



Article

MRDA-MGFSNet: Network Based on a Multi-Rate Dilated Attention Mechanism and Multi-Granularity Feature Sharer for Image-Based Butterflies Fine-Grained Classification

Maopeng Li ¹, Guoxiong Zhou ^{1,*}, Weiwei Cai ¹ , Jiayong Li ¹, Mingxuan Li ¹, Mingfang He ¹, Yahui Hu ² and LiuJun Li ³ 

- ¹ College of Computer & Information Engineering, Central South University of Forestry and Technology, Changsha 410004, China; t20182930@csuft.edu.cn (M.L.); vivitsai@csuft.edu.cn (W.C.); 20191200331@csuft.edu.cn (J.L.); 20191200321@csuft.edu.cn (M.L.); t20162306@csuft.edu.cn (M.H.)
- ² Plant Protection Research Institute, Henan Academy of Agricultural Sciences, Changsha 410125, China; huyah627@163.com
- ³ Department of Civil, Architectural and Environmental Engineering, University of Missouri-Rolla, Rolla, MO 65401, USA; llpwc@umsystem.edu
- * Correspondence: t20060599@csuft.edu.cn

Abstract: Aiming at solving the problems of high background complexity of some butterfly images and the difficulty in identifying them caused by their small inter-class variance, we propose a new fine-grained butterfly classification architecture, called Network based on Multi-rate Dilated Attention Mechanism and Multi-granularity Feature Sharer (MRDA-MGFSNet). First, in this network, in order to effectively identify similar patterns between butterflies and suppress the information that is similar to the butterfly's features in the background but is invalid, a Multi-rate Dilated Attention Mechanism (MRDA) with a symmetrical structure which assigns different weights to channel and spatial features is designed. Second, fusing the multi-scale receptive field module with the depthwise separable convolution module, a Multi-granularity Feature Sharer (MGFS), which can better solve the recognition problem of a small inter-class variance and reduce the increase in parameters caused by multi-scale receptive fields, is proposed. In order to verify the feasibility and effectiveness of the model in a complex environment, compared with the existing methods, our proposed method obtained a mAP of 96.64%, and an F_1 value of 95.44%, which showed that the method proposed in this paper has a good effect on the fine-grained classification of butterflies.

Keywords: butterfly classification; MRDA-MGFSNet; multi-rate dilated attention mechanism; multi-granularity feature sharer



Citation: Li, M.; Zhou, G.; Cai, W.; Li, J.; Li, M.; He, M.; Hu, Y.; Li, L. MRDA-MGFSNet: Network Based on a Multi-Rate Dilated Attention Mechanism and Multi-Granularity Feature Sharer for Image-Based Butterflies Fine-Grained Classification. *Symmetry* **2021**, *13*, 1351. <https://doi.org/10.3390/sym13081351>

Academic Editors: Whoi-Yul Kim and Moonsoo Ra

Received: 22 June 2021

Accepted: 23 July 2021

Published: 26 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Butterflies are an important part of the ecosystem. In recent years, people have gradually realized that certain species of butterflies can provide us with antibiotics that may be vital to saving lives, which shows that they have important medicinal values. At the same time, certain habits of butterflies are also very helpful to scientists who monitor global climate change. Studying the categories of butterflies is conducive to recording and studying their habits and to protect them from extinction. The recognition and classification of butterfly images is an important part of butterfly protection. However, the problem is that the natural environment is complex and harsh, and the background of images on the internet is also highly complex. However, on our actual image recognition process, we usually need slight local differences to separate butterfly subcategories from each other because they have little inter-class differences. For example, the patterns of the *Danaus genutia* and *Monarch butterfly* have high similarities, whose main differences are in the shape of the lateral band of the front wing and the shape of the hind wing, but the differences are relatively small. In order to clearly distinguish the butterfly category even

when the feature differences in the butterfly subcategories are very small, it is necessary to perform fine-grained classification of butterfly images. Fine-grained image classification is also called subcategory recognition, which usually has large inter-class differences, including object posture, brightness, dimension, cover, background and angle compared with inter-class differences. It is not an easy job to achieve fine-grained image classification based on a weakly supervised method, especially when you only have a limited amount of data in each category and there is no other manual label message for butterfly components. Moreover, it is easy for the images to be contaminated by noise when collecting and transmitting butterfly images, which decreases the quality of them or blurs them. At the same time, if we compare fine-grained image classification with coarse-grained image classification, we can find that the fine-grained image classification focuses more on smaller but significant local features to which it is harder to pay attention to due to the slight inter-class differences between subcategories and big intra-class differences. Extracting SIFT (Scale Invariant Feature Transform) [1], HOG (Histogram of Oriented Gradients) [2] or other local image features, and then using the Vector of Local Aggregation Descriptor (VLAD) [3], Fisher vector [4] or other coding models for feature coding, is what early artificial feature-based fine-grained image classification algorithms usually do. Due to the complexity of the selection process of artificial features and the restriction of expressive ability, the classification performance is poor. Nevertheless, the features gained from the Convolutional Neural Network (CNN) [5] have more powerful expressive capabilities than artificial works thanks to the rapid improvement of deep learning. Thus, scientists and researchers have proposed an enormous number of CNN algorithms which promote the improvement of fine-grained image classification algorithms. The CNN has made extraordinary accomplishments in the overall image classification and brought new development directions for fine-grained image classification in recent years and researchers began to choose CNN features for classification. However, due to the different structures of different CNN models, the recognition capabilities and recognition effects of butterfly images are very different.

