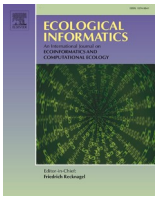
Ecological Informatics 81 (2024) 102637

Contents lists available at ScienceDirect

Ecological Informatics

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



SIAlex: Species identification and monitoring based on bird sound features

Lin Duan [a,b], Lidong Yang [a,b,*], Yong Guo [c]

[a] School of Digital and Intelligent Industry, Inner Mongolia University of Science and Technology, Baotou 014010, China
[b] Inner Mongolia Key Laboratory of Pattern Recognition and Intelligent Image Processing, Baotou 014010, China
[c] School of science, Inner Mongolia University of Science and Technology, Baotou 014010, China

ARTICLE INFO

Keywords:
Lightweight
Cascading activation function
Bird sound recognition
Structural re-parameterization
Nonlinear performance

ABSTRACT

The combination of deep learning and bird sound recognition is widely employed in bird species conservation monitoring. A complex network structure is not conducive for deploying bird sound recognition devices, resulting in problems such as long inference time and low efficiency. Using AlexNet as the backbone model, we explore the potential of shallow and straightforward models without complex connection techniques or attention mechanisms, named SIAlex, to recognise and classify 20 bird sound datasets, which are simultaneously validated on a 10 class UrbanSound8k dataset. Using the structural re-parameterization method, the number of model layers is reduced, computational efficiency is improved, and the inference time is significantly reduced, achieving a decoupling of training and inference time in the structure. To increase the nonlinearity of the model, a cascaded approach is utilised to increase the number of activation functions, thereby significantly improving the generalisation performance of the model. Simultaneously, in the classifier section, convolutional layer replaces the original fully connected layer, thereby reducing the inference time and increasing the feature extraction ability of the model, improving accuracy, and effectively recognising bird speech. The experimental data show that the SIAlex network on the Birdsdata dataset improves the accuracy to 93.66%, and the inference time for a piece of data is only 2.466 ms. The accuracy of the UrbanSound8k dataset reaches 96.04%, and the inference time for a piece of data is 3.031 ms. A large number of experimental comparisons have shown that the method proposed in this paper achieves good results in reducing the inference time of the model, bringing breakthroughs in the application of shallow, simple models.

1. Introduction

Combining bird voice recognition and deep learning is vital for monitoring bird species for conservation. It is an essential indicator for monitoring the ecosystem health and conducting practical ecological assessments for biodiversity conservation (Liu et al., 2023). Through population monitoring, scholars can understand the responses of local birds to environmental changes and conservation efforts. Monitoring bird movement in real-time is also the first step in monitoring the balance of the ecosystem (Xie et al., 2022). Many professionals have begun to conduct long-term observations of birds to protect their species. Therefore, training the collected data using appropriate network models and constructing an automatic bird recognition system are of profound significance.

Most bird datasets are standard natural source data that are manually labelled and tested by relevant professionals (Van Horn et al., 2015). However, owing to the rapid movement of birds and their sensitivity to the outside world, imaging devices may produce blurry effects when capturing bird images, making it challenging to identify bird species accurately. Considering that birds are difficult to capture from the perspective of obstacles when flying at high altitudes, it becomes challenging to collect data for detection through imaging devices (Yoshihashi et al., 2015). Moreover, pixel clarity and placement position of the devices have high requirements, resulting in high costs. For birds living in inaccessible high-altitude habitats, an increasing number of professionals use hearing (Fischer et al., 2023) and recordings to identify birds because of the difficulty of physical monitoring. This method, known as bioacoustic monitoring, can provide a passive and economical strategy for studying endangered bird populations. Compared with image acquisition, bird sound signals collected by audio recording devices have various characteristics, such as pitch, loudness, energy rate, entropy, and wavelength, which help better extract features from multiple angles and aspects. Fischer et al. (Fischer et al., 2023) demonstrated that more bird species are identified through recordings than

* Corresponding author at: School of Digital and Intelligent Industry, Inner Mongolia University of Science and Technology, Baotou 014010, China.
E-mail address: yld_nkd@imust.edu.cn (L. Yang).

https://doi.org/10.1016/j.ecoinf.2024.102637
Received 19 December 2023; Received in revised form 7 May 2024; Accepted 7 May 2024
Available online 13 May 2024

through standard observation-type data. Additionally, point-counting surveys revealed a significant difference in the number of species identified by observers and by audio recognition, with audio recognition being the most accurate. Therefore, an increasing number of researchers are investigating bird sound recognition methods. Combining bird sound datasets and deploying appropriate models plays a crucial role in bird sound recognition.

In recent years, neural networks have significantly improved the performance of models by increasing the complexity of network structures. The AlexNet network model proposed by Alex Krizhevsky et al. (Krizhevsky et al., 2017) demonstrates that the model depth is crucial for its performance. Stacking the convolution layer and activating function layers in the model and removing any intermediate layers resulted in a loss of approximately 2% in the top-level performance of the network. In the VGG network, Karen Simonyan et al. (Simonyan and Zisserman, 2015) validated the effect of model depth on training performance by stacking convolutional kernels with a size of 3 and training them with smaller parameter quantities and multiple scales under the same field of view. Simultaneously, this also opens up a new idea: reducing the memory space occupied by the model while increasing its complexity. This helps solve the problems of model training time and computing resources. Shazzadul Islam et al. (Islam et al., 2019) used VGG-16 to extract features from bird images and then used support vector machines for classification, achieving an accuracy of 89%. The ResNet network (He et al., 2016) optimises the network through residual modules by introducing skip connections between the traditional convolutional layers. This enables the network to learn and retain low-level feature information better, thereby avoiding gradient vanishing and representation bottlenecks. Sankupellay et al. (Sankupellay and Konovalov, 2018) used ResNet-50 to automatically recognise the spectrograms of 46 bird songs with an accuracy of 60%–72%.

An increasing number of scholars have integrated deep learning into bird voice recognition. Ágnes Incze et al. (Incze et al., 2018) transformed bird sounds into spectrograms and applied convolutional neural networks (CNNs) to classify bird sounds on the Xeno-canto dataset, demonstrating the potential of deep learning in the field of sound recognition. Xie et al. (Xie and Zhu, 2019) further confirmed the advantages of deep learning in bird sound classification; their method outperformed traditional methods in classifying 14 bird species. Gupta et al. (Gupta et al., 2021) explored the application of CNNs in large-scale bird classification and proposed a model that demonstrated how memory channel patterns change over time and the changes observed in spectrograms.

However, as time-varying images, bird song spectrograms are prone to ignoring important features of time-series data when using traditional CNNs for classification. To compensate for this deficiency, researchers have optimised the CNNs. Himawan et al. (Himawan et al., 2018) used a three-dimensional CNNs to simultaneously capture both long-term and short-term frequency information in bird sounds using a separate recurrent neural network to act on each filter in the last convolutional layer, resulting in an AUC score of 89.58%. Zhang et al. (Zhang et al., 2019) proposed an innovative method for bird song classification based on continuous frame sequences and spectrogram frame linear networks. Specifically, the method combines the sliding window algorithm, differential spectrogram, and gate control loop unit to capture the temporal dynamic characteristics of bird songs more accurately. Clark et al. (Clark et al., 2023) successfully detected 54 bird species by combining automatic recording devices and CNNs. This effectively solved the problems of noise interference and false positives, resulting in a total accuracy of 84.5% and average species accuracy of 85.1%. The phylogenetic perspective neural network proposed by Wang et al. (Wang et al., 2024) provides fine-layered and multi-layered labels for different birds. Using a hierarchical semantic embedding framework, feature information from different levels can be accurately captured and integrated, thereby achieving certain classification results on multiple bird datasets. Fu et al. (Fu et al., 2023) proposed an improved ACGAN model based on the residual structure and attention mechanism, called DR-ACGAN. Their model achieved stable training and generated high-quality bird singing spectrograms. It can generate dynamic convolution kernels with the MobileNetV2, ResNet18, and VGG16 models, resulting in new improvements.

Deep learning can be understood as building deep CNNs and using extensive sample data as the input. This results in a model with strong analytical and recognition capabilities. Researchers have long committed to improving the accuracy of bird sound recognition through complex and sophisticated methods to capture and distinguish bird sound characteristics more accurately. However, we found that well-designed shallow networks (Kahl et al., 2017), (Schlüter, 2018) can compete with extremely deep models in terms of performance (Sevilla and Glotin, 2017), (Lasseck, 2018). This study provides us with new ideas and insights.

In practical applications, as birds inhabit natural environments, the use of Internet of Things devices to remotely monitor bird populations has gradually become a trend. This technology not only provides a deeper understanding of the living habits of birds but also provides valuable data support for bird conservation and research. Therefore, bird sound recognition systems need to have high accuracy and should also consider the deployment cost and computational efficiency challenges caused by computational complexity. Although large and complex models may theoretically have higher recognition accuracy, they are often constrained by hardware resources, computational efficiency, and other factors in actual deployment and computation processes, making it difficult for them to perform as effectively as expected in practical scenarios.

