Ecological Indicators 150 (2023) 110258

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

A hierarchical birdsong feature extraction architecture combining static and dynamic modeling

Yanan Wang [a], Aibin Chen [a,*], Huaicheng Li [a], Guoxiong Zhou [a], Jizheng Yi [a], Zhiqiang Zhang [b]

[a] *Institute of Artificial Intelligence Application, College of Computer and Information Engineering, Central South University of Forestry and Technology, Changsha 410004, China*

[b] *Wildlife Conservation and Utilization Laboratory, College of Forestry, Central South University of Forestry and Technology, Changsha 410004, China*

ARTICLE INFO

Keywords:
Birdsong classification
Context modeling
Hierarchical refinement model
Pearson correlation coefficient
Self-built dataset
Complex background noise

ABSTRACT

To conserve bird biodiversity and monitor the distribution of species in the region, it is of tremendous necessity to identify birds by their songs and explore the rich ecological information birdsong contains. The audios recorded in the monitoring area generally have complex background noise, the characteristics of the song are not prominent and the biological spectrum information is not comprehensive, which brings some challenges to the identification of birds. This study proposes a hierarchical birdsong feature extraction architecture combining dynamic and static modeling to cope with complex environments as a modeling context. Firstly, six common speech features were extracted for the characteristics of birdsong. The Pearson correlation coefficient is then used to analyze the correlations between birdsong and human speech, examining the correlations between each feature in the presence and absence of environmental noise interference. Combined with the scatter plot matrix analysis, we conclude that Mel Frequency Cepstral Coefficient (MFCC) is more suitable comparing with other features when dealing with birdsong and can superiorly cope with a complex background noise. Secondly, a feature extraction architecture is built, which integrates static and dynamic modeling to fully explore the contextual relationship, to solve the problem of ignoring the internal structure information of the patch and losing some spatial information in the Transformer-type model. Finally, a hierarchical refinement module is designed to help extract more detailed features, as well as to optimize the computational cost of the Transformer-type model that requires many training data and has high complexity. The performance of the model can be detected with 93.67 % accuracy on the self-built birdsong dataset, 95.19 % accuracy on the public birdsong dataset Birdsdata and 97.02 % accuracy on the public environmental dataset UrbanSound8k.

1. Introduction

In the current social context of ecological civilization construction, the concept of ecological conservation has prompted the study of flora and fauna to become a popular field (Tao et al., 2022). In-depth study of the habits of flora and fauna is conducive to the dynamic protection of their living environment and achieving superior conservation effects. The classification of birdsongs is an important ecological research topic, with bird species serving as a key indicator for such classification. Ecological indicators of bird species are closely associated with the production and performance of their vocalizations, which encompass biological traits, behaviors, environmental factors, among others. The classification of regional birdsong involves the application of two ecological indicators. One aspect is species diversity. Analyzing and categorizing the vocalizations of different bird species can determine the level of bird species diversity in a given area, thereby understanding the biodiversity status of the ecosystem. The second aspect is habitat preference, as birdsongs are typically associated with the habitat in which they reside. Conducting classification studies on birdsongs from different habitats can provide data analysis for understanding bird preferences for specific habitats (Brooker et al., 2020). Especially for some endangered bird species, immediate in situ conservation should be implemented to restore and improve the environment of the original habitat (Yu et al., 2021; Ganatsas et al., 2022). This study aims to identify different categories of birdsongs and explore the rich ecological information in the regional birdsongs to provide a reference for measuring the balance of biomes (Farwell et al., 2021). Identifying birds by their calls and pictures are both good research directions, but it may

[*] Corresponding author.
E-mail address: hotaibin@163.com (A. Chen).

https://doi.org/10.1016/j.ecolind.2023.110258
Received 5 December 2022; Received in revised form 6 April 2023; Accepted 12 April 2023
Available online 21 April 2023

be difficult for us to capture pictures of birds in real-time in a natural environment, while the birdsongs can be easily captured and identified (Hussain et al., 2018; Peng et al., 2018; Kumar and Das, 2019). Research on birdsongs has been conducted from the primitive use of machine learning methods to detect and recognize them. For example, the Dynamic Time Warping algorithm (DTW) has high accuracy but poor generalization (Tan et al., 2015), and later a study improved the performance of the algorithm by setting the slope of the path and combining it with global control to enhance the recognition of birdsongs (Jiang et al., 2021). There are also some traditional modeling methods such as Hidden Markov Model (Lee et al., 2013), Gaussian Mixture Model (Kalan et al., 2015), and the Support Vector Machine (Fagerlund, and Seppo, 2007; Wei et al., 2020; Cinkler et al., 2022). With deep learning excelling in computer vision tasks, the audio classification task borrowed many methods to process audio efficiently. From the original classification based on audio 1D time series data to the proposed end-to-end modeling of audio feature map mapping labels for classification (Jaitly and Hinton, 2011; Dieleman and Schrauwen, 2014; Trigeorgis et al., 2016). Á et al. (2018) converted birdsongs into spectrograms and then used convolutional neural networks (CNN) to extract features for classification (Á et al., 2018). With feature extraction and generalization capabilities that other types of models do not have, CNN has become the main network architecture used. Xie and Zhu (2019) proposed a deep learning-based bird sound classification method, and experiments on the classification of 14 bird species proved that the deep learning method outperformed the traditional meth (Xie and Zhu, 2019). The birdsong spectrogram is a time-varying image with time-series data features ignored when using CNN for classification. Later, it was shown that combining recurrent neural networks (RNN) to process spectrograms could yield superior results. Some of the collected birdsongs are discontinuous in the time domain, and for discontinuous sequential signals, exploring the context dependence in birdsongs can help to superior identify them (Morita et al., 2021). Zhang et al. (2019) proposed a combination of sliding window algorithm differential spectrogram and GRU as a classifier to solve this problem (Zhang et al., 2019). The birdsongs that we collect in real-time from the natural environment are generally also accompanied by complex background noise or sounds of other species. A combination of signal detection methods used by different recognizers was proposed to improve recognition performance (Brooker et al., 2020). Alternatively, unsupervised sound separation techniques were used to separate background sounds from birdsongs before learning high-quality features. (Dai et al., 2021; Denton et al., 2022). There are also birdsong classifiers that used machine learning methods, passing each sound through a noise suppressor and a separate classification procedure (Mehyadin et al., 2021). Another research idea was to design a channel for sound detection and classification, learn from weak labels, classify birdsong by fine-grained features, and perform robustness analysis for background sounds (Conde et al., 2021). Deep learning models suffer from high computational complexity along with rapid development, and bird identification using traditional methods has once again become a research hotspot (Pahuja and Kumar, 2021; Tuncer et al., 2021).

The birdsong spectrogram is a single-channel image that does not have as many relatively distinct features as the RGB image. The features that exist are relatively regionalized and marginalized, and even most of the pixels are blank or background noise. Therefore, the spatial capture of time–frequency features cannot be ignored for audio spectrograms (Kim et al., 2020). How to extract the optimal features in birdsongs to help improve the recognition effect is the focus of research. Feature scoring was proposed to evaluate the contribution of each feature to the classification to select the optimal feature (Xu et al., 2021). Alternatively, the number of selectable features can be increased to improve birdsong classification accuracy by applying multi-scale feature fusion (Yan et al., 2021; Xie and Zhu, 2022). These methods entail many calculations and there is a randomness in their effects. To capture more global features, studies proposed adding attention to CNN, which was used to learn the weight distribution and then act on the features to achieve advanced results for the classification of birdsongs (Yang et al., 2022). However, there is randomness in reassigning weights after convolution and the interpretability is not strong. The emergence of the Transformer architecture replaced most RNN models, modeling global dependencies by using self-attention weighting to calculate the relationship between each node of the input data. The self-attention produces a more interpretable model, and each attention head learns to perform different tasks (Vaswani et al., 2017). The Transformer showed excellent performance in a range of categorized tasks, demonstrating the competitiveness of pure attention-based models (Kitaev et al., 2020). Among them, Vision Transformer (ViT) achieved the most advanced experimental results at that time under the limitation of lacking inductive biases (Dosovitskiy et al., 2020). However, much data training is required upfront, and the model complexity is high, requiring advanced equipment to support long training periods (Yuan et al., 2021). Therefore, the use of traditional Transformer-type models directly applied to audio classification may not work well and has limitations. Audio Spectrogram Transformer (AST) is an improvement based on the ViT applied in the field of audio classification with a simpler structure (Gong et al., 2021). Both of them used the traditional Transformer, which ignores the internal two-dimensional structure and local spatial information after slicing the spectrogram into multiple patches. Small-size patches are still images in nature, and there is spatial information inside them, so it is not reasonable to input them directly into the Transformer after a simple linear projection (Guo et al., 2022). More studies shown that the Transformer layer is not irreplaceable and that it is not the best choice (Rao et al., 2021; Wang et al., 2022).

To improve the number of optimal features in the capture spectrogram to cope with the complex background noise, and solve the problems of the Transformer-type audio classification model, this study proposes a lightweight feature extraction method that combines statistical knowledge to analyze audio features. The main contributions of this paper are as follows:

(1) In this study, six common speech features are extracted according to the characteristics of birdsong, and then the correlation between birdsong and human speech, and between each feature interfered by environmental noise and not can be analyzed using the Pearson correlation coefficient. Combined with the scatter plot matrix analysis, we conclude that MFCC is more suitable comparing with other features when dealing with birdsong and can superiorly cope with a complex environmental background.