Therefore, in this paper, we mainly study the following problems: (1) High background complexity of butterfly images. When most algorithms are applied to the actual butterfly classification problem, they are easily influenced by the natural environment. When the CNN performs feature extraction of butterfly images, it is easy to extract “interference features” which are similar to butterfly features, such as dead trees, flowers, plant leaves, rocks, etc. What was mentioned above will seriously interfere the normal butterfly feature recognition and extraction. (2) Similarities between butterfly subcategories. The shape and area of butterfly patterns of the same category are not exactly the same, which brings great difficulties to the feature extraction of neural networks, and it is often difficult to obtain satisfactory precision. Thus, there is an urgent need to perform a more detailed feature extraction in the neural network, and it is also necessary to build a deeper network layer for the neural network to learn more detailed features.

Aiming at solving the two problems mentioned above, Xin et al. [6] used SESADRN to focus on the image features of butterflies, but there were still some errors in focusing on butterfly features. Tan A et al. 2020 [7] used ResNet-101 and the Feature Pyramid Network (FPN) as a feature extraction network to extract butterfly features, and used Mask R-CNN to automatically perform butterfly identification and fine-grained classification using images captured in the natural environment, which achieved good results. However, due to the high complexity of the background, the details extracted by the Region Proposal Network (RPN) are not comprehensive enough, and even some extraction regions do not match, so the classification accuracy is still not enough. Therefore, we propose a fine-grained butterfly classification method based on MRDA-MGFSNet, which can more effectively identify different categories of butterflies.

In our implementation, the main contributions of this paper are as follows:

- (1) Aiming at solving the problems of high background complexity in some butterfly images and difficulty in identifying them caused by their small inter-class variance, we propose MRDA-MGFSNet, designed as follows:
 - a. A Multi-rate Dilated Attention Mechanism with a symmetrical structure suitable for fine-grained butterfly classification is proposed. This module assigns different weights to channel and spatial features to keep more important butterfly features and discard redundant information such as complex natural background information. At the same time, the module integrates dilated convolutions of different rates to expand the visual field of the network and obtain rich context information. It has a good effect on the problem that it is difficult to recognize butterfly images under the interference of a complex background.
 - b. A Multi-granularity Feature Sharer is designed. This module can effectively integrate the overall features of butterflies, save and extract similar spots and patterns and other feature information in butterflies. On the basis of effectively solving the recognition problem of the small inter-class variance of butterfly spots, by connecting a 2-dimensional channel-by-channel convolution and a 3-dimensional point-by-point convolution, it effectively compensates for the increase in parameters caused by the multi-scale structure, saves the training time and improves the efficiency of the network.
- (2) The method of this paper obtains a mAP of 96.64% for the recognition of five categories of butterflies, and the F_1 value reaches 95.44%. It has a good effect on distinguishing butterflies with similar patterns and spots and other features. It has good performance for butterfly classification in complex natural environments, enabling butterfly experts and scholars to better use this technology in the field of butterfly identification to record and study butterfly habits to protect butterflies from extinction, which can finally protect the ecosystem from damage.

The rest of this paper is as follows: Section 2 introduces the related work. Then, Section 3 introduces the materials and methods. Next, Section 4 introduces the experiment and result analysis. Finally, Section 5 concludes.

2. Related Work

In recent years, many experts, scholars and researchers have made great contributions to the issue of the fine-grained classification of butterflies. For example, Zhang J W [8] used 26 morphological features of the color features of the front and the front and rear wings to identify 43 butterfly specimens, thereby obtaining a relatively good precision. Liu F [9] established a radial-based neural network model by using the color features of the front and back of the butterfly, which also achieved a good result. Kaya Y et al. [10] proposed a Gabor filtering and an extreme learning machine (ELM) based image feature extraction method to recognize five categories of butterflies with higher precision. Combined with the artificial wing network classifier, Kaya Y and Kayci L [11] used the RGB color feature of the butterfly wing surface and gray-level symbiosis matrix feature to identify 14 butterfly species in Turkey. The above method was improved in 2014, combining the Gray-Level Co-occurrence Matrix (GLCM) with a polynomial logistic regression realized the automatic identification of 19 categories of butterflies. Kang S H et al. [12] proposed a butterfly recognition method, which is to extend the training set by observing butterfly images from multi-angles. Hernández-Serna et al. [13] devised 15 special features of plants, butterflies and fish from three sides of image morphology, geometric construction and pattern features, using neural networks for training and species identification. Zhou A M et al. [14] used CaffeNet to identify butterfly pattern images whose identification result was not significantly distinct from traditional SVM methods, but the identification accuracy of butterfly ecological images was much higher. The faster R-CNN algorithm was used by Juan-Ying X et al. [15] to identify and classify butterfly images captured in the natural environment, achieving good results. Fine-grained classification also has many applications in agriculture. For example, a surface defect identification of citrus based on the

KF-2D-Renyi and ABC-SVM algorithms was proposed by Aijiao Tan et al. [16] to better detect and classify citrus surface defects reaching an average accuracy of about 98%. Xiao Chen et al. [17] used a new Both-channel Residual Attention Network model(B-ARNet) to identify tomato leaf diseases and achieved an accuracy of about 88%. S. Huang et al. [18] proposed a Non-Local Progressive Average Denoising algorithm combined with a new parallel convolutional neural network to identify peach diseases, and achieved an average accuracy of 88%.