In view of this discovery, we re-examined the application strategies of deep learning, no longer pursuing model complexity but exploring the potential of shallow networks. Through meticulous algorithm optimisation and structural design, we strive to maintain the model performance while reducing the computational complexity and deployment costs, achieving a balance between speed and accuracy. The specific contributions of this study are as follows:

- We propose a lightweight SIAlex model that utilises AlexNet as the backbone, fully exploiting the performance of minimalist models. Ensuring a good balance between speed and accuracy.
- The method of cascading multiple activation functions fully introduces nonlinear factors such that the model approximates the nonlinear expression function of the learning features while also improving the gradient propagation.
- We use structural re-parameterization techniques to decouple the training structure from the inference structure. This enables independent optimisation of the training and reasoning stages. Moreover, it enables the training phase to capture more feature information, while the inference phase simplifies the model structure, thereby reducing computational costs and deployment difficulties.
- A dual improvement in model performance and efficiency is achieved by replacing the fully connected layer in traditional classifiers with convolutional layers and simplifying the model under structural re-parameterization.

The remainder of the paper is structured as follows: Section 2 introduces the relevant work of previous researchers and the processing steps of the dataset. Section 3 introduces the optimisation methods of the model. Section 4 introduces the experimental results and compares the test results of different networks. The discussion and conclusion are presented in Sections 5 and 6, respectively.

2. Related work

2.1. Experience of predecessors

Deep learning models have become a mainstream research direction

in bird recognition and have achieved significant results. Li et al. (Xiangxia et al., 2021) reviewed the research progress of fine-grained image classification algorithms based on deep learning for bird image classification from the perspective of strong and weak supervision. They introduced various typical algorithms with excellent classification performances. These algorithms are widely used in fine-grained image recognition, including cutting-edge deep learning models such as YOLO (Redmon et al., 2016), multi-scale CNNs (Nah et al., 2017), and generative adversarial networks (Goodfellow et al., 2014). In addition, they compared the classification performances of the latest data enhancement methods related to fine-grained images. They analysed the performance characteristics of different fine-grained recognition methods for complex scenes. Finally, they compared and summarised the classification performance of the algorithms and explored future development directions and challenges they face. However, the problems of high intra-class variance and low inter-class variance among birds in bird images and low model efficiency always exist. To address these problems, Wang et al. (Wang et al., 2023a) improved feature extraction and model compression using fine-grained methods based on attention mechanisms and decoupled knowledge distillation. This method achieved an accuracy of 87.6% and a computational complexity of 1.2 G.

With the deepening of research, bird speech has become an essential source of information, and the combination of acoustic features and machine learning technology has become widely used in bird sound classification. Kogan et al. (Kogan and Margoliash, 1998) demonstrated the role of machine learning models in bird sound recognition by recording indoor bird call audio using hidden Markov models as the basic model and incorporating dynamic time-warped technology. However, they found that a weakness in the performance of hidden Markov models was the misclassification of song units with short vocalisations or more variable structures such as the syllables of specific calls and plastic songs.

Stastny et al. (Stastny et al., 2018) explored the classification of bird sounds. Based on their work, they extracted the cepstral coefficients of the human factors and detected bird occurrence fragments. They used hidden Markov models to identify 18 bird species in 6 families, with an inter-species success rate of 81.2% and a family classification success rate of 90.45%. Jiang et al. (Jiang et al., 2021) improved dynamic time warming by classifying ten bird species with an accuracy of 92.17% and a classification time of 1.88 s. Based on a support vector machine, Han and Peng (Han and Peng, 2023) added an error correction output encoding fusion and proposed the ECOC-SVM model to classify 11 types of birds with a prediction time of 52 ms.

With further research on the cross domain of images and audio, Kumar et al. (Kumar et al., 2022) successfully applied deep transfer learning models to bird sound classification tasks. They used well-known deep learning models, including 50 layers residual neural network and densely connected networks, to extract the Mel frequency cepstral coefficient of the audio as the feature input. The method performed well for 22 categories of bird sound classification and achieved high accuracy. Wang et al. (Wang et al., 2023b) believes that extracting the Mel frequency cepstral coefficient features during bird singing has a good effect, and concurrently, using deep learning models to combine static and dynamic modelling improves accuracy. In the LDFSRE-NET (Hu et al., 2023b) and MFF-ScSEnet (Hu et al., 2023a) models designed by Hu et al., recognition was achieved by combining bird voice audio feature extraction with deep learning. This not only demonstrates the potential of deep learning in bird voice recognition, but also highlights the importance of feature extraction and collaborative improvement of network models.

In studying deep learning for bird sound recognition, many studies have achieved high accuracy by improving the feature extraction methods or model structures. However, the complexity of a model can affect its computational complexity and inference time. The method introduced in this article, while acknowledging that deepening the number of model layers is beneficial for performance improvement, utilises structural re-parameterization technology and the principles of convolutional layer and convolutional layer fusion to ensure the nonlinear performance and accuracy of the model, while reducing the number of model layers and computational complexity, significantly shortening the time. The activation function was improved by cascading and stacking activation functions to enhance the nonlinear computing ability of the model. Replacing the fully connected layer of the classifier with a convolutional layer can reduce the inference time and increase the robustness of the model. The improvement of the SIAlex model not only improves the accuracy but also dramatically shortens the testing time, making breakthroughs in shallow and simple models and making them easier to implement in practical applications.

2.2. Feature extraction

Before extracting bird sound features, corresponding preprocessing is required. The specific steps are shown in Fig. 1. The duration of each bird song segment in the dataset is different, so this article intercepts sample data at 2 s intervals to ensure that the duration of each sample data is the same. To eliminate the impact of amplitude differences in bird audio data on model training, the following min-max standardisation process was (Haga et al., 2019) performed on each intercepted bird sample data:

$$S(n)_t = \frac{S(n)_t - min\{s(n)\}}{max\{s(n)\} - min\{s(n)\}} \tag{1}$$

where $S(n)_t$ represents the normalised input signal at time t; $s(n)$ represents the original input signal at time x; $min\{.\}$ and $max\{.\}$ represent the minimum and maximum values, respectively.

To better recognise the characteristics of bird sounds, the Mel spectrum (Imai et al., 1983), which is widely used in speech recognition systems, was chosen as a feature for bird audio signals. The feature-extraction steps are as follows (Gupta et al., 2013):

1) Preprocessing: This step processed the signal through a filter that emphasised high frequency. The purpose of the pre-profession is to compensate for the loss of high-frequency components to enhance energy at high frequency.
2) Framing: The process of dividing speech signals into N shorter frames (10 ms–30 ms). A partial overlap M (M < N) exists between two adjacent frames to make the parameters between frames transition smoothly. Typical values used were M = 100 and $N = 256$.
3) Windowing: It is used to avoid discontinuity of signals generated from the framing process.
4) The Mel spectrum of the bird audio signal obtained in this paper is defined as follows:

$$feature(m) = \sum_{k=0}^{N-1} E(k) H_{\mathrm{m}}(k) \tag{2}$$

where $feature(m)$ is the energy characteristic that corresponds to the Mel filter; $E(k)$ is the signal energy spectrum; $H_{\mathrm{m}}(k)$ is the response of the Mel filter; and N is the FFT length. Fusion of features in the channel dimension to obtain a feature map.

3. The proposed method

3.1. AlexNet network architecture

The AlexNet network (Krizhevsky et al., 2017) was used as the backbone structure to compare the effectiveness of the methods proposed in this study. By adjusting the layout of the basic convolutional layer, normalisation layer, and activation function, the accuracy was improved while reducing the computational complexity of the model, thereby improving the computational efficiency. The details of the

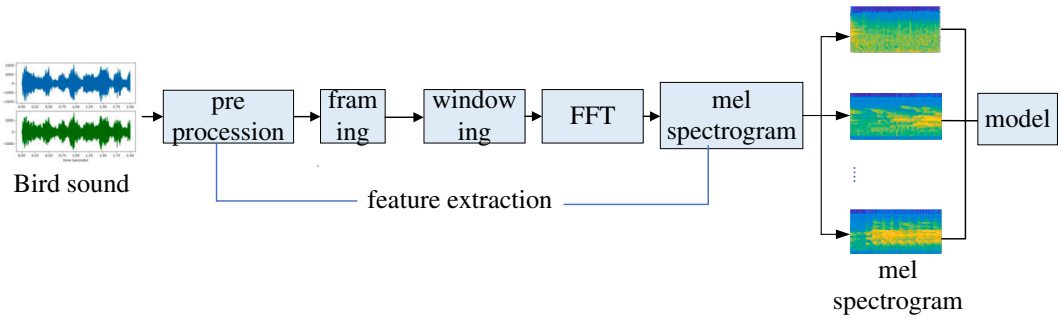


Fig. 1. Bird audio signal feature extraction process.

module improvements are shown in Fig. 2 and Fig. 3.

AlexNet is a deep CNN. Its structure can be divided into three main parts: conv-block, max-block and classifier. The conv-block comprises a series of convolution layers, normalisation layers, and activation function layers. It extracts features from the input data and captures local features through convolution layers. The normalisation layer is used to accelerate training and prevent overfitting, while the activation function layer uses the objective function to increase the nonlinear expression ability of the model. The max-block module is located behind the conv-block. Through the down-sampling operation, the size of the input feature map is decreased, effectively reducing the amount of data while retaining the essential features, thus reducing the amount of calculations of the model. Second, it can increase the robustness of the model and make it more adaptive to input characteristics. A classifier is a crucial part of a model for classification decisions. It comprises a complete connection layer and a dropout layer. The entire connection layer multiplies the output of the previous layer by a weight matrix. It performs nonlinear operations using an activation function by mapping the input to the final classification result. Simultaneously, dropout technology is used to prevent overfitting and improve the generalisation ability of the model.