(2) The main module design adopts CoTAttention (Li et al., 2022) as the encoder and Multilayer Perceptron (MLP) as the feedforward network to build a new feature extraction architecture CMS. This architecture can well handle the spectrogram with time series features and can learn the time–frequency features more comprehensively. It solves the problems of ignoring the internal structure information of patches and losing part of the spatial information that exists in the Transformer model.

(3) We design a hierarchical refinement module that halves the feature map resolution layer-by-layer while increasing the dimensionality to preserve the extracted refinement features. This shallow-to-deep feature extraction approach optimizes the computational cost problem of the Transformer type model which requires many training data and high complexity.

2. Materials and methods

2.1. Self-built dataset

The self-constructed birdsong dataset Birdsound15 for this study was downloaded from Xeno-canto (Xeno-canto, 2022), with a total of 15 categories of birdsongs common to southern China. While considering the existence of birds of the same genus and different species, two species of swallow intra-class were downloaded for the production of the dataset, namely Hirundo rustica and Cecropis daurica. They have some differences in appearance, but their habits and calls are similar, making it difficult for people to identify them by listening to their calls. This increases the difficulty of the dataset identification task and tests the feature extraction ability of the model. Each audio file downloaded ranged from 2 s to 10 min in duration, and after listening to it, two main problems were found. (1) For the longer audio, the proportion of birdsongs is not large, and most of them may be other sounds and blank parts. (2) For the shorter audio, there may be no sound or complete noise in the entire audio due to upload errors or data loss problems.

The editing software used to produce the dataset is Adobe Audition 2022, and we can manually listen to the audio while observing the waveform graph and spectrogram of the audio during the editing. This will help us to eliminate the audio in the above two cases and prevent the overall quality of the dataset from being affected. After eliminating some useless raw audio files, the parts of the downloaded audio containing birdsongs were manually intercepted in units of 4 s. The intercepted audio is interspersed with some natural and urban environmental sounds, or there is some noise from the operation of recording equipment. The presence of these factors more closely matches and conforms to the recognition of birdsongs in real environments, and also improves the noise immunity of the model. Finally, we edited out 13,417 birdsong clips, and the pictures, numbers, and song characteristics of each category of birds in the dataset are shown in Table 1.

Although there are some natural ambient sounds in the original audio, the number is small and the manual production of the intercepted audio tends to make the dataset relatively pure due to subjective factors. To more closely match the real natural environment as well as to improve the complexity and usefulness of the dataset, we incorporated ten types of ambient sounds into the edited dataset, that is, each type of birdsong was classified as being in ten environments at a ratio of approximately-one-tenth. The ten ambient sounds incorporated include car horns, aircraft roar, wind, bark, cluck, knock, thunder, water, voice, and pitter-patter. The distribution is shown in Fig. 1, and the numbers of the horizontal coordinates correspond to the birdsong in Table 1. We also present a comparison of the differences before and after audio fusion in four of these environments in Fig. 2.

2.2. Correlation analysis

Birds of the same species with different groups separated by long distances may change their calls slightly, and over time eventually, a new language will emerge. This process is much like the process by which humans develop different accents, dialects, and languages. Both are learned through individual vocalizations and thus convey information, so it is necessary to explore the contextual relationship of birdsong. Following the study of the human auditory mechanism, it was found that the human ear has different auditory sensitivity to different frequencies of sound waves. Mel frequency is based on human auditory characteristics, which are nonlinearly related to frequency. MFCC uses this relationship between them to calculate spectral characteristics. Other commonly used speech features are Log-Mel, Mel-spectrogram, Contrast, Chroma, Tonnetz, etc.

The Pearson correlation coefficient is a statistical measure used to quantify the linear correlation between two variables and is primarily employed in deep learning for model performance evaluation and feature selection. Regarding feature selection, the Pearson correlation coefficient can be utilized to measure the linear correlation between features and target variables, thereby assessing feature importance.

In this study, Pearson Correlation Coefficient and Scatter plot matrix are used as evaluation metrics for feature maps to calculate the similarity between analyzed features and categories.

$$\rho_{X,Y} = \frac{covX, Y}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{1}$$

The correlation coefficient is calculated as in Eq.1, where ρ represents the correlation coefficient is the quotient of the covariance and standard deviation between the two variables, X, Y represents the two data, $\sigma_X\sigma_Y$ is the sample standard deviation, and $\mathrm{cov}(\mathrm{X},\ \mathrm{Y})$ calculates the covariance.

Firstly, we analyze the average correlation of speech features between the human voice and 15 types of birdsongs using Pearson correlation coefficients, and the results are shown in Table 2. Following the conclusion of the Pearson Correlation Coefficient: a significance level<0.05 (p-value $<$ 0.05) is a significant correlation, Pearson Correlation, also commonly referred to as R-value, in the case of confirmation of significance, the higher the correlation coefficient indicates the closer the relationship between the two. The level of significance is primary, and if it is not significant, the correlation coefficient is not meaningful and may be caused by chance only. The analysis of the results shows that MFCC can reflect a strong correlation between human speech and birdsong.

Second, we visualize the correlation between birdsongs and human voices using a scatter plot matrix as shown in Fig. 3. The histogram represents the kernel density estimation plot and is used to see the distribution of MFCC feature values, with the horizontal axis corresponding to the value of the feature and the vertical axis corresponding to the density of the feature. Two acoustic features are paired and plotted as coordinate points on a scatter plot to measure their correlation. This enables a graphical representation and analysis of the relationship between the acoustic features. The correlation scatters plot in Fig. 3 shows that the scatter points of birdsong and human voice are evenly distributed around the diagonal, and there is a good correlation.

Finally, we randomly select the audio in the dataset and use Pearson correlation coefficients to analyze the correlation between birdsongs disturbed by different sounds and birdsongs without disturbance. The analysis in Table 3 reveals that the feature correlation of MFCC is the highest in changing the background environment, indicating that the recognition of birdsongs in variable environments can be superior handled by using MFCC as a feature map.

2.3. A hierarchical birdsong feature extraction architecture combining static and dynamic modeling

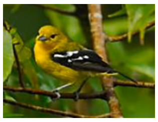
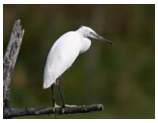
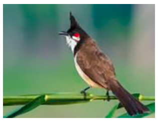
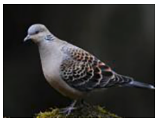
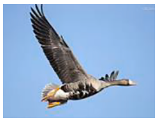
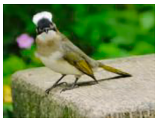
In this study, we analyze the characteristics of birdsong spectrograms in different environments and found that the proportion of the chirping part was small and not prominent, and there are many interference factors. And by using statistical knowledge to analyze the correlation between birdsongs and human speech, it was found necessary to explore the contextual relationship. In this paper, we propose a hierarchical birdsong feature extraction architecture combining dynamic and static modeling as shown in Fig. 4, which solves these problems.

The Transformer was originally used for Natural Language Processing tasks, so to convert the birdsong spectrogram into a word structure, the approach taken here is to split the image into small pieces, each of which is equivalent to a word in a sentence, and each piece is referred to here as a Patch. The proposed method differs from the main architecture of the traditional Transformer model by initially partitioning the spectrogram into a series of non-overlapping Patches, after which it is no longer linearly transformed into one-dimensional sequence data and embedded with location information, reducing the complexity of the model. The model is to slice the MFCC feature map obtained by preprocessing through a layer of convolution (Eq.2, X_{class} denotes trainable labels, X_P^N represents N patches with resolution P × P).

$$\mathrm{Z}_0 = \left[X_{\mathrm{class}} : X_P^1; X_P^2; \cdots; X_P^N\right] \tag{2}$$

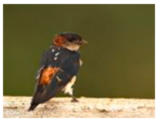
The convolution used for cutting feature maps has a larger convolution kernel and step size that distinguishes it from normal convolution. After input processing, the 2D data is directly imported into the feature extraction architecture CMS. The feature extraction is followed by the

Table 1
Introduction to the self-built dataset.

Scientific name	Image	Number	Voice features
Aegithalos caudatus		847	The call is weak and short, usually-two tones, sometimes several times in a row, with little variation.
Aegithina tiphia		844	The call consists of a single trill or a flute with a bursting tone that descends to an abrupt end.
Anas platyrhynchos		850	The call is loud and clear and could be heard far away.
Egretta garzetta		846	In the breeding nests croak, the rest of the time silent.
Eudynamys scolopaceus		852	The sound is noisy, crisp, and loud, usually higher and faster, and stops suddenly at the highest point.
Hirundo rustica		930	A sharp and urgent chirping sound.
Pycnonotus jocosus		919	Good at chirping, with a soft and pleasant sound.
Motacilla alba		904	The chirping sound is clear and loud.
Streptopelia chinensis		847	The call is loud, nods when whimpering, and will whimper repeatedly.
Zosterops japonicus		926	The call is crisp, the male has a tail note and will be relatively long, and the female is brief and powerful.
Anser albifrons		919	High and noisy cackling sound, flying with different scales of “lyo-lyok” pleasant call.
Pycnonotus sinensis		1036	The chirping sound is pleasant and varied.
Cuculus canorus bakeri		916	The chirping is loud, and the rough, monotonous sound can be heard from a great distance.
Dicrurus hottentottus		920	A pleasant loud and clear chirp, with occasional coarse and piercing calls.

(continued on next page)

Table 1 (*continued*)

Scientific name	Image	Number	Voice features
Cecropis daurica		915	The sharp and urgent chirping sound is slightly louder than that of the Hirundo rustica.