The above results certify that the CNN features can play a better character in fine-grained image classification than traditional methods, whereas butterfly recognition needs more research and its classification accuracy can still improve. Furthermore, the existing butterfly fine-grained classification algorithms are generally based on butterfly specimens, and tend to be simple image classification tasks with weaker ecological expansion capabilities for images, which needs further research. Therefore, we propose a suitable model for butterfly image identification and classification in the natural environment in this paper, namely the fine-grained butterfly classification based on MRDA-MGFSNet.





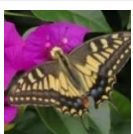
3. Experimental Materials and Methods

3.1. Data Acquisition

The first part of the dataset came from some websites such as Baidu Images, Google Images, personal photography collections, blogs, and social media (5176 images). Some images with poor quality or unclear target objects were removed, leaving 4535 images. Almost all butterfly images in the dataset were captured in the natural environment, only a small part of which were butterfly specimen images. Based on the classification labels in the original websites, the authors re-filtered and categorized the images with reference to professional books. The second part of the dataset came from Kaggle, GitHub, and Google dataset search engines, some websites provided by research reports, etc.; a total of 11,249 images were collected. Although the images are from authoritative websites and have labels, some classification errors were inevitable in the dataset. The author checked them and reclassified them. There were a total of 15,784 images obtained from these two parts of the dataset retrieval mentioned above.

The quantitative distribution of the five categories of butterflies we collected is shown in Table 1.

Table 1. Quantitative distribution of butterfly dataset.

	Example	Number	Proportion (%)
<i>Argynnis hyperbius</i>		3014	19.10
<i>Monarch butterfly</i>		2983	18.90
<i>Polygonia caureum</i>		3205	20.31
<i>Danaus genutia</i>		3311	20.98
<i>Papilio machaon</i>		3271	20.72

The background of the images is basically a complex natural background. It can be seen from Table 1 that the butterflies in the dataset are all in a complex natural background. Due to the fact that there are many similar features, including spots, patterns, shapes, edges, etc., among these five categories of butterflies, we chose them as our research object. Studying the classification of these five categories of butterfly also has reference significance for the study of other butterfly categories. The backgrounds of images in the same category are also very different while the shape and color of the target (butterfly) in each category are very similar.

3.2. MRDA-MGFSNet

In the collected butterfly images, we found that, as shown in Figure 1a–d,f, the butterflies have features that are very similar to the backgrounds. For example, the shape and color of the butterfly in Figure 1a are very close to the stone in the background, and the color and shape of the butterflies in Figure 1b–d,f are almost blended with the flower in the backgrounds. These complex backgrounds were easily recognized as part of the butterfly by the neural network, causing recognition errors and reducing the recognition accuracy. In addition, the patterns of some petals were similar to butterfly patterns. For example, the horizontal and vertical spider web like patterns on the left and right wings of the *Monarch butterfly* had similar patterns to the edges of the petals. It is difficult to distinguish those features by using only spatial attention or channel attention, but if we fuse the two, then the neural network will have a greater possibility to finish this work. For another example, as shown in Figure 1c–e, *Polygonia c-aureum* and *Argynnis hyperbius* have similar spots, both of which are black in color and similar in size. The identification between them may require more information, such as the white spots on the left and right wing ends of *Argynnis hyperbius* may help distinguish it from *Polygonia c-aureum* and the *Monarch butterfly* can be distinguished from the *Polygonia c-aureum* by the black and white spots on the left and right wings. Similarly, as shown in Figure 1c,f,g, they have similar tail shapes, and they need to be distinguished by more patterns and spot features. When encountering these problems, if our network only used a single scale, it may have caused part of the information to be lost. For example, if *Argynnis hyperbius* lacks the features of the white spots, the neural network is very likely to recognize it as *Polygonia c-aureum*.

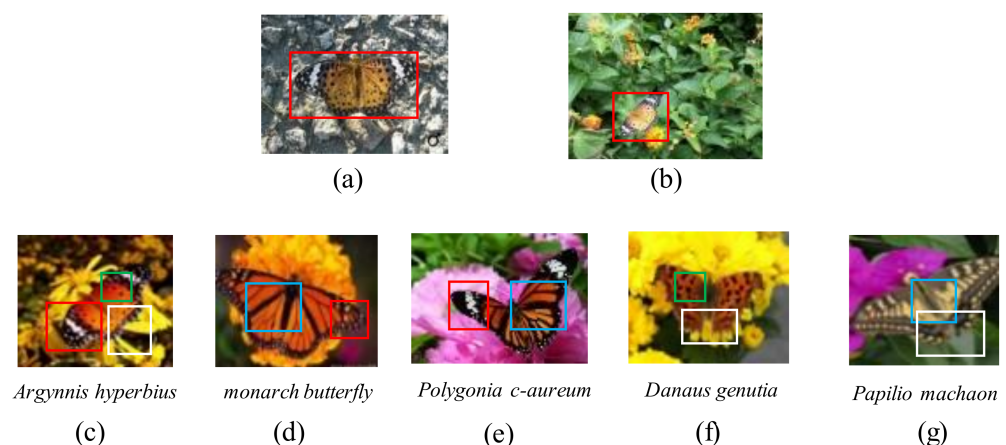


Figure 1. Differences comparison. (a,b) Butterfly images with complex background; (c–g) Butterfly species with similar features.

Therefore, the problem of difficulty in recognition caused by the complexity of the backgrounds and small inter-class variance of butterflies needed to be solved urgently.