3.2. Improvement method

3.2.1. Merging of convolution layers

In deep learning, the improvement in model performance usually depends on the addition of a hierarchical structure to the model. The nonlinear ability of the model can be enhanced to train a more generalised model by adding convolution modules or introducing residual connection techniques. However, more complex and deeper models often incur higher computational costs and require longer training time. To solve these problems, the AlexNet network was designed to be lightweight and optimised to achieve deep training and short-term testing (Zhang et al., 2022). Specifically, the two convolution layers are merged in the model deployment mode while ensuring that the nonlinear factors caused by the activation function are not reduced. By reducing the number of layers in the model, the computational efficiency is improved (Ding et al., 2021; Zhao et al., 2017) and the inference time of the model is significantly shortened.

First, a single convolution layer is merged with a single Batch Normalisation (BN) layer. When training the model, the BN layer can accelerate the network convergence and prevent overfitting. By normalising the BN layer, the problems of gradient disappearance and explosion can be effectively solved. The method has been applied to many advanced network models, such as ResNet, MobileNet, Xception, and ShuffleNet, which use BN technology to optimise model performance. However, although the BN layer plays an active role in the training process, additional operations are added to the network forward inference, which affects the performance of the model and occupies more memory or memory space. To solve this problem, the BN layer parameters can be merged into a convolution layer to improve the forward inference speed of the model. Through the above optimisation steps, we can further improve the efficiency and performance of deep learning models and promote the development of related fields.

In the convolutional layer, the weight relationship is as follows:

$$y_{conv} = wx + b \tag{3}$$

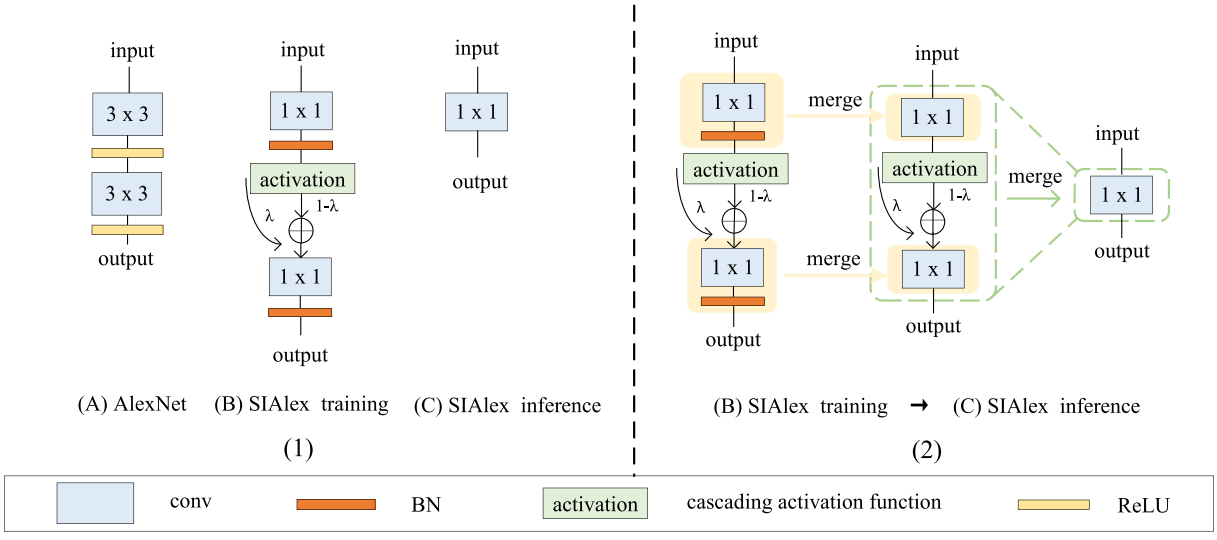


Fig. 2. SIAlex's optimisation process for convolutional layers and activation functions. Among them, part (1) compares the convolutional module structures of the original AlexNet and SIAlex in the training and inference stages. Part (2) shows the process of structural re-parameterization.

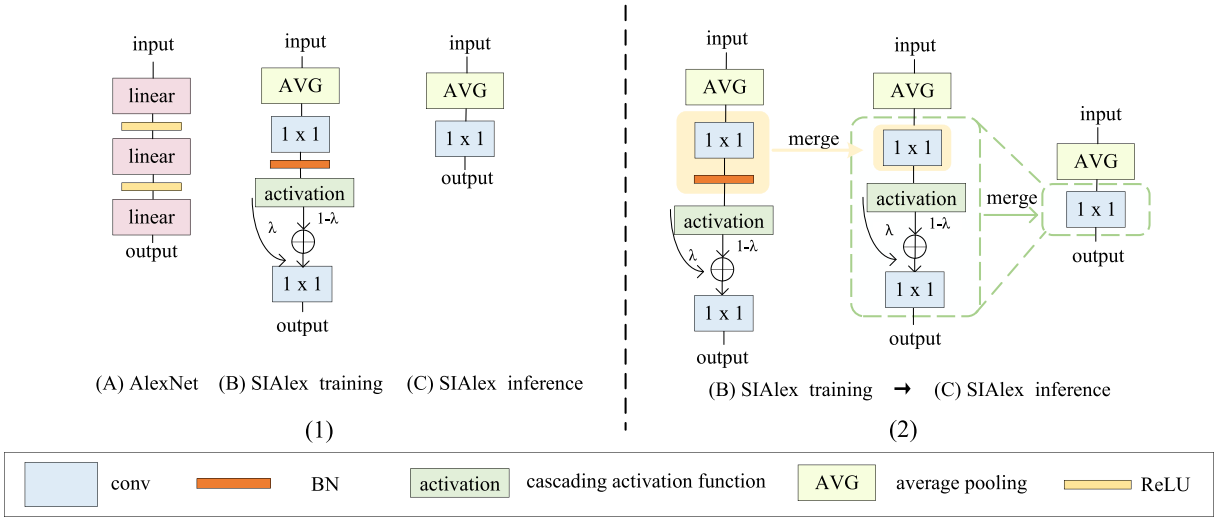


Fig. 3. SIAlex's optimisation process for the classifier part. Among them, part (1) compared the classifier structures of the original AlexNet and SIAlex in the training and inference stages. Part (2) shows the process of structural re-parameterization.

where w represents the weight, b represents the offset, x represents the input features, and y_{conv} represents the output features after passing through the convolutional layer.

In the BN layer, the definitions of the input and output features are as follows:

$$y_{bn} = \gamma\left(\frac{x_{bn} - \mu}{\sqrt{\sigma^2 + \varepsilon}}\right) + \beta \tag{4}$$

where μ represents the mean that can be expressed as $\frac{1}{m}\sum_{i=1}^{m}(x_{bn})_i$, and σ represents the variance that can be expressed as $\frac{1}{m}\sum_{i=1}^{m}\left((x_{bn})_i - \mu\right)^2$. γ is the scaling factor, β is the offset, and ε is a smaller number to prevent the denominator from becoming zero.

When the input features x passes through the convolutional layer, the input feature of the BN layer is the output feature of the convolutional layer, which is called an $x_{bn} = y_{conv}$. Therefore, the output feature of x after processing by the convolutional and BN layers is:

$$y_{bn} = \gamma\left(\frac{x_{bn} - \mu}{\sqrt{\sigma^2 + \varepsilon}}\right) + \beta \tag{5}$$

$$= \gamma\left(\frac{y_{conv} - \mu}{\sqrt{\sigma^2 + \varepsilon}}\right) + \beta \tag{6}$$

$$= \gamma\left(\frac{wx + b - \mu}{\sqrt{\sigma^2 + \varepsilon}}\right) + \beta \tag{7}$$

$$= \gamma\left(\frac{w}{\sqrt{\sigma^2 + \varepsilon}}\right)x + \gamma\left(\frac{b - \mu}{\sqrt{\sigma^2 + \varepsilon}}\right) + \beta \tag{8}$$

We assign the weights of the merged convolutional layer and the bn layer to a new convolutional layer, and the merged convolutional layer can be represented as:

$$y_{conv+bn} = w_{conv+bn}x + b_{conv+bn} \tag{9}$$

among them, $w_{conv+bn} = \gamma\left(\frac{w}{\sqrt{\sigma^2+\varepsilon}}\right)$, and $b_{conv+bn} = \gamma\left(\frac{b-\mu}{\sqrt{\sigma^2+\varepsilon}}\right) + \beta$.

The schematic diagram of parameter merging is shown in Fig. 4.