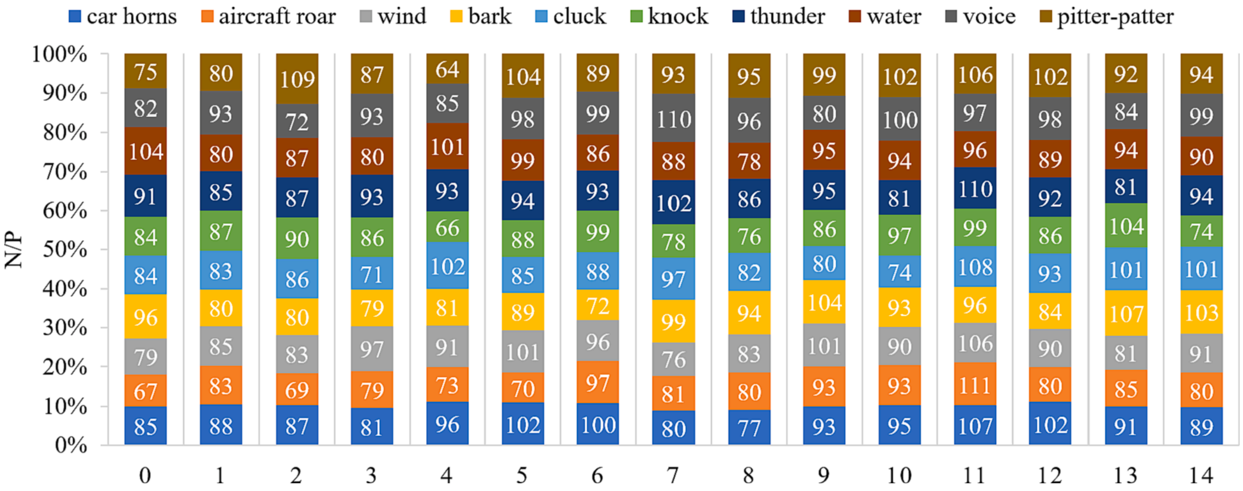


Fig. 1. Distribution of birdsong in different environments.

hierarchical refinement module, which reduces the resolution of the feature map to improve the training speed. (Eq.3, ι represents the current model depth, d represents the depth of the model, δ represents the FHS module depth).

$$Z_\iota = \mathrm{CMS\_H}_\iota \begin{cases} Z_\iota^1 = CLP(\mathrm{BN}(Z_{\iota-1})) \\ Z_\iota^2 = MLP(\mathrm{LN}(Z_\iota^1)) \\ Z_\iota^3 = SL(Z_\iota^2) \\ Z_\delta^4 = FHS(Z_\iota^3) \end{cases} \iota = 1\cdots d, \delta = 1\cdots d-1 \tag{3}$$

$$\mathrm{y} = \mathrm{GL}(Z_\iota^o) \tag{4}$$

The CMS block is cycled four times and the first three times go through the hierarchy refinement module. The model is finally input to the global average pooling layer and the linear layer for classification (Eq.4, Z_ι^o is the final output and y represents the classification result). Table 4 shows the pseudo-code for the model CMS-H. The details of the feature extraction module and the hierarchy refinement module are described in Section 2.3.1 and Section 2.3.2.

2.3.1. *Feature extraction architecture*

The birdsong spectrogram is a grayscale image characterized by time-series information, unlike RGB images that can clearly distinguish the components in the image. And there are problems in that the song features are not prominent, the noise interference is serious, and the biological spectrum information is not comprehensive. Birdsong is similar to human speech in that it conveys information through sound, and there is a certain correlation between sound fragments. The proposed feature extraction architecture combines static and dynamic modeling to fully explore the relationship between the context of birdsong and incorporates the ability to capture information at long and short distances with improved perception at both local and global levels.

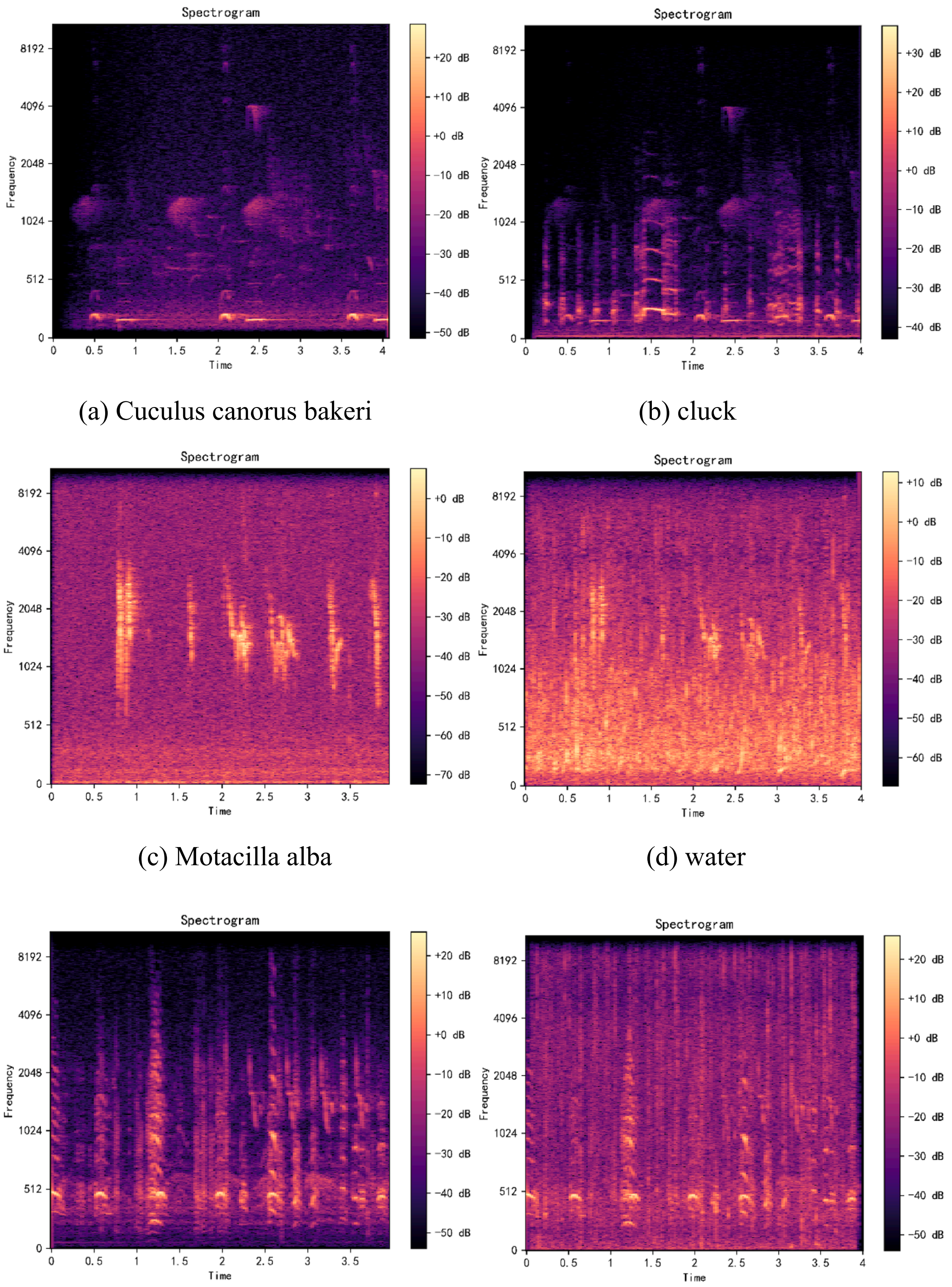
The input data needs to be normalized before target feature extraction, because the input data is after the convolution operation, and the normalization process makes the distribution of each feature similar. CoTAttention aggregates the mining of contextual information and traditional Self-Attention learning into a single structure, followed by a linear projection layer to convert two-dimensional image data into one-dimensional sequence data. Most traditional Transformer-based architectures act directly on the 2D feature map using independent query-key pairs to obtain the attention matrix, which ignores the rich context between keys. CoTAttention uses convolution to encode the context of the input key to generate a static context representation matrix (K^1). After that, it is concatenated with query to obtain attention map (A) after two 1 × 1 convolutions ($W_\theta W_\delta$). The dynamic context representation matrix (K^2) is generated by multiplying it with the mapped value. Finally, the static and dynamic contexts are fused as the output (Li et al., 2022).

$$\mathrm{A} = [\mathrm{K}^1, \mathrm{Q}]\mathrm{W}_\theta \mathrm{W}_\delta \tag{5}$$

$$\mathrm{K}^2 = \mathrm{V} \otimes \mathrm{A} \tag{6}$$

The learning rate needs to be chosen concerning the size of the input layer values, while the data normalization operation makes it easy to choose the learning rate. We propose to use a combination of batch normalization and layer normalization before and after CoTAttention. This normalizes the elements inside both channels and samples, which makes it easier to learn the optimal parameters and makes the model training converge more smoothly.