In order to solve the above problems, an MRDA-MGFSNet was designed, the core idea of which is to use the MGFS structure to make the main network have multi-scale receptive fields to prevent the loss of subtle features. At the same time, the MRDA module was used to focus on the important features of butterflies, abandon invalid background

information, and effectively reduce the amount of parameter calculations and reduce the training time of the network.

MRDA-MGFSNet has three parts, and the model is defined as follows:

1. The first part was used to extract features, of which there were $64 \ 7 \times 7$ convolution kernels, stride was 2, whose purpose was to quickly extract various edge features and reduce the size of the image to half of the original size. The function of a maxpool of 3×3 size was to retain the main features while reducing the amounts of parameters and calculations, preventing over-fitting, and improving the generalization ability of the model.
2. The second part was composed of 16 MRDA modules and MGFS modules (explained in detail below). The MGFS module was composed of $2 \ 1 \times 1$ convolutions and $4 \ 3 \times 3$ convolutions of different scales, which were used to pay attention to the similar spots and patterns and other small feature information of butterflies. The 3×3 convolution used a two-dimensional channel-by-channel convolution and a three-dimensional point-by-point convolution, and its purpose was to reduce the amount of parameter calculations and speed up network training. The MRDA module first, respectively, passed three dilated convolutions with rate = 1, 2, 3, and then used the channel attention mechanism which consists of a max pooling layer and two 1×1 convolutional layers. Then, we used the spatial attention mechanism which consists of an average pooling layer, a max pooling layer and two 3×3 convolutions. It assigned different weights to channel and spatial features, whose role was to distinguish similar patterns in the butterfly's feature maps and suppress the background information that was similar to the features of the butterfly but was invalid, and enhance the expressive ability of the network. Finally, the feature map obtained in the first layer was added to the module after the attention mechanism, and the PRelu activation function was used to enhance the nonlinear expression ability of the network.
3. In the last part, an average pooling down-sampling layer was connected to a fully connected layer and, finally the, the output was converted into a probability distribution through softmax to obtain the classification result of the butterfly image.

The overall structure of the MRDA-MGFSNet is shown in Figure 2.

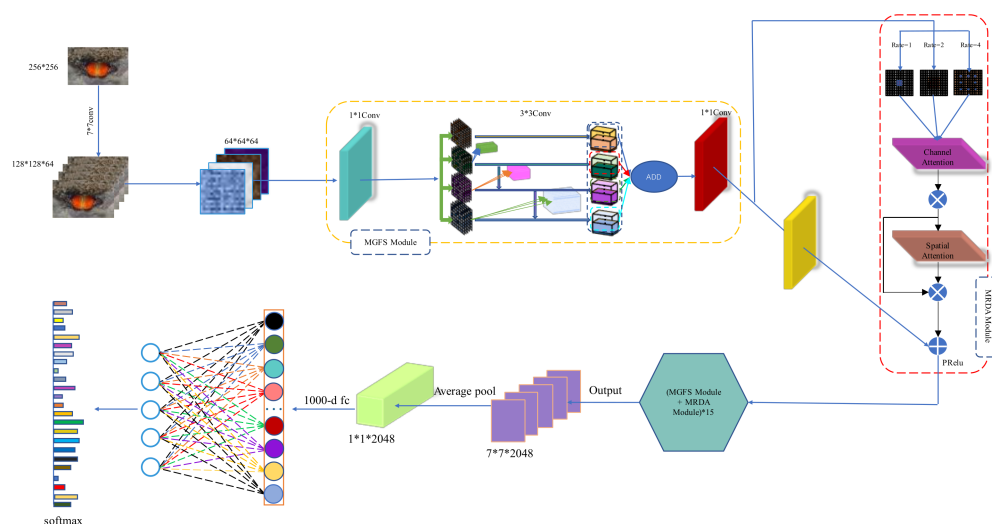


Figure 2. The overall structure of the MRDA-MGFSNet.

3.2.1. Multi-Rate Dilated Attention Mechanism (MRDA)

The spatial attention mechanism [19] pays attention to the importance of the spatial location of the feature (spatial feature), generating spatial attention weights for the output feature map, and strengthening or suppressing different spatial location features based on the feature weights.

The channel attention mechanism [20] focuses on the importance of different feature channels (edge features, because it is a complete image of different C channels convolved by the convolution kernel). In a convolutional neural network, an image feature matrix (H, W, C) is generated after a two-dimensional image is passed through a convolution kernel, where H and W represent the image spatial scale, that is, height and width, and C represents the image feature channel. By modeling the importance of each feature channel, assigning weight to channel features, and strengthening or suppressing different channels according to task requirements.

The attention of the spatial domain is to ignore the information in the channel domain and treat the image features in each channel equally. This approach limits the spatial domain transformation method to the original image feature extraction stage, and cannot be well explained when it applies to other layers of the neural network. The channel attention mechanism directly performs an average pool for all the information in one channel globally, while ignoring the local information in each individual channel.

Due to the detailed features of the butterfly, the complex background has a great influence on it. Therefore, not only the spatial position needs to be paid attention to, but also the images in each channel cannot be treated equally, but should be given different weights. Therefore, combining the two ideas, MRDA was designed:

Spatial attention uses a symmetrical multi-scale structure and uses dilated convolutions with different rates. This structure enables the data stream of butterfly features to be transmitted in a symmetrical manner. The symmetrical characteristic enables MRDA to have a more comprehensive ability to retain the complete butterfly features. Compared with standard convolution, dilated convolution can expand the receptive field of convolution and capture multi-scale information without introducing additional parameters. In this way, the network's visual perception domain can be expanded and rich contextual information can be obtained. At the same time, the PRelu activation function was used to improve the learning convergence effect of the network. The channel attention module used max pooling that can retain more butterfly texture features.