In the second step, the activation function is integrated into the convolutional layer process. The activation function plays a crucial role in the convolutional layer of deep network models and plays a crucial role in nonlinear training. To merge two convolutional layers, it is necessary to first handle the nonlinear relationship between these two convolutional layers. The activation function was improved to gradually

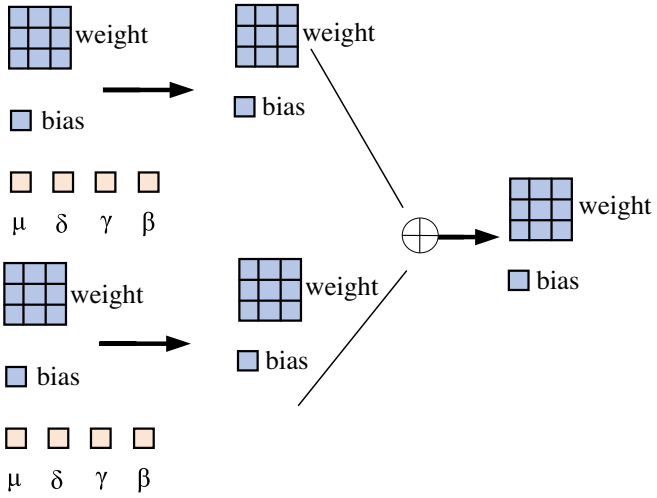


Fig. 4. Principle of convolution layer merging.

adapt the model to merging during the training process. The improved activation function can be gradually transformed into an identity map, laying the foundation for merging two convolutional layers while maintaining its training effect on the model. Thus, the two convolutional layers could be merged smoothly without losing the training effect of the activation function on the model.

Create the original activation function $A(x)$. The improved activation function is $A'(x)$. The expression for the improved activation function is as follows:

$$A'(x) = (1 - \lambda)A(x) + \lambda(x) \tag{10}$$

where, λ is a hyperparameter used to adjust the balance relationship between the nonlinear factors of the activation function and the identity mapping of $A'(x)$ based on the number of training iterations, which can be expressed as $\lambda = \frac{e}{E}$. Here, e represents the current number of rounds of training and E represents the total number of rounds of model training.

Starting training, $\lambda = \frac{e}{E} \to 0$, and $A'(x) \to A(x)$. At this point, a strong nonlinear factor exists between the two convolutional layers of the network, which can achieve better training of the model. When $e \to E$, $\lambda \to 1$, and $A'(x) \to x$. Nonlinear factors are gradually integrated into the

convolutional layer, and no activation function exists between the two convolutional layers, which can be merged.

The third step is to merge convolutional layers. Let W_1 and W_2 serve as weights for the two convolutional layers, and let $im2col()$ represent the transformation of the input into a matrix that corresponds to the shape of the kernel. Upon merging the convolutional and BN layers, the two convolutional layers without activation functions can be merged using the following formula. The output function after merging the two convolutional layers is as follows:

$$Y_{conv+conv} = W_1^*(W_2^* x) = W_1 \cdot W_2 \cdot im2col(x) = (W_1 \cdot W_2)^* x \tag{11}$$

3.2.2. Cascading activation function

In deep learning, model optimisation often focuses on enhancing the performance of complex deep networks by flexibly selecting different activation functions. This study reduces the complexity of the model through the fusion of convolutional layers, thereby reducing the computational complexity and improving the computational efficiency. However, it also brings about a crucial problem: a simple network structure leads to weak nonlinearity, which limits performance. To solve the vanishing gradient problem in the ReLU activation function, this study replaced the ReLU activation function with a Leaky ReLU (Dubey and Jain, 2019; Xu et al., 2020) activation function. The Leaky ReLU function provides a non-zero slope to the negative values in the input data, as defined below:

$$Leaky\_ReLU(x) = \begin{Bmatrix} x, x \geq 0 \\ ax, x < 0 \end{Bmatrix} \tag{12}$$

where, x represents the input data and a represents the hyperparameter of the activation function.

To further enhance the nonlinear expression ability of the model, an innovative method of cascading activation functions (Chen et al., 2023; Mhaskar and Poggio, 2016) is adopted. The method enables the model to exhibit richer and deeper nonlinear characteristics when processing complex data using multiple cascading activation functions (Eldan and Shamir, 2016). Using this optimisation strategy, we can effectively improve the performance of the model and provide powerful tools for solving various complex practical problems.

Specifically, we denote the original activation function in the model as $A(x)$. The stacking of multiple activation functions is represented by $A_n(x)$ and a powerful activation function module is constructed by cascading multiple activation functions. A mathematical cascading formula is used to represent the following equation:

$$A_n(x) = \sum_{i=1}^{n} a_i A(x + b_i) \tag{13}$$

where n represents the number of activation functions, and a_i and b_i represent the scale and bias of each activation function, respectively.

By combining the bird sound feature recognition applications, the activation function module can be expressed using the following formula:

$$A_n(x_{h,w,c}) = \sum_{i,j \in \{-n,n\}} a_{i,j,c} A(x_{i+h,j+w,c} + b_c) \tag{14}$$

where h represents the height of the feature dimension; w represents the width of the feature dimension; and c represents the number of channels in the feature dimension.

In the subsequent experimental section, we demonstrate how to determine the number of cascading activation functions required to achieve optimal model performance. Using this method, breakthroughs have been achieved in bird sound feature recognition.

3.2.3. Improvement of classifier

Fig. 5 shows the AlexNet classifier module, and Fig. 6 shows the improved SIAlex classifier module.

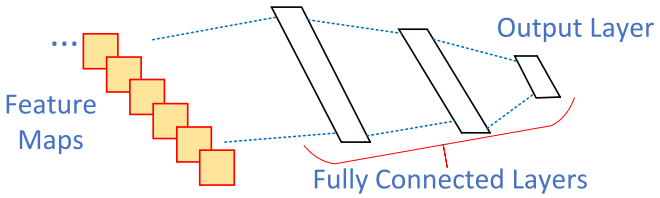


Fig. 5. Basic classifier structure.

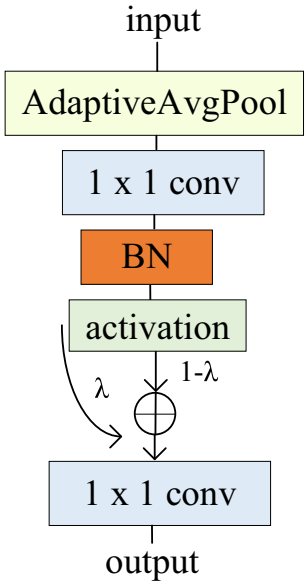


Fig. 6. Classifier module for SIAlex.

To demonstrate the effectiveness of the proposed method on the AlexNet model, the classifier module was modified. The traditional fully connected layer is abandoned and adopts a 1 × 1 convolutional layer for the channel expansion (Springenberg et al., 2015). This innovative modification successfully simulated the functionality of the fully connected layer while solving the problems of spatial structure destruction and significant computational parameters in the fully connected layer. By applying convolutional layers to the classifiers, a nonlinear transformation was achieved. In addition, applying the convolutional layer fusion principle proposed in this study to the improved classifier enables the model to improve both the nonlinear performance and testing efficiency.

3.2.4. The network structure of SIAlex

The model designed in this study is based on AlexNet as the backbone, and has been improved in the convolutional layer, activation function, and classifier module. The overall architecture of the SIAlex model is presented in Table 1. The model comprises conv-blocks, classifier, several max-blocks. Meanwhile, in Fig. 7, we demonstrate the basic flowchart of the model using SIAlex-3 as an example. Schematic diagrams of the conv-block and max-block under structural re-parameterization are shown in Fig. 2 and Fig. 3.

4. Experiments

4.1. Dataset and experimental configuration

To fully validate the effectiveness and generalisation of the model, experiments were conducted on the Birdsdata and Urbansound8k datasets.

Table 1
Overall architecture of the SIAlex model, where n represents the number of max-blocks stacked in SIAlex-n. S represents step size, k represents convolutional kernel size, [input, output], represents the number of input and output channels.

	SIAlex-3	SIAlex-4	SIAlex-5
conv-block	conv,k = 4,s = 4, [3512] conv,k = 1,s = 1, [512,512]	conv,k = 4,s = 4, [3512] conv,k = 1,s = 1, [512,512]	conv,k = 4,s = 4, [3512] conv,k = 1,s = 1, [512,512] conv,k = 1,s = 1, [512,512]
max-block	conv,k = 1,s = 1, [512,512] conv,k = 1,s = 1, [512,512] maxpool,k = 2,s = 2 conv,k = 1,s = 1, [1024,1024] conv,k = 1,s = 1, [1024,2048] maxpool,k = 2,s = 2 conv,k = 1,s = 1, [2048,2048] conv,k = 1,s = 1, [2048,4096] maxpool,k = 2,s = 2	conv,k = 1,s = 1, [512,512] conv,k = 1,s = 1, [512,1024] maxpool,k = 2,s = 2 conv,k = 1,s = 1, [1024,1024] conv,k = 1,s = 1, [1024,2048] maxpool,k = 2,s = 2 conv,k = 1,s = 1, [2048,2048] conv,k = 1,s = 1, [2048,4096] maxpool,k = 2,s = 2 conv,k = 1,s = 1, [4096,4096] conv,k = 1,s = 1, [4096,4096 maxpool,k = 2,s = 2]	conv,k = 1,s = 1, [512,512] maxpool,k = 2,s = 2 conv,k = 1,s = 1, [512,512] conv,k = 1,s = 1, [512,1024] maxpool,k = 2,s = 2 conv,k = 1,s = 1, [1024,1024] conv,k = 1,s = 1, [1024,2048] maxpool,k = 2,s = 2 conv,k = 1,s = 1, [2048,2048] conv,k = 1,s = 1, [2048,4096] maxpool,k = 2,s = 2 conv,k = 1,s = 1, [4096,4096] conv,k = 1,s = 1, [4096,4096] maxpool,k = 2,s = 2]
classifier	adaptiveavgpool conv,k = 1,s = 1, [4096,20] conv,k = 1,s = 1, [20,20]	adaptiveavgpool conv,k = 1,s = 1, [4096,20] conv,k = 1,s = 1, [20,20]	adaptiveavgpool conv,k = 1,s = 1, [4096,20] conv,k = 1,s = 1, [20,20]

1) Birdsdata dataset: This dataset was jointly launched by Bainiao Data and Beijing Zhiyuan Artificial Intelligence Research Institute. After testing the cutting quality, a 30-min standard sound was selected and organised. The text adopts the common part of the Birdsdata dataset, which contains 14,311 natural audio clips, and all of its sound data are designed and collected in natural scenes. The Birdsdata natural sound detection dataset primarily targets the detection problem of collecting sound categories in nature and provides a diverse and practical benchmark for object detection research. Each audio length was standardised and cropped to 2 s, and audio files less than 2 s were cleared to ensure data consistency and accuracy. Twenty types of bird sounds were used. Audio saved in the wav format. The specific categories of bird sound datasets are listed in Table 2.