Secondly, the feedforward network of the main architecture uses a single hidden layer MLP, which contains multiple layers of nodes and the connections of nodes in adjacent layers are equipped with weights, and the purpose of learning is to assign the correct weights to the edge features in the birdsong feature map (Eq. (7), w_k is the weight, b_l is the bias). Highly parallel processing and nonlinear global action can be obtained by using MLP. And for single-channel images, there is no need to add more hidden layers to improve the perceptron's capability. Otherwise, it will increase the number of parameters and also bring the problem of long feedback time, resulting in a longer training time and lower efficiency. The perceptron is a linear regression network that brings overfitting problems and poor generalization ability. Adding Gaussian error linear units to the hidden layer increases the nonlinear factor of the neural network model and improves the sparsity of the



(a) Cuculus canorus bakeri

(b) cluck

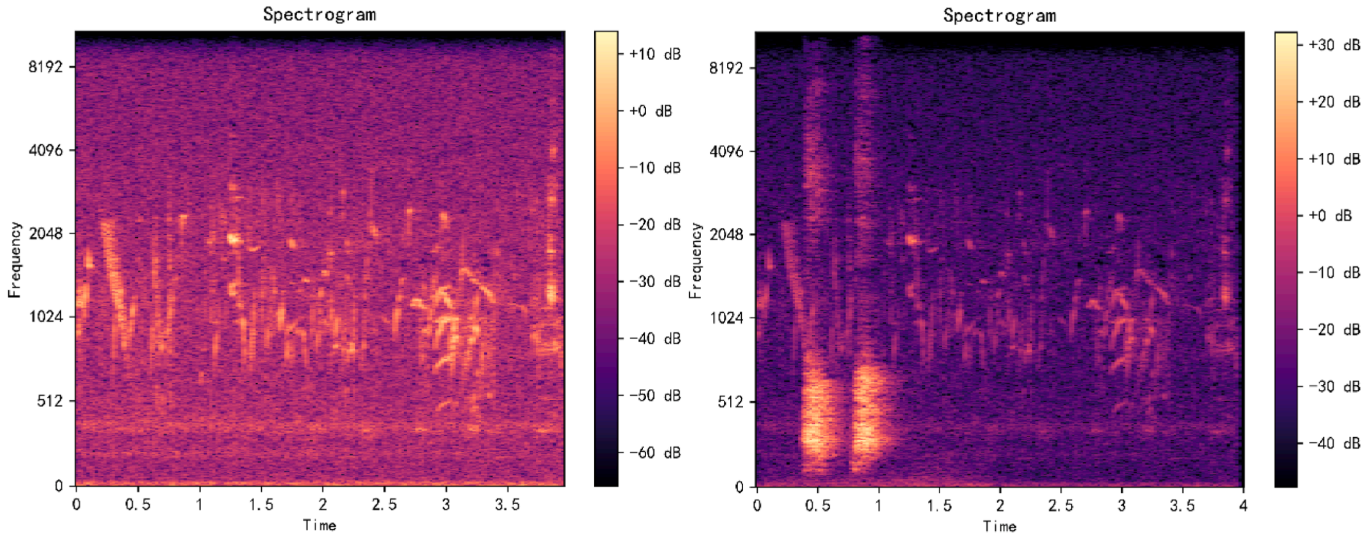
(c) Motacilla alba

(d) water

(e) Anser albifrons

(f) pitter-patter

Fig. 2. Spectrogram comparison before and after fusion.



(g) Cecropis daurica

(h) bark

Fig. 2. (*continued*).

Table 2
Comparison of human voice correlation analysis.

Feature	Pearson Correlation R	Significance Level P
MFCC	0.8828	<0.001
Mel-spectrogram	−0.0131	0.421
Contrast	−0.0926	2.431
Chroma	0.1107	1.822
Log-Mel	0.0467	2.277
Tonnetz	−0.1710	0.897

model to avoid gradient disappearance. As well as adding regularization to avoid overfitting and improve the generalization ability of the model. Finally, one-dimensional sequence data is transformed into two-dimensional image data by stretching layer.

$$\begin{cases} w_k \to w_k^{'} = w_k - \dfrac{\eta}{m}\sum_j \dfrac{\partial C_{X_j}}{\partial w_k} \\ b_l \to b_l^{'} = b_l - \dfrac{\eta}{m}\sum_j \dfrac{\partial C_{X_j}}{\partial b_l} \end{cases} \tag{7}$$

2.3.2. Hierarchy refinement module

Transformer models generally have high complexity, require a long training time, and rely on pre-trained visual models to improve performance, which limits scalability in audio tasks. For this reason, we design a layer refinement module FHS behind the feature extraction architecture to reduce the model complexity and training time by halving the feature map resolution layer by layer while increasing the dimensionality to preserve the extracted features. The algorithm for refining the feature map is a convolution with a convolution kernel size of 2 × 2 and a step size of 2. The initial Patch size used in the model is 4 × 4, and 3136 Patches are obtained by dividing the feature map at a resolution of 224 × 224, which means that the feature map is transformed into a size of 56 × 56 as the initial input (Eq. (8), H_{in}, W_{in} are the sizes of the input feature map, P_h, P_w are the Patch sizes).

$$\text{Patches} = \frac{H_{in}}{P_h} \times \frac{W_{in}}{P_w} \tag{8}$$

After three layers of feature extraction and a hierarchical refinement module, the feature map resolution is reduced to 7 × 7, and the last layer does not need to slice the extracted feature map again. This hierarchical extraction approach helps to capture more features and save them inside more channels. It also optimizes the computational cost problems that exist in Transformer-type models that require many training data and are generated by high complexity.

3. Results and analysis

3.1. Experimental environment

To verify the validity of the self-built dataset and the generalization of the model, we conduct experiments on three datasets. The main experiments rely on the self-built birdsong dataset BirdSound15, whereas the public datasets Birdsdata and Urbansound8k are used for generalization experiments. To ensure a fair comparison across experiments, only simple fine-tuning of model parameters was performed. Each experiment randomly disrupted the trained samples, and both divided the training and test sets under the ratio of 8:2. We set the batch size to 16, set the initial learning rate to 0.01, and use the same optimizer stochastic gradient descent (SGD) as well as the learning rate decay strategy exponential decay. The experiments were also conducted using the same original parameter settings to ensure the authority of the original settings of the parameters in the comparison model in the proposed paper. To make training faster and take up less computational space, we use mixed precision for training to improve computational efficiency without sacrificing precision. This study uses Pytorch 1.11.0 framework to build the network model, and the programming environment is python 3.9. The hardware environment is CPU: Intel i9, GPU: NVIDIA GeForce RTX 3090.

3.2. Model performance index

We measure the complexity of the model by using the params and the Floating-Point Operations (FLOPs). Where Params is used to describe the size of the model, similar to the space complexity in algorithms. Params are only relevant to the defined network structure and are calculated as shown in Eqs. (9)–(10).

$$Params = (K_h \times K_w \times C_{in} + 1) \times C_{out} \tag{9}$$

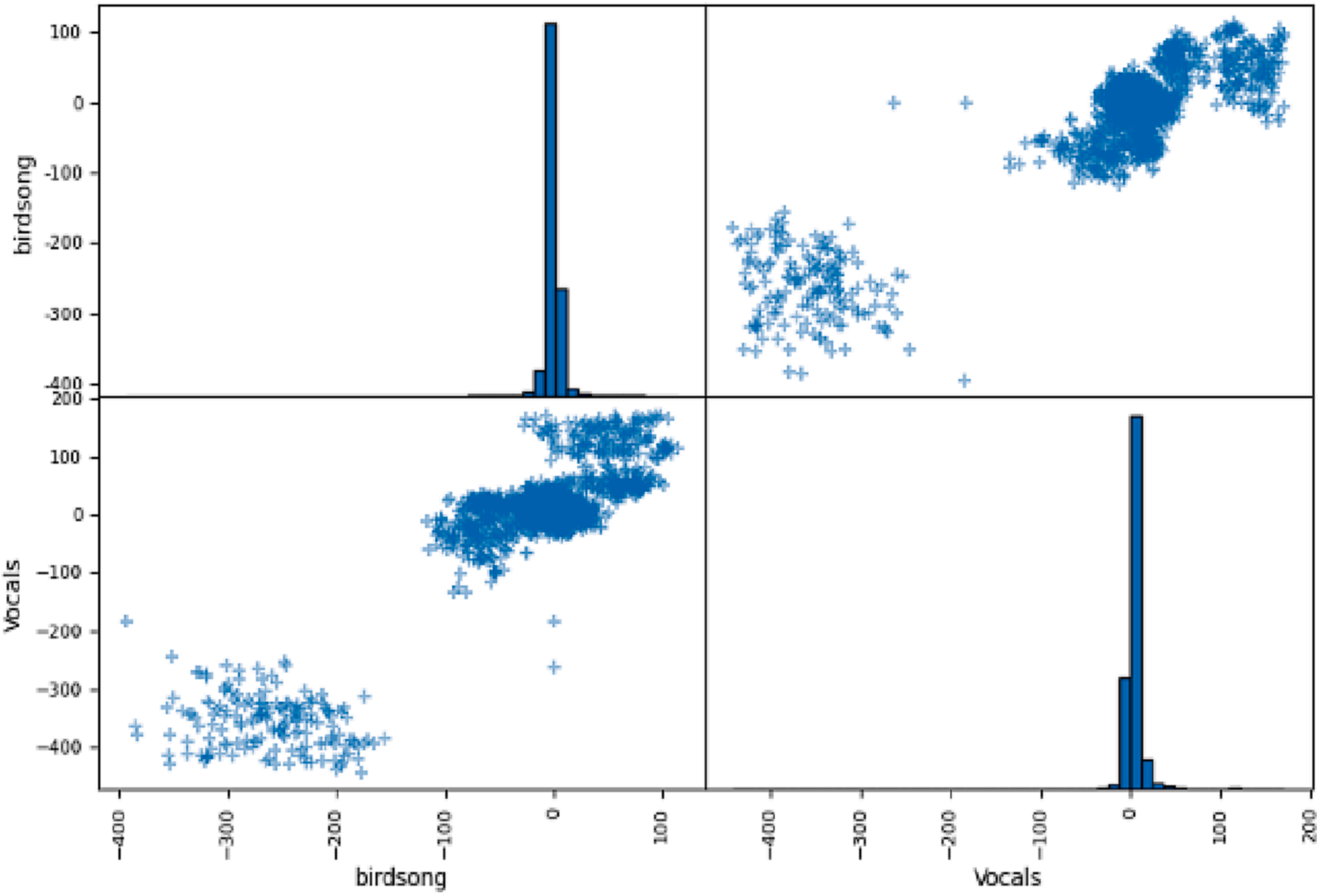


Fig. 3. Scatter plot matrix.