The structure of MRDA is shown in Figure 3.

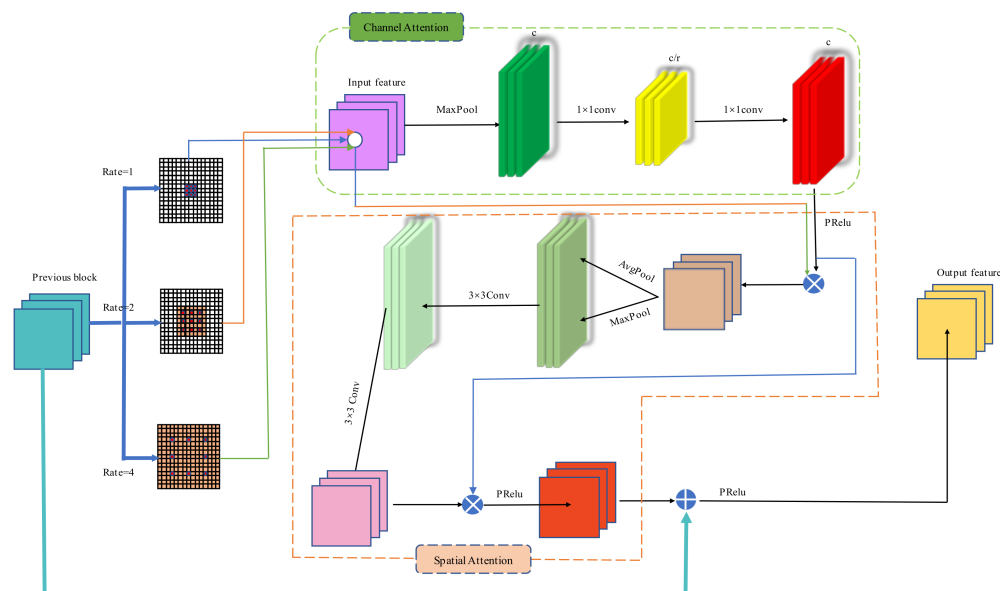


Figure 3. Multi-rate Dilated Attention Mechanism.

MRDA first uses multi-scale and rate = 1, 2, 4 dilated convolution on the input feature maps to expand the receptive field of view and obtain richer butterfly detail information. Then, the attention mechanism we designed was added to obtain three different output feature maps.

The upper part of Figure 3 shows the channel attention module designed in this paper. The max pooling layer used in this paper can retain more butterfly pattern features. After the pooling layer, a 1×1 convolutional layer was added to perform dimensionality reduction operations to reduce the number of channels. A 1×1 convolution kernel was also placed at the output to increase the dimensionality, and the dimensionality reduction and dimensionality increase operations were used to exchange information between channels. Then, the PRelu activation function was used to obtain the result of the channel attention. The definition of channel attention module is shown in Equation (1):

$$CA(F) = \sigma(f(\text{MaxPool}(F))) \quad (1)$$

The lower part of Figure 3 shows the spatial attention module designed in this paper. The feature maps output by the channel attention module was used as the input of the spatial attention module, and the input feature map was the channel compressed using average pooling and max pooling, and then concat operation was performed and two 3×3 convolutions were used to extract receptive fields. Finally, the spatial attention feature maps were generated through the PRelu activation function.

The definition of spatial attention module is shown in Equation (2):

$$DSA(F) = \sigma\left(f^{3 \times 3}(\text{concat}[\text{MaxPool}(F); \text{AvgPool}(F)])\right) \quad (2)$$

The dilated convolution used in the MRDA module in this paper is a method of data sampling on feature maps. It can increase the receptive fields without affecting the resolution to make up for the loss of information. Receptive fields refer to the area size mapped on the original image by the pixels on the feature map output by each layer of the network. The calculation method of the receiving field is shown in Equation (3):

$$r_{i+1}^2 = [(r_i - 1) + (2l + 1)]^2 \quad (3)$$

In Equation (3), r_i represents the side length of the receptive field of the i -th layer, and l represents the coefficient of the dilated convolution.

As shown in Figure 4, in the case of the same core size, different coefficients can lead to different receptive fields. In Figure 4a, the coefficient was 1, which was no different from traditional convolution. In Figure 4b, the coefficient was 2, and the receptive fields were expanded to 7×7 . In Figure 4c, the coefficient was 4, and the receptive fields were expanded to 15×15 .

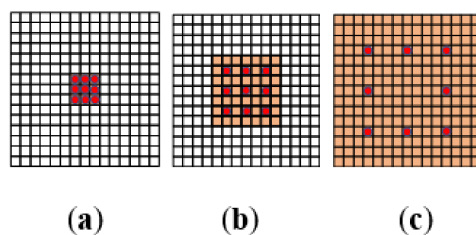


Figure 4. Examples of dilated convolution with different coefficients. (a) The coefficient is 1; (b) the coefficient is 2; (c) the coefficient is 4.

Dilated convolution makes convolution calculations have a wider view and can capture longer dependencies at the same computational cost. Dilated convolution is suitable for situations that require a wider view and do not use multiple convolutions or larger convolution kernels. Therefore, in the feature fusion and down-sampling part of the network, we chose an expanded convolution with a convolution kernel size of 3×3 to increase the receptive field without changing the size of the feature map to improve the efficiency of the feature extraction of the network.

