability of MUSCA across various spectrogram types. The model consistently outperforms classical single-scale architectures such as CNN, ResNet18 and ResNet50, especially under challenging acoustic conditions.

Despite its promising performance, this study acknowledges several limitations. First, the manual annotation of birdsong data remains time-consuming and labor-intensive, posing challenges to the scalability of large-scale ecological research. Future work will focus on developing automated techniques for birdsong endpoint detection and species labeling to accelerate dataset construction and reduce human effort. Second, although down-sampling bird audio reduces computational load and training time, it may also result in the loss of harmonic information critical for accurate species identification. Balancing efficiency and acoustic fidelity will be an important consideration in future improvements. Lastly, the integration of multi-scale feature extraction and attention mechanisms, while enhancing recognition accuracy, inevitably increases the model's computational complexity. As a next step, we plan to explore model compression strategies and investigate lightweight alternatives, such as MobileNet-based attention modules, to reduce the computational footprint without compromising performance.

Author Contributions: Conceptualization, L.G.; Data curation, L.G., W.L. and Z.W.; Funding acquisition, G.D., D.L., Y.Z. and Y.Y.; Investigation, L.G. and W.L.; Methodology, L.G., G.D., D.L. and Y.Z.; Supervision, G.D., D.L. and Y.Z.; Validation, D.L., Y.Z. and Y.Y.; Writing—original draft, L.G.; Writing—review and editing, L.G., G.D., D.L., Y.Z. and Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Agricultural Joint Fund of Yunnan Province under Grant no: 202301BD070001-086, the Scientific Research Foundation of the Education Department of Yunnan Province, China under Grant no: 2022J0495, the National Natural Science Foundation of China under Grant no: 31860332, the National Natural Science Foundation of China under Grant no: 32360388, and Research on the Application of Multi-Target Swarm Intelligence Algorithms with the Multi-Modal in Biological Data.

Data Availability Statement: All datasets included in this work are public datasets, ensuring transparency and accessibility.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Priyadarshani, N.; Marsland, S.; Castro, I. Automated birdsong recognition in complex acoustic environments: A review. *J. Avian Biol.* **2018**, *49*, jav-01447. [\[CrossRef\]](#)
2. Qi, J.; Gage, S.; Joo, W.; Napoletano, B.; Biswas, S. Soundscape characteristics of an environment: A new ecological indicator of ecosystem health. *Wetl. Water Resour. Model. Assess.* **2008**, 201–211.
3. Anusha, P.; ManiSai, K. Bird species classification using deep learning. In Proceedings of the 2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP), Hyderabad, India, 21–23 July 2022; pp. 1–5.
4. Browning, E.; Gibb, R.; Glover-Kapfer, P.; Jones, K.E. *Passive Acoustic Monitoring in Ecology and Conservation*; WWF-UK: Woking, UK, 2017; 76p.
5. Koh, C.-Y.; Chang, J.-Y.; Tai, C.-L.; Huang, D.-Y.; Hsieh, H.-H.; Liu, Y.-W. Bird Sound Classification Using Convolutional Neural Networks. In Proceedings of the Clef (Working Notes), Lugano, Switzerland, 9–12 September 2019.
6. Ranjan, R.; Thakur, A. Analysis of feature extraction techniques for speech recognition system. *Int. J. Innov. Technol. Explor. Eng.* **2019**, *8*, 197–200.
7. Potamitis, I.; Ntalampiras, S.; Jahn, O.; Riede, K. Automatic bird sound detection in long real-field recordings: Applications and tools. *Appl. Acoust.* **2014**, *80*, 1–9. [\[CrossRef\]](#)
8. Liu, J.; Zhang, Y.; Lv, D.; Lu, J.; Xie, S.; Zi, J.; Yin, Y.; Xu, H. Birdsong classification based on ensemble multi-scale convolutional neural network. *Sci. Rep.* **2022**, *12*, 8636. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Noumida, A.; Rajan, R. Deep learning-based automatic bird species identification from isolated recordings. In Proceedings of the 2021 8th International Conference on Smart Computing and Communications (ICSCC), Kochi, India, 1–3 July 2021; pp. 252–256.
10. Hu, S.; Chu, Y.; Wen, Z.; Zhou, G.; Sun, Y.; Chen, A. Deep learning bird song recognition based on MFF-ScSEnet. *Ecol. Indic.* **2023**, *154*, 110844. [\[CrossRef\]](#)
11. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
12. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [\[CrossRef\]](#)
13. Griffin, D.; Lim, J. Signal estimation from modified short-time Fourier transform. *IEEE Trans. Acoust. Speech Signal Process.* **1984**, *32*, 236–243. [\[CrossRef\]](#)
14. Incze, A.; Jancsó, H.-B.; Szilágyi, Z.; Farkas, A.; Sulyok, C. Bird Sound Recognition Using a Convolutional Neural Network. In Proceedings of the 2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY), Subotica, Serbia, 13–15 September 2018; pp. 000295–000300.
15. Yan, N.; Chen, A.; Zhou, G.; Zhang, Z.; Liu, X.; Wang, J.; Liu, Z.; Chen, W. Birdsong classification based on multi-feature fusion. *Multimed. Tools Appl.* **2021**, *80*, 36529–36547. [\[CrossRef\]](#)
16. Xie, S.; Lu, J.; Liu, J.; Zhang, Y.; Lv, D.; Chen, X.; Zhao, Y. Multi-view features fusion for birdsong classification. *Ecol. Inform.* **2022**, *72*, 101893. [\[CrossRef\]](#)
17. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.

18. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, October 22–29 2017; pp. 2980–2988.
19. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [[CrossRef](#)]
20. Zhang, Q.; Hu, S.; Tang, L.; Deng, R.; Yang, C.; Zhou, G.; Chen, A. SDFIE-NET—A self-learning dual-feature fusion information capture expression method for birdsong recognition. *Appl. Acoust.* **2024**, *221*, 110004. [[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.