Table 3
Feature contrast selection.

Feature	Environmental Sound		Voice	
	R	P	R	P
MFCC	0.9604	<0.001	0.9536	<0.001
Mel-spectrogram	0.9339	<0.001	0.9192	<0.001
Contrast	0.6715	1.483	0.8714	<0.001
Chroma	0.3866	5.658	0.7438	<0.001
Log-Mel	0.3131	<0.001	0.6873	<0.001
Tonnetz	0.0839	0.007	0.1733	1.938

$$Params = (C_{in} + 1) \times C_{out} \tag{10}$$

For the convolutional layer in Eq. (9), K_h, K_w represent the length and width of the convolutional kernel, respectively, and C_{in}, C_{out} represent the number of input and output channels, respectively. For the fully connected layer as in Eq.10, C_{in}, C_{out} denote the number of nodes in the input and output respectively, and the convolution kernel with bias term is added by 1.

FLOPs are used to describe the execution efficiency of a model and measure the complexity of the model. It is similar to the time complexity in algorithms and is often used as an indirect measure of the speed of a neural network model. Taking the convolutional layer as an example without considering the activation function layer, the FLOPs are calculated as follows:

$$FLOPs = \left(2 \times C_{in} \times K^2 - 1\right) \times H \times W \times C_{out} \tag{11}$$

$$FLOPs = (2I - 1)O \tag{12}$$

K is the size of the convolution kernel, *H, and W* represents the size of the output feature map, C_{in}, C_{out} represent the number of input and output channels respectively. *I, O* represents the number of input and output dimensions.

To evaluate the performance of the method, we employed four commonly used classification evaluation metrics, namely accuracy, precision, recall, and F1 score.

3.3. *Model parameter setting*

The number of parameters of the model is mainly determined by the predefined input and output channel dimensions, as well as the number of invoked modules. We tuned the model by changing the input and output dimensions of the hierarchical refinement module. As shown in Table 5, there is a subtle effect of adjusting the dimension size on the results, and the optimal dimension setting is [64,128,256,512]. Then we set the dimensions to the best combination and transform the resolution of the layer division feature map for the experiment. The results are shown in Table 6, and the model works best when the Patch size is initially set to 4 × 4, which means the feature map resolution is 56 × 56.

3.4. *Ablation experiment*

In the second part of this study, the correlation analysis of six features common to speech is performed, and it is concluded that birdsong is suitable for using MFCC as an input feature for recognition. Table 7 again demonstrates experimentally that MFCC is the best for the recognition of birdsongs in complex environments. On this basis, we conducted ablation experiments on the model and the results are shown in Table 8. The results show a difference of 2 ± 0.05 % between the best results using only the feature extraction module and the hierarchical

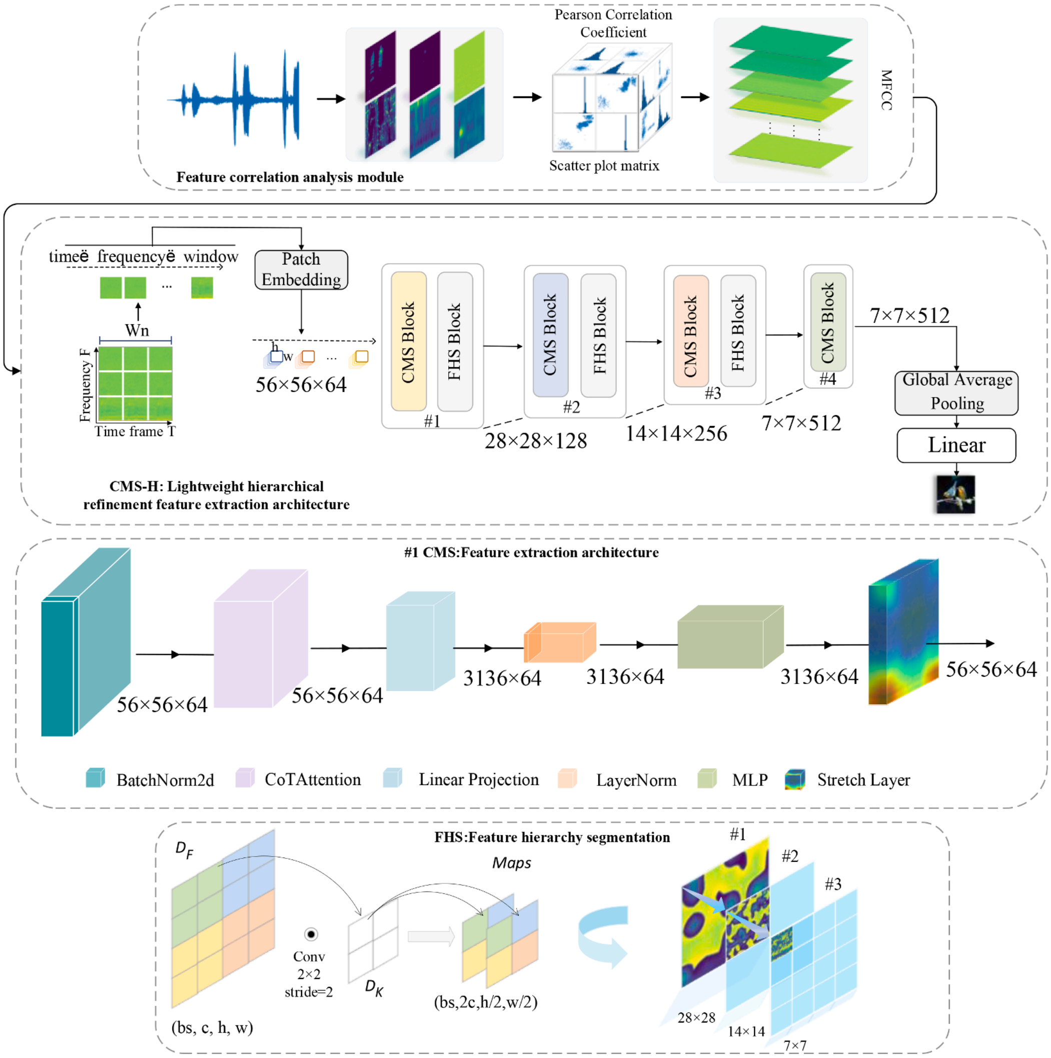


Fig. 4. Schematic diagram.

Table 4
Pseudo-code of CMS-H.

```
Algorithm: CMS-H, is our proposed algorithm.
Require: C, the number of channels. P, the patch size. N, the total number of patches.
  D, the embedded dimension. GL, the average pooling, and linear classification layer.
  Input: Self-built birdsong dataset Birdsound15
  Output: Linear classification result
Feature extraction and selection:
  MFCC ← Pearson Correlation Coefficient {MFCC, Log-Mel,
  Contrast, Mel-spectrogram, Chroma, Tonnetz}
  MFCC to Patch Embedding:
  $Z_0 \leftarrow \left[X_{class} X_P^1; X_P^2; \cdots; X_P^N\right] \leftarrow Conv(MFCC)$
  for $\iota = 1$ to 4 do
  if $\iota \geq 3$ then
  $Z_\iota \leftarrow \mathrm{CMS\_H}_\iota(Z_{\iota-1})$
  $Z_\iota^{'} \leftarrow$ FHS $(Z_\iota)$
  else
  $Z_\iota \leftarrow \mathrm{CMS\_H}_\iota(Z_{\iota-1})$
  end
  return$y = \mathrm{GL}\left(Z_\iota^o\right)$
```

model using also the layer refinement module after adjusting the number of modules. The introduction of a hierarchical refinement feature map approach can coordinate the feature extraction architecture to extract more global and local birdsong features. The ablation experiments also confirm that increasing the number of modules does not improve the feature extraction capability of the model.

Table 5
Comparison of model adjustment dimensions.

Channels	FLOPs(G)	Params(M)	Accuracy (%)
[32,64,128,256]	0.15	1	92.48
[64,128,256,512]	0.53	2.9	93.67
[96,192,384,768]	1.2	9.3	93.19

Table 6
Comparison of model adjustment feature map size.

Feature Size	FLOPs(G)	Params(M)	Accuracy (%)
[112,56,28,14]	2.1	2.9	93.41
[56,28,14,7]	0.53	2.9	93.67
[32,16,8,4]	0.17	2.9	89.89
[16,8,4,2]	0.05	2.9	84.01

3.5. *Comparison experiments*

In this section, we compare the performance of the proposed method with the current mainstream models on a self-built dataset. And a detailed performance analysis of the proposed method is conducted to explore the effectiveness of the method. The results are shown in

Table 7
Experimental comparison of six features.

Features	Birdsound15
MFCC	93.67 %
Mel-spectrogram	85.04 %
Contrast	65.07 %
Chroma	78.22 %
Log-Mel	92.02 %
Tonnetz	–

Table 8
Ablation experiment.