3.2.2. Multi-Granularity Feature Sharer (MGFS)

The use of MGFS was to solve the limitation of identifying butterflies on a single scale, improve the adaptability of the network, effectively integrate the more comprehensive features of butterflies, and save and extract information such as similar spots between butterflies. The structure of Multi-granularity Feature Sharer is show in Figure 5 and the description is as follows:

- (1) Generally, larger convolution kernels have a stronger ability to perceive large target objects, and small-size convolution kernels are better at extracting features of small targets. However, the quality of butterfly images varies. Some were butterfly specimens and had few backgrounds information, and some had complex backgrounds and the targets were not easy to find. Therefore, we increased branches of different sizes of receptive fields and used convolution kernels with sizes of 3×3 , 5×5 , and 7×7 to improve the recognition accuracy.
- (2) The MGFS structure divided the feature maps obtained after 1×1 convolution into 4 scales on average, of which 3×3 convolution used depthwise separable convolution to reduce the amount of parameter and calculation.
- (3) Using the PRelu activation function to replace the ReLU or Sigmoid activation function to improve the learning convergence effect of the network.
- (4) As the number of butterfly images was relatively small, the group normalization (GN) that was not affected by the batch size was used to replace the batch normalization (BN) layer to improve the network convergence effect, and the batch size was set to 10.

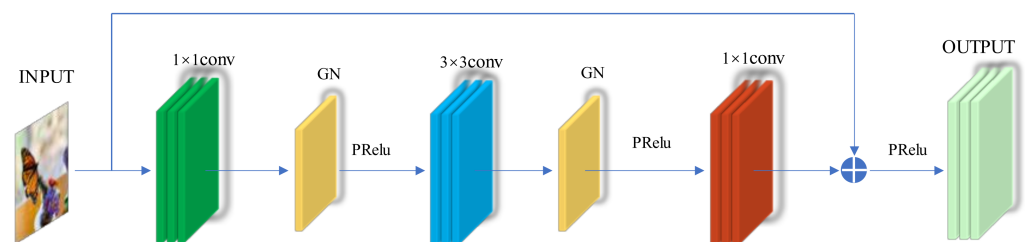


Figure 5. Multi-granularity Feature Sharer structure composition.

In this paper, the structure enabled the neural network to learn more detailed features of butterflies, such as spots of similar color and size and their spatial distribution, and greatly improved the accuracy of the recognition of subtle features among butterflies. Figure 6 is a schematic diagram of the detailed design of the MGFS structure.

The multi-scale receptive field also brought along the problem of an increase in the amount of calculation parameters. At the same time, the multi-scale structure as well as multiple 1×1 and 3×3 small-size convolution kernel structures were used in this paper; the network was also deeper. Therefore, we introduced the depthwise separable convolution used by F. Chollet [21] to construct convolutional neural networks whose work enables large and complex neural networks to run more efficiently. As shown in Figure 7, the idea of depthwise separable convolution was to separate the traditional convolution operation into two steps: first, depthwise convolution was performed, that is, a one-to-one 2-D convolution was performed on each channel of the input feature map to reduce parameter calculations; then, using the 1×1 size convolution kernel to continue the traditional convolution (3-D convolution) operation to combine the features of each channel, also known as point-wise convolution. The structure of the depthwise separable convolution is shown in Figure 6.

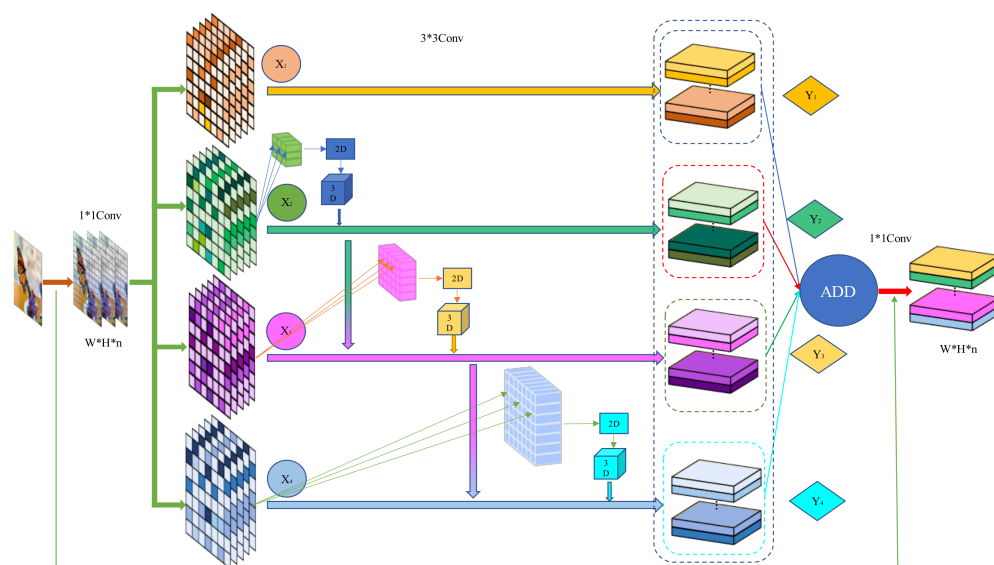


Figure 6. Detailed structure of Multi-granularity Feature Sharer.