2) UrbanSound8k dataset: This is a public dataset prepared by Salamon et al. (Salamon et al., 2014) for automatic urban environmental sound classification research. Audio is saved in the wav format. These classes included air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren, and street music. It contains 8732 audio clips from 10 categories, each lasting 4 s. The sampling rate of the audio clips is 44.1 kHz and the clips were recorded in stereo. The UrbanSound8k dataset provides a CSV file that contains metadata for each audio clip, such as category labels and recording devices. These metadata help researchers better understand and analyse the sound samples in the dataset.

We used Mel spectrograms as input features for these datasets. For dataset partitioning, the training and testing sets were divided at an 8:2 ratio.

The evaluation indicators of the experiment mainly included the testing accuracy, inference time, precision, recall, specificity, F1 score, and sensitivity. In the experiment, the batch size was set to 16, and the learning rate was 0.00005, which was optimised using the Adam optimiser. For the AlexNet and SIAlex models, the number of epochs was set as 200. A cross-entropy loss function was adopted the cross entropy loss function.

Table 2
The Birdsdata dataset collected 20 types of bird sounds both indoors and outdoors in the form of actual collection and network collection.

Types of birds	Number of samples	Types of birds	Number of samples
grey goose	759	goshawk	733
whooper swan	800	eurasian Buzzard	290
mallard	766	western Yangji	680
greenwing	602	coot	460
grey partridge	29	black-winged stilt	786
western quail	738	fengtou Wheat Chicken	814
chicken	797	green Sandpiper	710
red throated diver	835	redshank	790
heron	850	pratincole	825
common Cormorant	852	sparrow	1195

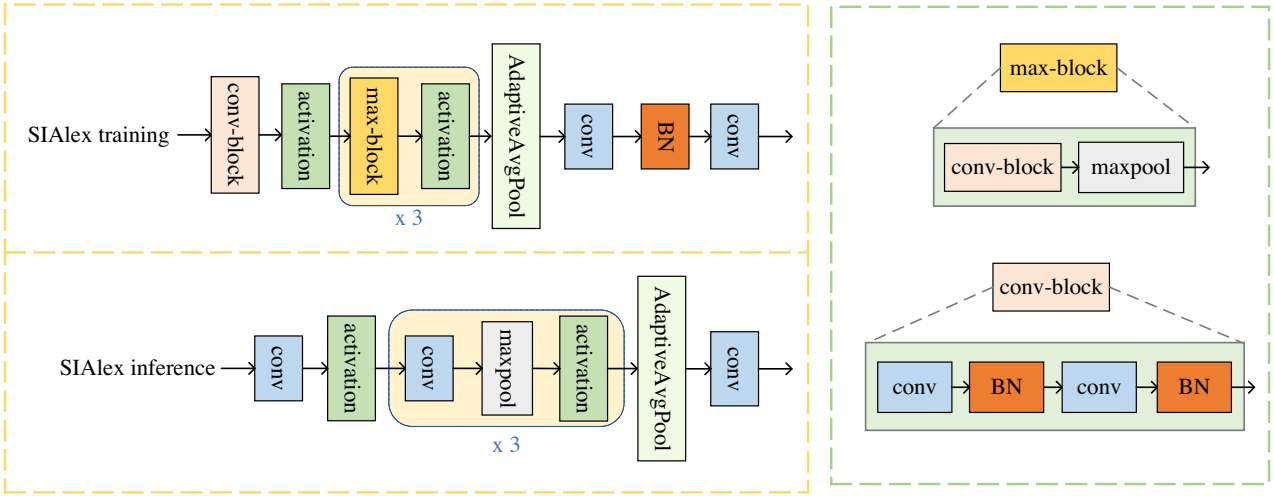


Fig. 7. Structural flowchart of SIAlex.

4.2. Ablation experiment

4.2.1. Selection of activation function and classifier

Table 3 presents comparative data on the optimisation of the two models, activation function selection, and classifier design. It can be observed that under the primary classifier, by comparing ReLU in AlexNet with Leaky ReLU, replacing the activation function increases the accuracy by 0.92%, indicating that the Leaky ReLU function better solves the problem of gradient vanishing in the ReLU function.

Comparing the two scenarios using fully connected layers and convolutional layers instead of fully connected layers in the classifier, the accuracy increased by approximately 1.96%. Considering that the convolutional fusion method reduces the number of model layers and improves the computational efficiency, the activation function of Leaky ReLU and the classifier with convolutional layers are used as the classification part of the model.

4.2.2. Cascading activation function

It can be observed that when using a single activation function, that is, $n = 0$, the testing accuracy can only reach 83.34%. Concurrently, $n = 1$ indicates that when two activation functions start cascading, the accuracy increases to 89.77%, 6.43% higher than ordinary activation functions. Improvement in nonlinear factors is helpful for model training. As n increases, the accuracy of the model gradually improves, whereas the testing time increases slightly. Here, we introduce the data on the testing time before and after convolutional layer merging. The experiment shows that although the accuracy at $n = 4$ is higher than that at $n = 3$, the testing time after merging the convolutional layers does not significantly decrease. Considering the evaluation indicators of the testing accuracy and inference time, $n = 3$ is used as the number of cascading activation functions in the experiment. The experimental data are listed in Table 4.

4.2.3. Merge of convolutional layers

In Table 5, a comparison is made between the SIAlex model and the convolutional layer merging method. After merging the convolutional layers, the testing time was significantly reduced by approximately 60%. They effectively improve recognition efficiency. Experimental data show that when convolutional layers and BN layers are applied to the model and the convolutional layers and convolutional layers are fused, the testing time is reduced while ensuring accuracy. When bird sound recognition is deployed on mobile or detection devices, the model reduces costs, while ensuring the efficiency and accuracy of bird sound recognition.

4.2.4. Comparison with existing model methods

Classic deep learning models, such as ResNet and DenseNet (Zhu and Newsam, 2017), and lightweight deep learning models, such as MobileNet (Sandler et al., 2018), ShuffleNet (Zhang et al., 2018), and EfficientNet (Tan and Le, 2019), were selected to demonstrate the effectiveness of the proposed model. We train the bird sound recognition dataset using the above model, record the testing accuracy, and testing time of different models, and compare them with the model proposed in this study. The experimental data are shown in Fig. 8.

Compared to other lightweight models, SIAlex improved the accuracy to 93.66% with a testing time of only 2.466 ms. SIAlex's multi-indicator evaluation for each category is shown in Fig. 10, and Fig. 9 presents the confusion matrix more intuitively.

Table 3
Comparison of AlexNet under different activation function and classifier.

	ReLU	Leaky ReLU	Basic classifier	classifier	acc(%)
AlexNet	✓		✓		80.46
AlexNet		✓	✓		81.38
AlexNet		✓		✓	83.34

Table 4
Comparison of the number of different cascading activation functions. The testing time before and after model merging are compared as data evidence.

n	cascading activation	conv merge	latency (ms)	acc (%)
n = 0	✘	✘	3.357	83.34
	✓	✓	1.959	
n = 1	✓	✘	3.399	89.77
	✓	✓	2.089	
n = 2	✓	✘	3.426	91.35
	✓	✓	2.129	
n = 3	✓	✘	3.447	93.18
	✓	✓	2.187	
n = 4	✓	✘	3.603	93.23
	✓	✓	2.269	

Table 5
Effect of convolutional layer merging on test time.

model	conv merge	latency (ms)	acc(%)
SIAlex-3	✘	3.447	93.18
	✓	2.187	
SIAlex-4	✘	3.536	93.37
	✓	2.345	
SIAlex-5	✘	3.632	93.66
	✓	2.466	

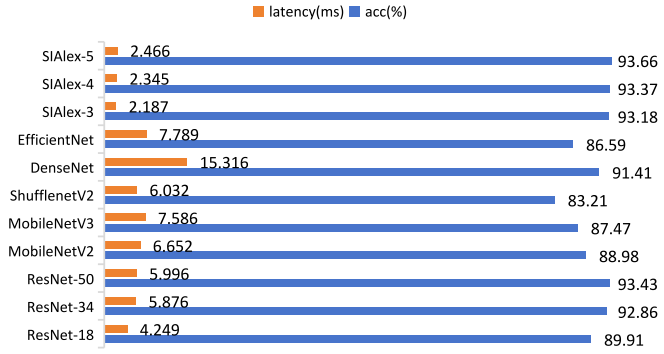


Fig. 8. Comparison of various models, where acc represents testing accuracy, and latency represents testing time for a piece of data.