Model	Blocks	FLOPs(G)	Params(M)	Accuracy (%)
CMS	4	4.6	5.9	91.69
	6	6.8	9	89.96
	10	11.4	14.6	88.22
CMS-H	[1,1,1,1]	0.53	2.9	93.67
	[1,2,2,1]	0.78	3.7	92.19
	[2,3,3,2]	1.2	6	91.33

Table 9
Comparison of our method with other methods on self-constructed datasets.

Model	Pre-trained	FLOPs(G)	Params(M)	Accuracy (%)
EfficientNet-B7	√	5.3	66.4	92.48
(Zhang et al., 2019)	×	7.6	24.2	85.31
(Kong et al., 2020)	√	30	81	91.26
ViT-B/16	√	55.4	85	52.21
ViT-L/16	√	190.7	303	65.70
AST-S	√	8.7	22.3	85.41
AST-B	√	34.7	87	83.78
AST-P	√	38	87	83.52
CMS-H	×	0.53	2.9	93.67

Table 9, where the proposed method achieves the best recognition of birdsongs in complex environments, outperforming the EfficientNet and Transformer audio classification models AST using a mixed model scale approach. The results also show the high performance of the proposed method with a small number of parameters and FLOPs.

3.5.1. Analysis of results

For the dataset Birdsound15, birdsong is in at least one environment, and there are birds of different species of the same genus which make identification difficult. The complex environment challenges the model's ability to extract features even more, and the results from Table 9 show that our method can cope with these challenges superior to other methods. We also give the confusion matrix generated by the proposed method for the identification of 15 birdsongs as shown in Fig. 5, and more experimental details metrics as shown in Table 10. The comprehensive analysis shows that our method also has recognition errors sometimes where the interference is high or the birdsongs are not prominent. We also show the correct test results of the models EfficientNet, AST, and CMS-H on the self-built dataset, as shown in Fig. 6 for the proposed method with a higher linear straight line than the comparison model.

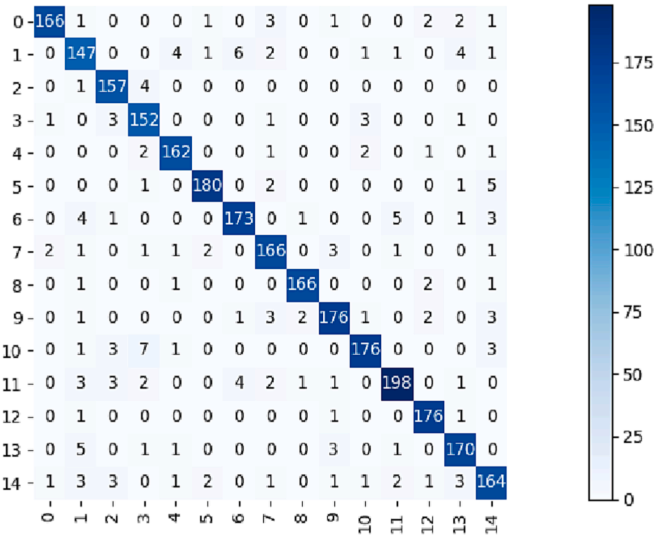


Fig. 5. Confusion Matrix.

In this section, we also used the heat map to judge how well the representative model assigns weights to the target features. The heat map is obtained by averaging the gradient maps, obtaining a scalar as a weight and multiplying it with the corresponding feature maps separately, and then summing all the results. We randomly selected two birdsong audios containing tapping and wind sounds in the dataset, and the spectrogram is shown in Fig. 7(a), with the birdsong sound in the red box. The model parameters obtained by pre-training on the spectrogram are then loaded, and the visualization results are shown in Fig. 7. We judge the performance of the model by looking at the heat map of the model's weight assignment to each component of the spectrogram. It can be seen from the figure that the proposed hierarchical model has a superior weight assignment mechanism, putting more attention on the critical regions.

3.5.2. Intra-class birdsongs test analysis

To verify the ability of the proposed method to capture spatial detail information, we added the analysis of intra-class birdsong recognition, and the results are shown in Fig. 8. From the test results of the three models in the two intra-class birdsongs, CMS-H performed the best in testing the correct identification of the two intra-class birds, which was significantly better than the comparison model. And it can be seen from Fig. 8(a) that CBS-H has the lowest number of mismeasurements for both intraclass birds, while AST highlights the problems that exist, ignoring local spatial information and having more mismeasurements. In a comprehensive analysis, the method of extracting features by hierarchical refinement and combining dynamic and static modeling can focus on more locally detailed spatial information.

3.6. Generalization experiments

In this section, we use the public birdsong and environmental sound datasets for generalization experiments to test the generalization and versatility of the model and also to illustrate the effectiveness of our self-

Table 10
Detailed indicators.

Categories	Precision (%)	Recall (%)	F1 (%)
Aegithalos caudatus	92.11	96.47	94.24
Aegithina tiphia	87.94	86.44	87.18
Anas platyrhynchos	96.97	94.13	95.53
Egretta garzetta	93.45	92.36	92.90
Eudynamys scolopaceus	94.74	94.75	94.74
Hirundo rustica	95.63	94.16	94.89
Malacorhynchus membranaceus	89.84	91.31	90.57
Motacilla alba	89.89	88.41	89.14
Streptopelia chinensis	96.43	95.32	95.87
Zosterops japonicus	91.76	89.79	90.76
Anser albifrons	96.63	93.48	95.03
Pycnonotus sinensis	90.74	94.72	92.69
Cuculus canorus bakeri	96.13	94.57	95.34
Dicrurus hottentottus	92.06	89.59	90.81
Cecropis daurica	89.92	91.26	90.59

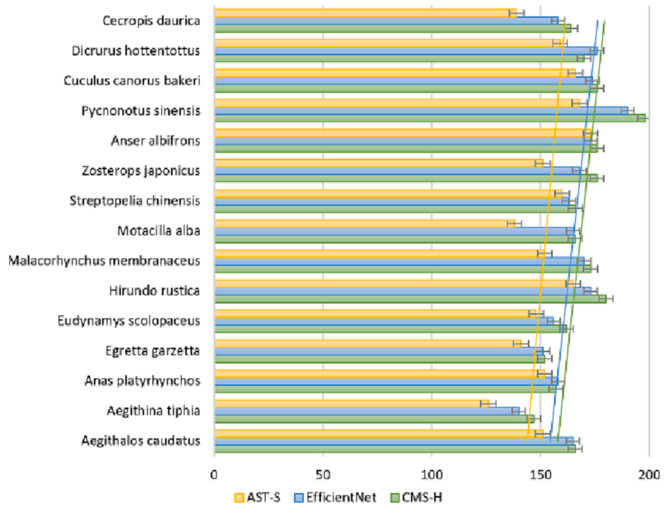
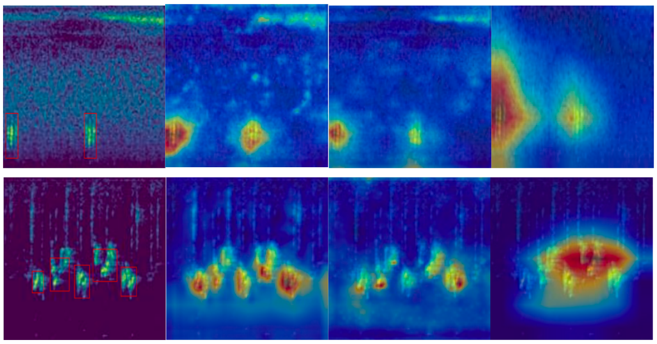


Fig. 6. Comparison of test results.



(a)Input (b)CMS-H (c)AST (d)EfficientNet

Fig. 7. Comparison of model weight assignment.

built datasets.

(1) The Birdsdata dataset (public portion) contains 14,311 nature audio tracks, each of 2 s duration, for a total of 20 species of birds commonly found in China.

(2) Urbansound8k is a widely used public dataset for urban environmental sound classification studies, containing 8732 audios of 10 categories of environmental sounds, each in about 4 s.

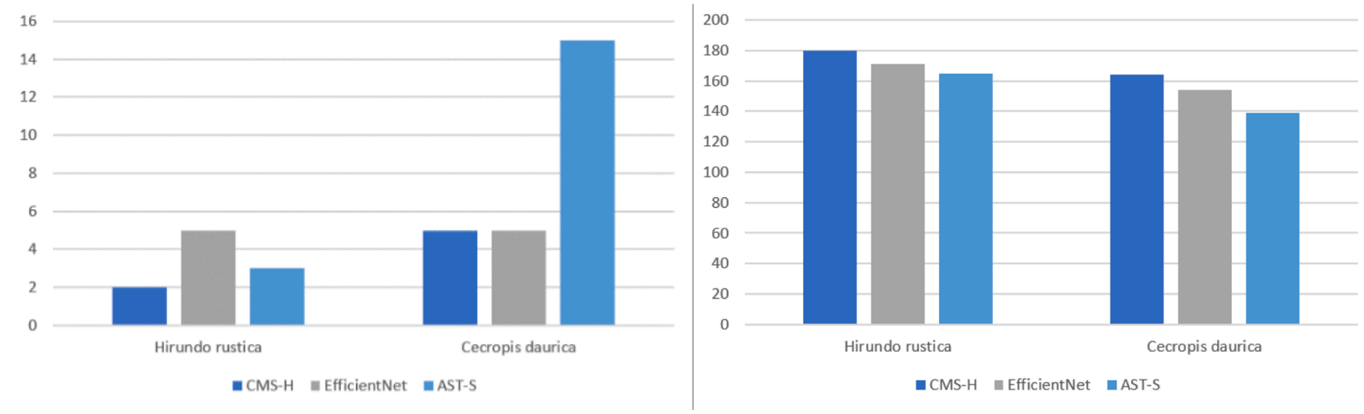
The experimental results on Birdsdata are shown in Table 11. The proposed method in this paper continues to show superior results compared to other methods, achieving an accuracy of 95.19 % on the test set. The experimental results also illustrate that the performance of the recently proposed Transformer audio classification model is inferior to that of the CNN model in the task of recognition of birdsongs. In contrast, the method proposed in this study compensates for the problem of missing some spatial information in the Transformer model and demonstrates the good performance and generalization ability of the lightweight model. We also show the experiments on Urbansound8k and the results are shown in Table 12. From the experimental results, we can see that EfficientNet achieves similar results to the proposed method in this paper both achieving 97 % ± 0.5 accuracy on the test set, but the comprehensive performance is not as good as our proposed method. The

Table 11
Generalization experiments on Birdsdata.