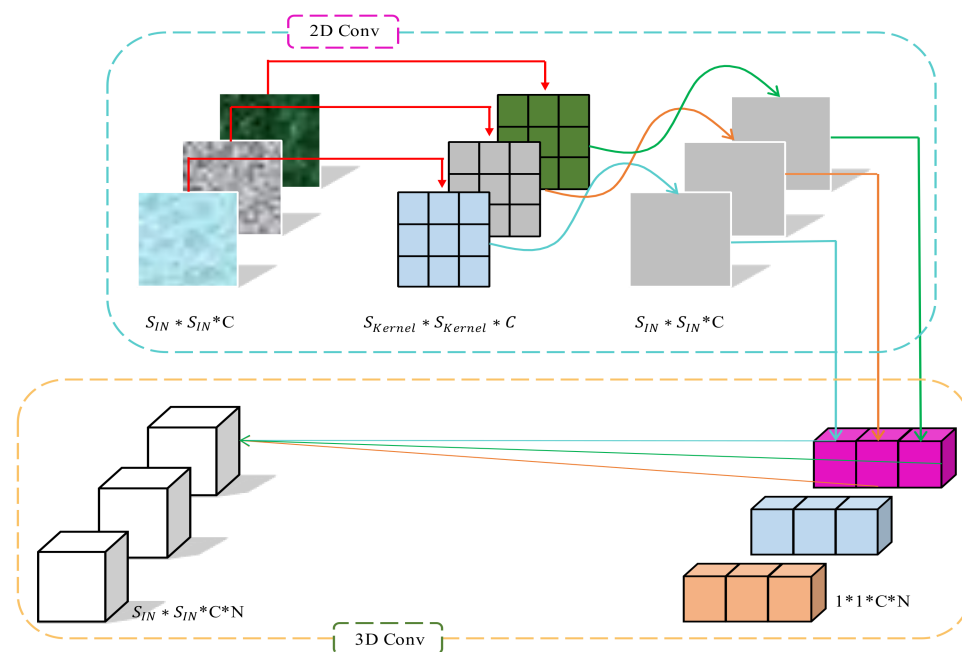


Figure 7. Depthwise separable convolution structure.

Suppose that the size of the input feature map is $S_{IN} * S_{IN}$ the number of channels is C , the size of the convolution kernel is $S_{Kernel} * S_{Kernel}$ and there are a total of N , the calculation amounts of traditional convolution and depthwise separable convolution are shown in Equations (4) and (5):

$$\text{Traditional} = S_{IN} * S_{IN} * C * N * S_{Kernel} * S_{Kernel} \tag{4}$$

$$\text{DSC} = S_{IN} * S_{IN} * C * S_{Kernel} * S_{Kernel} + S_{IN} * S_{IN} * C * N \tag{5}$$

Therefore, the calculation ratio of depthwise separable convolution and traditional convolution is:

$$\text{ratio} = \frac{S_{IN} * S_{IN} * C * S_{Kernel} * S_{Kernel} + S_{IN} * S_{IN} * C * N}{S_{IN} * S_{IN} * C * N * S_{Kernel} * S_{Kernel}} = \frac{1}{N} + \frac{1}{S_{Kernel} * S_{Kernel}} \tag{6}$$

It can be seen that the reduction in the calculation amount of the depthwise separable convolution is related to the size of the two-dimensional convolution kernel used $S_{Kernel} * S_{Kernel}$ and the number N of the three-dimensional convolution kernel. In practice, the depthwise separable convolution generally uses a 3×3 size convolution kernel. If the output channel was 64, the calculation amount of the depthwise separable convolution can be calculated by Equation (6), which is only about 1/10 of the traditional convolution parameter calculation amount.

4. Results and Analysis

4.1. Experimental Environment and Preparation

The hardware information is as follows: the processor was AMD4800h, the GPU was RTX2060, and the video memory was 6 GB.

The software information is as follows: CUDA Toolkit 10.0; CUDNN V7.5.0; Pycharm2020.1, PyTorch; MATLAB (R2019a).

The unified input size of images was 256×256 , and a total of 15,784 images were obtained. We used the 10-fold cross-validation method for training in this paper so the images were divided into 11,364 as the training set and 1263 as the validation set according to a 9:1 ratio in the method. Additionally, the number of images of the test set was 3157.

4.2. Results and Analysis

As the images of our dataset were not enough, for the accuracy and reliability of the model, we used the 10-fold cross-validation method for training in this paper. Cross-validation is also called loop estimation. Most of the samples were taken out of a given modeling sample to build a model, and a small part of the sample was left for prediction with the newly established model, and the forecast error of this small part of the sample was calculated and their sum of squares was recorded. This process continued until all samples were predicted once and only once. For example, using the 10-fold cross-validation divided the butterfly dataset into ten parts, and took turns to train nine parts and one part for validation, and the average of the results of 10 times was used as an estimate of the accuracy of the algorithm. We repeated the 10-fold cross-validation 10 times in this paper to obtain a higher accuracy and reliability. The advantage of this method is that it repeatedly uses randomly generated sub-samples for training and verification at the same time, and each result is verified once. In order to verify the recognition effect of MGFS on butterfly subtle differences such as similar spots, we used the 10-fold cross-validation method to allocate 90% of the training sample and the verification sample to 10%. The dataset included all the images of five categories of butterflies: *Argynnis hyperbius*, *Monarch butterfly*, *Polygonia caureum*, *Danaus genutia* and *Papilio machaon*. Any other experimental environment was the same.

a. Ablation experiment

First, the experiment adopted the 10-fold cross-validation method, the accuracy of all ablation experiments was the average value of ten times of the 10-fold cross-validation.