Compared with other models, the model proposed in this study has advantages in terms of accuracy and testing time. Ensuring the accuracy of bird sound recognition dramatically reduces the testing time and improves the recognition efficiency.

The experimental comparison in ResNet shows that as the number of layers in the network model increases, the number of non-linear layers of the network also increase, which improves the model performance and leads to a specific improvement in accuracy. However, the testing time of the model also increases. By comparing the SIAlex models with different module layers, it was verified that the accuracy improves with an increase in the number of module layers under the same conditions. However, increasing the number of layers results in higher accuracy, but the testing time also increases. Fig. 11 and Fig. 12 show the performances of the proposed model's training and testing sets, respectively, which converged quickly. The model converges in 200 epochs.

For the DenseNet network, the accuracy is 91.41%, which is the closest to the accuracy of SIAlex. However, its testing time is as long as 15.316 ms. The analysis suggests that the reason for this result may be the use of dense blocks and transition layer structures in the DenseNet network, where dense blocks are composed of multiple convolutional blocks; each block uses the same number of output channels, and the

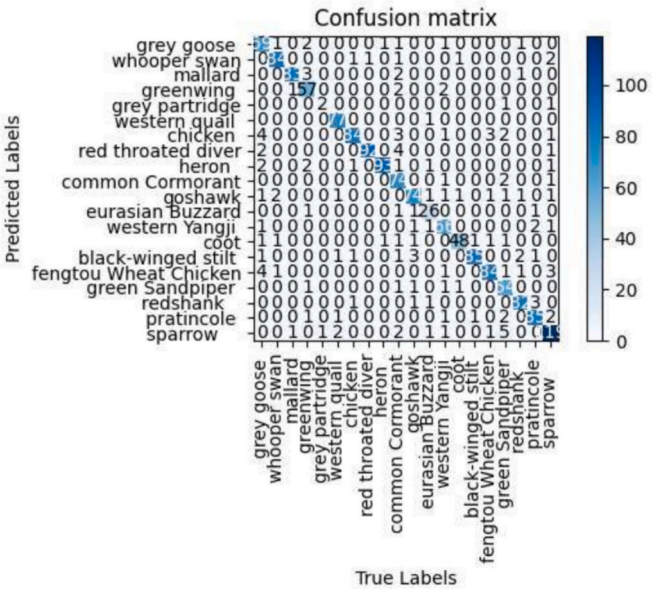


Fig. 9. The confusion matrix of SIAlex.

input and output of each block are connected in the channel dimension through a loop. However, excessive use can lead to overly complex models owing to an increase in the number of channels caused by each dense block connection. Therefore, it is necessary to balance the model complexity and performance, which can also affect the testing time.

Through a comprehensive comparison, the SIAlex model is optimised using a lightweight model approach, which significantly improves the model's testing accuracy indicators and reduces the testing time, thereby improving the recognition efficiency of the model. This also taps into the potential of shallow and straightforward models, enabling them to perform better at a relatively low model complexity and contributing to the low-cost development of bird sound recognition for practical device deployment.

4.3. Generalisation experiments

In the generalisation experiment section, we conduct experiments on the UrbanSound8k dataset, an urban environmental classification dataset that better reflects the effectiveness and generalisation of the SIAlex model. The experimental results are shown in Fig. 13. Fig. 14 shows the evaluation of SIAlex on 10 types of data in UrbanSound8k.

The accuracy of the SIAlex model can reach 96.04%, and its inference time is only 3.031 ms. Compared with existing technologies, our method further improves the classification accuracy, which once again verifies the effectiveness of our method. In summary, the SIAlex model balanced accuracy and inference time on the UrbanSound8k dataset, providing a new solution for audio event classification tasks and demonstrating enormous potential for real-time applications. This achievement provides new references for research and practice in related fields, and lays the foundation for future model optimisation and performance improvement.

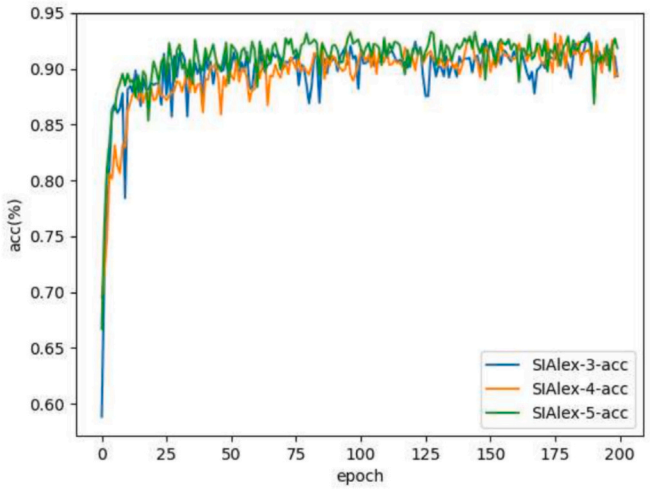


Fig. 11. Accuracy of SIAlex.

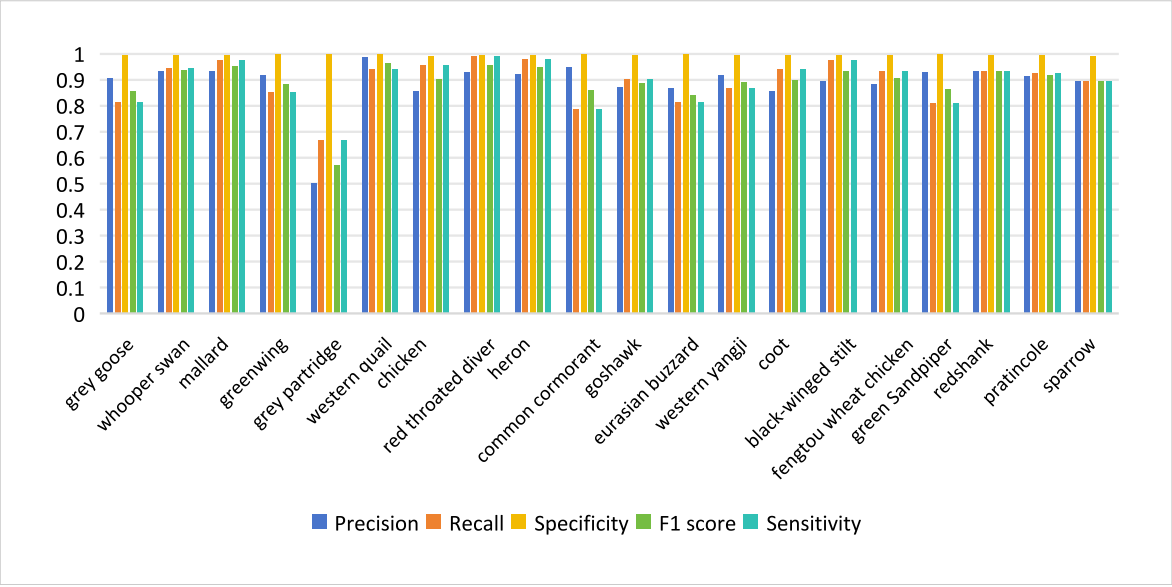


Fig. 10. Conduct a comprehensive evaluation of SIAlex on each category of the bird song dataset using five evaluation metrics: Precision, Recall, Specificity, F1 score, and Sensitivity.

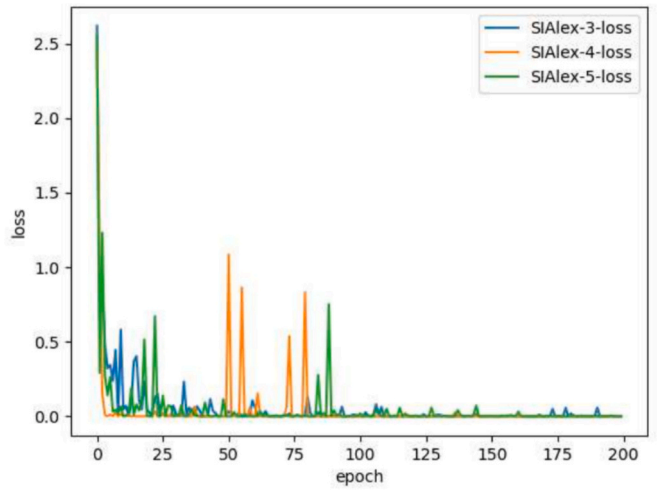


Fig. 12. Loss of SIAlex.

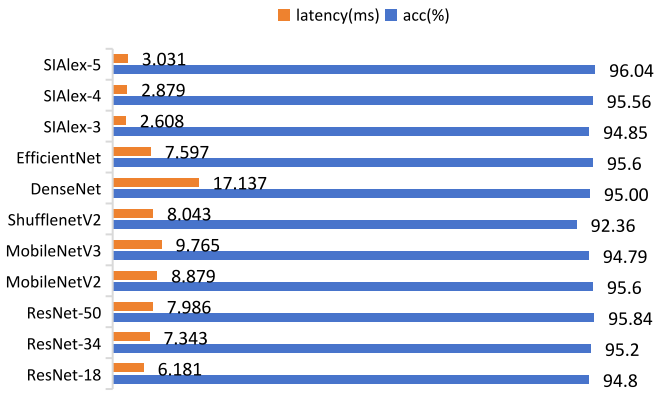


Fig. 13. Comparison of accuracy and testing time between SIAlex and other models on UrbanSound8k.