Model	Pre-trained	FLOPs(G)	Params(M)	Accuracy (%)
EfficientNet-B7	√	5.3	66.4	94.49
(Zhang et al., 2019)	×	7.6	24.2	87.65
(Kong et al., 2020)	√	30	81	94.34
(Liu et al., 2021)	×	–	–	92.20
AST-S	√	8.7	22.3	89.36
AST-B	√	34.7	87	89.85
AST-P	√	38	87	88.14
CMS-H	×	0.53	2.9	95.19

Table 12
Generalization experiments on Urbansound8k.

Model	Pre-trained	FLOPs (G)	Params (M)	Accuracy (%)
EfficientNet-B7	√	5.3	66.4	96.74
(Zhang et al., 2019)	×	7.6	24.2	91.65
(Su et al., 2020)	×	–	11.3	93.4
(Mushtaq and Su, 2020)	×	–	3.2	95.3
EAT-M	×	–	25.5	90
AudioCLIP	×	–	–	90.07
AST-S	√	8.7	22.3	94.28
AST-B	√	34.7	87	94.79
AST-P	√	38	87	94.39
CMS-H	×	0.53	2.9	97.02



(a) Number of misclassifications (b) Number of correct tests

Fig. 8. Intra-class test results.

audio classifier AST also shows good recognition of environmental sounds, reaching the performance of most CNN models. In contrast, some recently proposed audio classification models (EAT-S (Gazneli et al., 2022), AudioCLIP (Guzhov et al., 2022)) have a general performance on Ubransound8k, which is not satisfactory as AST. The combined performance of these two datasets fully validates the performance of our method and the effectiveness of the self-built dataset. It also shows that our proposed method has good generalizability and can be widely used in birdsong and environmental sound recognition tasks.

4. Discussion

From the experimental performance on birdsong and environmental sound datasets, the proposed method in this study can be widely applied. However, this method has not been trained on large datasets yet, and direct application on small datasets will generate overfitting problems. The experimental results of the proposed method on two small datasets are shown in Table 13. Among them, ESC-50 is an environmental sound dataset, which contains 50 semantic classes, each containing 40 audios 5 s long. Birdsmall is a self-built small birdsong dataset, containing 900 audios of 15 bird species in North China. The experimental results show that the proposed method has an overfitting problem when dealing with small datasets, while the proposed method still has potential for improvement by analyzing the training situation. The overfitting problem will be corrected by data enhancement and migration learning in the future, while simple and reliable strategies will be designed to improve the accuracy of recognition and the performance of the algorithm.

The spatial distribution of bird species is also a significant subject of ecological research, yet our understanding in this area is currently limited. To improve our understanding of bird distribution patterns, we plan to employ Variograms, combined with birdsong recognition, to conduct our next analysis. This visualization and quantification tool can be employed to explore spatial data distribution patterns and spatial autocorrelation (Al, 2000; Zawadzki et al., 2005; T. et al., 2016). At present, we are establishing a real-time monitoring system in local forest areas. In future studies, we plan to input the collected bird observations as spatial data into the variance function model. Within this model, we will compute the semivariance values between any two bird observation points at different spatial distances, creating a semivariogram graph that increases with distance. This graph will reflect the spatial correlation strength between bird observation points, which can be used to determine the spatial distribution patterns of bird populations. Furthermore, we can explore the relationships between these patterns and environmental factors.

5. Conclusion

In this paper, we propose a simple and efficient hierarchical model CMS-H to cope with the recognition of birdsong in complex background noise. Compared with previous methods, we combine statistical knowledge to analyze the correlation of six features of birdsong using the Pearson correlation coefficient and scatter plot matrix to select features suitable for birdsong recognition. Based on the analysis, we design the feature extraction architecture CMS, which incorporates dynamic and static modeling to obtain the dependency relationship between birdsong contexts. The insensitivity of the Transformer model to the structural information inside the region and the problem of missing spatial information is solved. In the experiments, we found that using only stacked feature extraction architectures does not cope well with complex environments. Therefore, we design a hierarchical refinement module to reduce the model size while extracting the subtle features faster. The experimental results show that the proposed method performs well in the recognition of birdsongs in complex environments and achieves advanced experimental results.

In practical applications, our method can identify birds and environmental sounds, and then study the distribution in the area to explore its rich ecological information. Especially for endangered bird species, we can provide more information for timely detection and conservation. Through the discussion and analysis of the experiments, the proposed method still has certain shortcomings. For small sample datasets, the model cannot learn sufficient feature space structure to easily generate overfitting problems. The follow-up research work will focus on extending the parameters of the network to improve the fitting ability to cope with the task of detecting and recognizing small sample species sounds. Also consider using the migratory learning capability of the model to learn sufficient feature space hierarchy to improve the generality of the model.

Table 13
Small sample dataset experiment.

Model	Birdsmall		ESC-50	
	Train	Test	Train	Test
Proposed Method	100 %	70.0 %	100 %	71.15 %

CRediT authorship contribution statement

Yanan Wang: Investigation, Writing – original draft. **Aibin Chen:** Funding acquisition, Supervision. **Huaicheng Li:** Data curation, Investigation. **Guoxiong Zhou:** Methodology, Validation. **Jizheng Yi:** Software. **Zhiqiang Zhang:** Formal analysis.

Declaration of Competing 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.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no. 62276276), the Hunan Provincial Natural Science Foundation of China (Grant no. 2022JJ31022), and the Undergraduate Education Reform Project of Hunan Province (Grant no. HNJG-2021-0532).

References

Á, I., Jancsó, H., Szilágyi, Z., Farkas, A., Sulyok, C., 2018. Bird Sound Recognition Using a Convolutional Neural Network. In: In 2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY), pp. 000295–000300. https://doi.org/10.1109/SISY.2018.8524677.

Al, C., 2000. Modern Spatiotemporal Geostatistics. Modern spatiotemporal geostatistics.

Brooker, S.A., Stephens, P.A., Whittingham, M.J., Willis, S.G., 2020. Automated detection and classification of birdsong: An ensemble approach. Ecol. Ind. 117 https://doi.org/10.1016/j.ecolind.2020.106609.

Cinkler, T., Nagy, K., Simon, C., Vida, R., Rajab, H., 2022. Two-Phase Sensor Decision: Machine-Learning for Bird Sound Recognition and Vineyard Protection. IEEE Sens. J. 22, 11393–11404. https://doi.org/10.1109/jsen.2021.3134817.

Conde, M.V., Shubham, K., Agnihotri, P., Movva, N.D., Bessenyei, S., 2021. Weakly-Supervised Classification and Detection of Bird Sounds in the Wild. A BirdCLEF 2021 Solution. arXiv preprint arXiv:2107.04878. 10.48550/arXiv, 2107.04878.

Dai, Y.S., Yang, J., Dong, Y.W., Zou, H.P., Hu, M.Z., Wang, B., 2021. Blind source separation-based IVA-Xception model for bird sound recognition in complex acoustic environments. Electron. Lett 57, 454–456. https://doi.org/10.1049/ell2.12160.

Denton, T., S. Wisdom, and J. R. Hershey. 2022. Improving Bird Classification with Unsupervised Sound Separation. Pages 636-640 *in* ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 10.48550/arXiv.2110.03209.

Dieleman, S., Schrauwen, B., 2014. End-to-end learning for music audio. In: *in* 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6964–6968. https://doi.org/10.1109/ICASSP.2014.6854950.

Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and S. Gelly. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv: 2010.11929. 10.48550/arXiv.2010.11929.

Fagerlund, and Seppo., 2007. Bird Species Recognition Using Support Vector Machines. Eurasip Journal on Advances in Signal Processing 2007, 038637. https://doi.org/10.1155/2007/38637.

Farwell, L.S., Gudex-Cross, D., Anise, I.E., Bosch, M.J., Olah, A.M., Radeloff, V.C., Razenkova, E., Rogova, N., Silveira, E.M.O., Smith, M.M., Pidgeon, A.M., 2021. Satellite image texture captures vegetation heterogeneity and explains patterns of bird richness. Remote Sens. Environ. 253 https://doi.org/10.1016/i.rse.2020.112175.

Ganatsas, P., Tsakaldimi, M., Oikonomakis, N., Davis, M., Manios, C., Broumpas, C., 2022. Reduction, degradation and restoration of Salix alba habitat in the Kerkini National Park, northern Greece; an important habitat for endangered bird species. Ecol. Eng. 179 https://doi.org/10.1016/j.ecoleng.2022.106593.