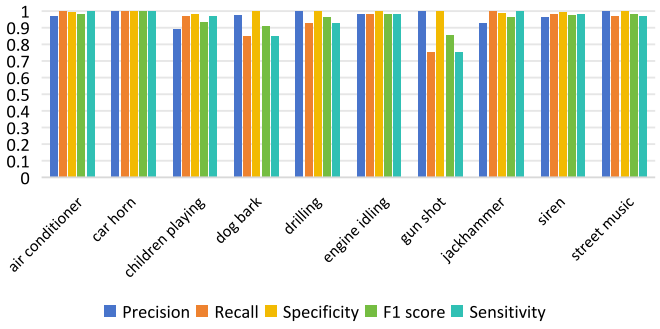


Fig. 14. Conduct a comprehensive evaluation of SIAlex on each category of UrbanSound8k dataset using five evaluation metrics: Precision, Recall, Specificity, F1 score, and Sensitivity.

5. Discussion

The SIAlex model proposed in this study achieved significant results in bird sound recognition. In addition to improving the recognition accuracy, the model also significantly reduces inference time, providing strong technical support for the deployment of bird sound recognition equipment.

To explore the performance potential of the minimalist models in depth, AlexNet was selected as the backbone model. We optimised the algorithm and structure of its basic convolutional layers, thereby achieving dual improvements in accuracy and speed. In this process, we improved the activation function and classifier of the original AlexNet and validated them through ablation comparison experiments. As indicated in Table 3, after replacing the ReLU activation function, the accuracy of the model improved by 0.92%. To further enhance the nonlinear expression ability of the model, we adopted a method of cascading activation functions method. By introducing multiple activation functions, the nonlinear characteristics of the model were enhanced, and the model could learn the feature information more deeply at different levels during the training phase. As indicated in Table 4, when the number of activation functions $n = 3$, the performance improvement of the model is the most significant.

We used structural re-parameterization techniques to decouple the model into the training and inference stages. This strategy enables the independent optimisation of accuracy and speed. Specifically, in the inference stage, we implemented merging operations between convolutional layers and bn layers, as well as between convolutional layers, effectively reducing the number of model layers and significantly reducing computational complexity. As shown in Table 5, this strategy significantly reduces the inference time of the model and improves the recognition efficiency. Additionally, we improved the classifier by replacing the fully connected layer in traditional classifiers with convolutional layers, which is beneficial for simplifying the model structure and achieving a dual improvement in accuracy and speed.

When performing bird sound recognition on the Birdsdata dataset, as shown in Fig. 10, we noticed that the comprehensive indicators, such as the grey partition, are relatively low. This was because of the limited amount of bird sound data available for this breed. To overcome this limitation and improve the recognition performance of grey partitions, we plan to expand the dataset in future research to include more diverse bird song samples. Through this strategy, we aim to cover various bird sounds more comprehensively, thereby improving the accuracy and reliability of bird sound recognition.

This study conducted experiments on the Birdsdata and UrbanSound datasets, fully demonstrating our significant advantages in terms of accuracy and inference time through comparison with other methods. The specific data are presented in Fig. 8 and Fig. 13. Our model significantly reduces the inference time while maintaining high accuracy. Compared with other networks, the experimental data show that the SIAlex network on the Birdsdata dataset improves the accuracy to 93.66%, and the inference time for a single data is only 2.466 ms. The accuracy of the UrbanSound8k dataset reaches 96.04%, and the inference time for a single data is 3.031 ms. Meanwhile, it can be observed that compared with existing lightweight models (Sandler et al., 2018), (Zhang et al., 2018), (Tan and Le, 2019), the proposed model demonstrates superior performance.

However, compared to other methods (Wang et al., 2023b), (Kahl et al., 2021), (Tang et al., 2023), there is still room for improvement in accuracy. This is primarily owing to the insufficient feature learning of the model. To further improve the performance of the model, we can draw on the research ideas of relevant literature and adopt various strategies to optimise it. On one hand, we can perform hierarchical label processing on the bird sound dataset (Wang et al., 2024), (Swaminathan et al., 2024) to describe the feature information of bird sounds more finely, thereby helping the model better learn and recognise different categories of bird sounds. Multi-scale recognition (Zhang et al., 2024a) can be performed on datasets or to fully learn bird song features through methods such as multi-granularity fusion (Zhang et al., 2024b). On the other hand, we can also consider expanding the dataset size and using richer bird sound data for training to enhance the generalisation ability and recognition accuracy of the model.

6. Conclusions

The SIAlex model proposed in this study, with its simplicity and efficiency at its core, is used for the automatic recognition and monitoring

of bird sounds. The model achieves a balance between speed and accuracy improvement by decoupling the model structure into training and inference stages through structural re-parameterization, achieving the convenience of low-cost computing and device deployment. To further improve the accuracy, we adopted a cascaded activation function to introduce nonlinear factors, thereby improving the learning ability of the model for bird sound features. The validation conducted on the Birdsdata and UrbanSound8k datasets demonstrated the generalisation ability of the model and significant advantages in terms of inference time and accuracy.

However, we also observed that the evaluation metrics may be affected when dealing with rare birds because of the uneven distribution of the dataset. Therefore, we plan to address this challenge by expanding the dataset. Considering that the current model is still shallow, its processing ability for large-scale datasets must be improved. Therefore, we will actively explore innovative methods for bird song feature extraction to achieve more accurate and efficient bird song recognition and monitoring in the future, thereby making breakthroughs and developments in ecology.

Funding

This work is supported by National Natural Science Foundation of China (62161040) (62201298), Science and Technology Project of Inner Mongolia Autonomous Region (2021GG0023), Supported by Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region (NJYT22056), Natural Science Foundation of Inner Mongolia Autonomous Region (2021MS06030) and Science and Technology Project of Inner Mongolia Autonomous Region (2023YFSW0006), Supported by the Fundamental Research Funds for Inner Mongolia University of Science and Technology (2023RCTD029).

CRediT authorship contribution statement

Lin Duan: Writing – original draft, Investigation, Formal analysis, Data curation, Conceptualization. **Lidong Yang:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition. **Yong Guo:** Validation, Investigation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Lidong Yang reports financial support was provided by National Natural Science Foundation of China (62161040) (62201298). Lidong Yang reports financial support was provided by Science and Technology Project of Inner Mongolia Autonomous Region (2021GG0023). Lidong Yang reports financial support was provided by Supported by Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region (NJYT22056). Lidong Yang reports financial support was provided by Natural Science Foundation of Inner Mongolia Autonomous Region (2021MS06030). Lidong Yang reports financial support was provided by Science and Technology Project of Inner Mongolia Autonomous Region (2023YFSW0006). Lidong Yang reports financial support was provided by Supported by the Fundamental Research Funds for Inner Mongolia University of Science & Technology (2023RCTD029).

Data availability

The data for this work are published on Kaggle and are available at https://www.kaggle.com/datasets/lllll00/urbansound8k and https://www.kaggle.com/datasets/lllll00/birdsdata.

References

Chen, H., Wang, Y., Guo, J., Tao, D., 2023. Vanillanet: the power of minimalism in deep learning. arXiv:2305.12972.

Clark, M.L., Salas, L., Baligar, S., Quinn, C.A., Snyder, R.L., Leland, D., Schackwitz, W., Goetz, S.J., Newsam, S., 2023. The effect of soundscape composition on bird vocalization classification in a citizen science biodiversity monitoring project. Ecol. Inform. 75, 102065 https://doi.org/10.1016/j.ecoinf.2023.102065.

Ding, X., Zhang, X., Han, J., Ding, G., 2021. Diverse branch block: building a convolution as an inception-like unit arXiv:2103.13425.

Dubey, A.K., Jain, V., 2019. Comparative study of convolution neural network's relu and leaky-relu activation functions. In: Applications of Computing, Automation and Wireless Systems in Electrical Engineering: Proceedings of MARC 2018. Springer, pp. 873–880.

Eldan, R., Shamir, O., 2016. The power of depth for feedforward neural networks. arXiv:1512.03965.

Fischer, S., Edwards, A.C., Garnett, S.T., Whiteside, T.G., Weber, P., 2023. Drones and sound recorders increase the number of bird species identified: a combined surveys approach. Eco. Inform. 74, 101988 https://doi.org/10.1016/j.ecoinf.2023.101988.

Fu, Y., Yu, C., Zhang, Y., Lv, D., Yin, Y., Lu, J., Lv, D., 2023. Classification of birdsong spectrograms based on dr-acgan and dynamic convolution. Ecol. Inform. 77, 102250 https://doi.org/10.1016/j.ecoinf.2023.102250.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial networks. arXiv:1406.2661.

Gupta, S., Jaafar, J., Ahmad, W.W., Bansal, A., 2013. Feature extraction using mfcc. SIPIJ 4, 101–108. https://api.semanticscholar.org/CorpusID:1546219.

Gupta, G., Kshirsagar, M., Zhong, M., Gholami, S., Ferres, J.M.L., 2021. Comparing recurrent convolutional neural networks for large scale bird species classification. Sci. Rep. 11.