Gazneli, A., Zimerman, G., Ridnik, T., Sharir, G., Noy, A., 2022. End-to-End Audio Strikes Back. Boosting Augmentations Towards An Efficient Audio Classification Network. arXiv e-prints. 10.48550/arXiv, 2204.11479.

Gong, Y., Chung, Y.-A., Glass, J., 2021. AST: Audio spectrogram transformer. Interspeech. 10.48550/arXiv, 2104.01778.

Guo, J., K. Han, H. Wu, Y. Tang, X. Chen, Y. Wang, and C. Xu. 2022. CMT: Convolutional neural networks meet vision transformers. Pages 12175-12185 *in* Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 10.48550/arXiv.2107.06263.

Guzhov, A., Raue, F., Hees, J., Dengel, A., 2022. Audioclip: Extending Clip to Image, Text and Audio. In: in ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 976–980. https://doi.org/10.1109/ICASSP43922.2022.9747631.

Hussain, M., J. J. Bird, and D. R. Faria. 2018. A study on cnn transfer learning for image classification. Pages 191-202 *in* UK Workshop on computational Intelligence. Springer. 10.1007/978-3-319-97982-3_16.

Jaitly, N., Hinton, G., 2011. Learning a better representation of speech soundwaves using restricted boltzmann machines.*in* 2011. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.1109/ICASSP.2011.5947700.

Jiang, H., Qiao, Q., Zheng, H., Wang, R., Zhu, H., 2021. Birdsong Recognition Based on Improved DTW. J. Phys. Conf. Ser. 1739, 012038 https://doi.org/10.1088/1742-6596/1739/1/012038.

Kalan, A.K., Mundry, R., Wagner, O.J.J., Heinicke, S., Boesch, C., Kuhl, H.S., 2015. Towards the automated detection and occupancy estimation of primates using passive acoustic monitoring. Ecol. Ind. 54, 217–226. https://doi.org/10.1016/j.ecolind.2015.02.023.

Kim, J.-h., J.-w. Jung, H.-j. Shim, and H.-j. Yu. 2020. Audio Tag Representation Guided Dual Attention Network for Acoustic Scene Classification. Pages 76-80 *in* DCASE.

Kitaev, N., Kaiser, U., Levskaya, A., 2020. Reformer. The Efficient Transformer. 10.48550/arXiv, 2001.04451.

Kong, Q.Q., Cao, Y., Iqbal, T., Wang, Y.X., Wang, W.W., Plumbley, M.D., 2020. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. Ieee-Acm Transactions on Audio Speech and Language Processing 28, 2880–2894. https://doi.org/10.1109/taslp.2020.3030497.

Kumar, A., and S. D. Das. 2019. Bird Species Classification Using Transfer Learning with Multistage Training. Pages 28-38 *in* Computer Vision Applications. Springer Singapore, Singapore. 10.1007/978-981-15-1387-9_3.

Lee, C.H., Hsu, S.B., Shih, J.L., Chou, C.H., 2013. Continuous Birdsong Recognition Using Gaussian Mixture Modeling of Image Shape Features. IEEE Trans. Multimedia 15, 454–464. https://doi.org/10.1109/tmm.2012.2229969.

Li, Y., Yao, T., Pan, Y., Mei, T., 2022. Contextual Transformer Networks for Visual Recognition. In: IEEE transactions on pattern analysis and machine intelligence pp. https://doi.org/10.1109/tpami.2022.3164083.

Liu, H., Liu, C., Zhao, T., Liu, Y., 2021. Bird Song Classification Based on Improved Bi-LSTM-DenseNet Network. In: *in* 2021 4th International Conference on Robotics, Control and Automation Engineering (RCAE). IEEE, pp. 152–155. https://doi.org/10.1109/RCAE53607.2021.9638962.

Mehyadin, A.E., Abdulazeez, A.M., Hasan, D.A., Saeed, J.N., 2021. Birds sound classification based on machine learning algorithms. Asian J Res Comput Sci:1–11. https://doi.org/10.9734/AJRCOS/2021/v9i430227.

Morita, T., Koda, H., Okanoya, K., Tachibana, R.O., 2021. Measuring context dependency in birdsong using artificial neural networks. PLoS Comput. Biol. 17, e1009707.

Mushtaq, Z., Su, S.F., 2020. Environmental sound classification using a regularized deep convolutional neural network with data augmentation. Appl. Acoust. 167 https://doi.org/10.1016/j.apacoust.2020.107389.

Pahuja, R., Kumar, A., 2021. Sound-spectrogram based automatic bird species recognition using MLP classifier. Appl. Acoust. 180 https://doi.org/10.1016/j.apacoust.2021.108077.

Peng, Y.X., He, X.T., Zhao, J.J., 2018. Object-Part Attention Model for Fine-Grained Image Classification. IEEE Trans. Image Process. 27, 1487–1500. https://doi.org/10.1109/tip.2017.2774041.

Rao, Y., Zhao, W., Zhu, Z., Lu, J., Zhou, J., 2021. Global Filter Networks for Image Classification. Page arXiv:2107.00645. 10.48550/arXiv, 2107.00645.

Su, Y., Zhang, K., Wang, J.Y., Zhou, D.M., Madani, K., 2020. Performance analysis of multiple aggregated acoustic features for environment sound classification. Appl. Acoust. 158 https://doi.org/10.1016/j.apacoust.2019.107050.

T., Subba, and Rao. 2016. Statistics for Spatial Data, Revised Edition, by Noel Cressie. Published by Wiley Classics Library, John Wiley, 2015. Total number of pages: 928. ISBN: 978-1-119-11518-2. Journal of Time Series Analysis 37:288-288. 10.1111/jtsa.12168.

Tan, L.N., Alwan, A., Kossan, G., Cody, M.L., Taylor, C.E., 2015. Dynamic time warping and sparse representation classification for birdsong phrase classification using limited training data. J. Acoust. Soc. Am. 137, 1069–1080. https://doi.org/10.1121/1.4906168.

Tao, Q., Gao, G.H., Xi, H.H., Wang, F., Cheng, X.B., Ou, W.X., Tao, Y., 2022. An integrated evaluation framework for multiscale ecological protection and restoration based on multi-scenario trade-offs of ecosystem services: Case study of Nanjing City, China. Ecological Indicators 140. https://doi.org/10.1016/j.ecolind.2022.108962.

Trigeorgis, G., F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou. 2016. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. Pages 5200-5204 *in* 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 10.1109/ICASSP.2016.7472669.

Tuncer, T., Akbal, E., Dogan, S., 2021. Multileveled ternary pattern and iterative ReliefF based bird sound classification. Appl. Acoust. 176 https://doi.org/10.1016/j.apacoust.2020.107866.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention Is All You Need. Page arXiv:1706.03762. 10.48550/arXiv.1706.03762.

Wang, G., Zhao, Y., Tang, C., Luo, C., Zeng, W., 2022. When Shift Operation Meets Vision Transformer: An Extremely Simple Alternative to Attention Mechanism. Page arXiv: 2201.10801. 10.48550/arXiv, 2201.10801.

Wei, P.C., He, F.C., Li, L., Li, J., 2020. Research on sound classification based on SVM. Neural Comput. & Applic. 32, 1593–1607. https://doi.org/10.1007/s00521-019-04182-0.

Xeno-canto, https://www.xeno-canto.org. World wild bird sounds network (accessed 5 Dec 2022).

Xie, J., Zhu, M.Y., 2019. Handcrafted features and late fusion with deep learning for bird sound classification. Eco. Inform. 52, 74–81. https://doi.org/10.1016/j.ecoinf.2019.05.007.

Xie, J., Zhu, M.Y., 2022. Sliding-window based scale-frequency map for bird sound classification using 2D-and 3D-CNN. Expert Syst. Appl. 207 https://doi.org/10.1016/j.eswa.2022.118054.

Xu, H.F., Zhang, Y., Liu, J., Lv, D.J., 2021. Feature Selection Using Maximum Feature Tree Embedded with Mutual Information and Coefficient of Variation for Bird Sound Classification. Math. Probl. Eng. 2021 https://doi.org/10.1155/2021/8872248.

Yan, N., Chen, A.B., Zhou, G.X., Zhang, Z.Q., Liu, X.Y., Wang, J.W., Liu, Z.H., Chen, W.J., 2021. Birdsong classification based on multi-feature fusion. Multimed. Tools Appl. 80, 36529–36547. https://doi.org/10.1007/s11042-021-11396-9.

Yang, F., Jiang, Y., Xu, Y., 2022. Design of Bird Sound Recognition Model Based on Lightweight. IEEE Access 10, 85189–85198. https://doi.org/10.1109/access.2022.3198104.

Yu, X.Y., Zhu, W.B., Wei, J.X., Jia, S.F., Wang, A.D., Huang, Y.B., Zhao, Y.J., 2021. Estimation of ecological water supplement for typical bird protection in the Yellow River Delta wetland. Ecol. Ind. 127 https://doi.org/10.1016/j.ecolind.2021.107783.

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z., Tay, F.E., Feng, J., Yan, S., 2021. Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet. Page arXiv:2101.11986. 10.48550/arXiv, 2101.11986.

Zawadzki, J., C. J. Cieszewski, M. Zasada, and R. C. Lowe. 2005. Applying geostatistics for investigations of forest ecosystems using remote sensing imagery. Silva Fennica 39:599. 10.14214/sf.369.

Zhang, X., Chen, A.B., Zhou, G.X., Zhang, Z.Q., Huang, X.B., Qiang, X.H., 2019. Spectrogram-frame linear network and continuous frame sequence for bird sound classification. Eco. Inform. 54 https://doi.org/10.1016/j.ecoinf.2019.101009.