Haga, A., Takahashi, W., Aoki, S., Nawa, K., Yamashita, H., Abe, O., Nakagawa, K., 2019. Standardization of imaging features for radiomics analysis. J. Med. Investig. 66, 35–37. https://doi.org/10.2152/jmi.66.35.

Han, X., Peng, J., 2023. Bird sound classification based on ecoc-svm. Appl. Acoust. 204, 109245 https://doi.org/10.1016/j.apacoust.2023.109245.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 770–778.

Himawan, I., Towsey, M.W., Roe, P., 2018. 3d convolutional recurrent neural networks for bird sound detection. In: Workshop on Detection and Classification of Acoustic Scenes and Events.

Hu, S., Chu, Y., Tang, L., Zhou, G., Chen, A., Sun, Y., 2023a. A lightweight multi-sensory field-based dual-feature fusion residual network for bird song recognition. Appl. Soft Comput. 146, 110678 https://doi.org/10.1016/j.asoc.2023.110678.

Hu, S., Chu, Y., Wen, Z., Zhou, G., Sun, Y., Chen, A., 2023b. Deep learning bird song recognition based on mff-scsenet. Ecol. Indic. 154, 110844 https://doi.org/10.1016/j.ecolind.2023.110844.

Imai, S., Sumita, K., Furuichi, C., 1983. Mel log spectrum approximation (mlsa) filter for speech synthesis. Electr. Commun. Jpn. 66, 10–18. https://doi.org/10.1002/ecja.4400660203.

Incze, Jancsó H.B., Szilágyi, Z., Farkas, A., Sulyok, C., 2018. Bird sound recognition using a convolutional neural network. In: 2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY), pp. 000295–000300. https://doi.org/10.1109/SISY.2018.8524677.

Islam, S., Khan, S.I.A., Abedin, M.M., Habibullah, K.M., Das, A.K., 2019. Bird species classification from an image using vgg-16 network. In: Proceedings of the 7th International Conference on Computer and Communications Management. Association for Computing Machinery, New York, NY, USA, pp. 38–42. https://doi.org/10.1145/3348445.3348480.

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.

Kahl, S., Wilhelm-Stein, T., Hussein, H., Klinck, H., Kowerko, D., Ritter, M., Eibl, M., 2017. Large-scale bird sound classification using convolutional neural networks. In: Conference and Labs of the Evaluation Forum.

Kahl, S., Wood, C.M., Eibl, M., Klinck, H., 2021. Birdnet: a deep learning solution for avian diversity monitoring. Ecol. Inform. 61, 101236 https://doi.org/10.1016/j.ecoinf.2021.101236.

Kogan, J.A., Margoliash, D., 1998. Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden markov models: a comparative study. J. Acoust. Soc. Am. 103, 2185–2196. https://doi.org/10.1121/1.421364.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. Imagenet classification with deep convolutional neural networks. Commun. ACM 60, 84–90. https://doi.org/10.1145/3065386.

Kumar, Y., Gupta, S., Singh, W., 2022. A novel deep transfer learning models for recognition of birds sounds in different environment. Soft. Comput. 26, 1003–1023. https://doi.org/10.1007/s00500-021-06640-1.

Lasseck, M., 2018. Audio-based bird species identification with deep convolutional neural networks. In: Conference and Labs of the Evaluation Forum. URL: https://api.semanticscholar.org/CorpusID:51941579.

Liu, Z., Zhou, Y., Yang, H., Liu, Z., 2023. Urban green infrastructure affects bird biodiversity in the coastal megalopolis region of Shenzhen city. Appl. Geogr. 151, 102860 https://doi.org/10.1016/j.apgeog.2022.102860.

Mhaskar, H., Poggio, T., 2016. Deep vs. shallow networks : an approximation theory perspective arXiv:1608.03287.

Nah, S., Hyun Kim, T., Mu Lee, K., 2017. Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3883–3891.
Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788.
Salamon, J., Jacoby, C., Bello, J.P., 2014. A dataset and taxonomy for urban sound research. In: Proceedings of the 22nd ACM International Conference on Multimedia. https://doi.org/10.1145/2647868.2655045.
Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C., 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520.
Sankupellay, M., Konovalov, D., 2018. Bird call recognition using deep convolutional neural network, resnet-50. https://doi.org/10.13140/RG.2.2.31865.31847.
Schlüter, J., 2018. Bird identification from timestamped, geotagged audio recordings. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (Eds.), Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10–14, 2018. CEUR-WS.org.
Sevilla, A., Glotin, H., 2017. Audio bird classification with inception-v4 extended with time and time-frequency attention mechanisms. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (Eds.), Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11–14, 2017. CEUR-WS.org.
Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.
Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M., 2015. Striving for simplicity: the all convolutional net arXiv:1412.6806.
Stastny, J., Munk, M., Juranek, L., 2018. Automatic bird species recognition based on birds vocalization. Eurasip J. Audio Speech Music Process. 2018, 1–7. https://doi.org/10.1186/s13636-018-0143-7.
Swaminathan, B., Jagadeesh, M., Vairavasundaram, S., 2024. Multi-label classification for acoustic bird species detection using transfer learning approach. Ecol. Inform. 80, 102471 https://doi.org/10.1016/j.ecoinf.2024.102471.
Tan, M., Le, Q., 2019. Efficientnet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, PMLR, pp. 6105–6114.
Tang, Q., Xu, L., Zheng, B., He, C., 2023. Transound: hyper-head attention transformer for birds sound recognition. Ecol. Inform. 75, 102001 https://doi.org/10.1016/j.ecoinf.2023.102001.
Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., Belongie, S., 2015. Building a bird recognition app and large scale dataset with citizen scientists: the fine print in fine-grained dataset collection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 595–604.
Wang, K., Yang, F., Chen, Z., Chen, Y., Zhang, Y., 2023a. A fine-grained bird classification method based on attention and decoupled knowledge distillation. Animal 13, 264. https://doi.org/10.3390/ani13020264.
Wang, Y., Chen, A., Li, H., Zhou, G., Yi, J., Zhang, Z., 2023b. A hierarchical birdsong feature extraction architecture combining static and dynamic modeling. Ecol. Indic. 150, 110258 https://doi.org/10.1016/j.ecolind.2023.110258.
Wang, Q., Song, Y., Du, Y., Yang, Z., Cui, P., Luo, B., 2024. Hierarchical-taxonomy-aware and attentional convolutional neural networks for acoustic identification of bird species: a phylogenetic perspective. Ecol. Inform. 80, 102538 https://doi.org/10.1016/j.ecoinf.2024.102538.
Xiangxia, L., Xiaohui, J., Bin, L., 2021. Deep learning method for fine-grained image categorization. J. Front. Comput. Sci. Technol. 15, 1830–1842. https://doi.org/10.3778/j.issn.1673-9418.2103019.
Xie, J., Zhu, M., 2019. Handcrafted features and late fusion with deep learning for bird sound classification. Ecol. Inform. 52, 74–81. https://doi.org/10.1016/j.ecoinf.2019.05.007.
Xie, S., Marzluff, J.M., Su, Y., Wang, Y., Meng, N., Wu, T., Gong, C., Lu, F., Xian, C., Zhang, Y., 2022. The role of urban waterbodies in maintaining bird species diversity within built area of Beijing. Sci. Total Environ. 806, 150430 https://doi.org/10.1016/j.scitotenv.2021.150430.
Xu, J., Li, Z., Du, B., Zhang, M., Liu, J., 2020. Reluplex made more practical: leaky relu. In: 2020 IEEE Symposium on Computers and Communications (ISCC). IEEE, pp. 1–7.
Yoshihashi, R., Kawakami, R., Iida, M., Naemura, T., 2015. Construction of a bird image dataset for ecological investigations. In: 2015 IEEE International Conference on Image Processing (ICIP). IEEE, pp. 4248–4252.
Zhang, X., Zhou, X., Lin, M., Sun, J., 2018. Shufflenet: an extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6848–6856.
Zhang, X., Chen, A., Zhou, G., Zhang, Z., Huang, X., Qiang, X., 2019. Spectrogram-frame linear network and continuous frame sequence for bird sound classification. Ecol. Inform. 54, 101009 https://doi.org/10.1016/j.ecoinf.2019.101009.
Zhang, C.B., Xiao, J.W., Liu, X., Chen, Y.C., Cheng, M.M., 2022. Representation compensation networks for continual semantic segmentation. arXiv:2203.05402.
Zhang, Y., Liu, T., Yu, P., Wang, S., Tao, R., 2024a. Sfsanet: multi-scale object detection in remote sensing image based on semantic fusion and scale adaptability. IEEE Trans. Geosci. Remote Sens. 1–1 https://doi.org/10.1109/TGRS.2024.3387572.
Zhang, Y., Liu, Y., Wu, C., 2024b. Attention-guided multi-granularity fusion model for video summarization. Expert Syst. Appl. 249, 123568 https://doi.org/10.1016/j.eswa.2024.123568.
Zhao, L., Wang, J., Li, X., Tu, Z., Zeng, W., 2017. Deep convolutional neural networks with merge-and-run mappings. arXiv:1611.07718.
Zhu, Y., Newsam, S., 2017. Densenet for dense flow. In: 2017 IEEE International Conference on Image Processing (ICIP). IEEE, pp. 790–794.