We used the MRDA-MGFSNet (Basic + MGFS + MRDA) network for butterfly fine-grained classification, and then tested and recognized the butterfly categories. In this paper, under the same experimental environment, we used CNN [5], AlexNet-fc6 [22], VGG16 [23], DenseNet-161 [24], Resnet-50 [25] and other models to train our butterfly dataset. In addition, we called the network with an ordinary convolutional layer and residual structure except for MGFS and MRDA in this paper, as the Basic network and multi-scale network have the same multi-scale structure as MGFS and MRDA on the basis of the Basic network. The training loss of CNN is shown in the black solid line in Figure 8, and the training loss of ResNet-50 is shown in the blue solid line in Figure 8. The training loss of the AlexNet-fc6 is shown by the solid green line in Figure 8, the loss of VGG-16 is shown by the solid yellow line in Figure 8, and the method used in this paper is shown by the solid gray line in Figure 8. In this Figure, the horizontal axis represents the number of training epochs, and the vertical axis represents the loss. It can be seen that, compared with

the CNN, AlexNet, VGG, and ResNet models, the MRDA-MGFSNet-based model had a lower loss value and a better training effect.

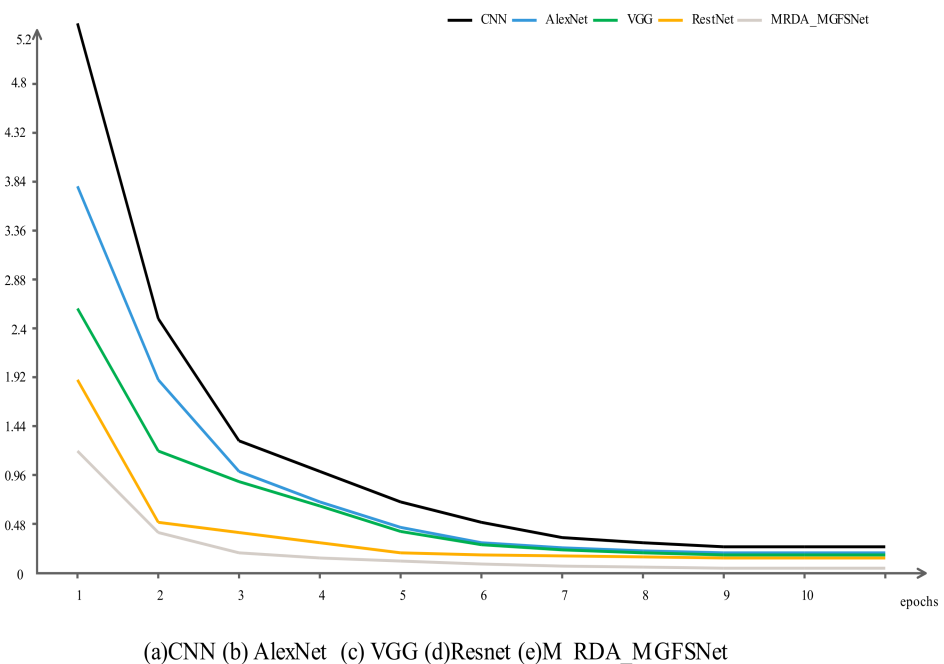


Figure 8. Loss curve of various methods.

Table 2 shows the test accuracy of each butterfly category corresponding to each method and the overall accuracy (based on the maximum category score higher than 0.5). From the ablation experiment in Table 2, it can be clearly seen that the MGFS and MRDA proposed in this paper improved the recognition accuracy of various butterflies to a certain extent. The accuracy of *Argynnis hyperbius*, *Danaus genutia*, *Papilio machaon* was basically above 95%, while the classification accuracy of the *Monarch butterfly* and *Polygonia caureum* was relatively low. This is because simple specimen images accounted for a relatively larger proportion of the images of the first three categories of butterflies mentioned above than the others', and the image quality was relatively good. On the contrary, the images with complex backgrounds of the *Monarch butterfly* accounted for a larger proportion, and the image backgrounds of *Polygonia caureum* were also more complicated, so the training results of these two categories of butterflies were relatively poor. Compared with the basic network architecture used in this paper, MRDA had better results (+5.19%, +5.45%) for the *Monarch butterfly* and *Polygonia caureum*, which had more complex backgrounds that were similar to their own features. MGFS had a good recognition effect (+4.84%, +4.43%, +3.64%) for the three categories of butterflies, *Polygonia caureum*, *Papilio machaon* and *Argynnis hyperbius*, whose spots were important information and patterns were few. The superposition of MGFS and MRDA made the network slightly improve the comprehensive recognition accuracy of the two problems, which also proves that the two structures proposed in this paper had different effects for each problem. The effect of the fusion was slightly reduced, but the comprehensive performance improved. At the same time, we also designed a separate multi-scale network and a complete structure of the MGFS network ablation experiment in the experiment. In order to obtain a richer receptive field, a separate multi-scale calculation increases the amount of parameter calculations, which slows down the training speed of the network, and the addition of depthwise separable convolutions forms the MGFS module. Under the same experimental environment, compared with the multi-scale network training time, the network training time of the MGFS module was reduced by 1 h 38 min 29 s under the same experimental environment conditions. Experiments show

that the MGFS module composed with depthwise separable convolution could effectively reduce the training time of the network to save experimental resources.

Table 2. Individual and overall accuracy and training time of each method.

Network	<i>Argynnis hyperbius</i>	<i>Monarch butterfly</i>	<i>Polygonia caureum</i>	<i>Danaus genutia</i>	<i>Papilio machaon</i>	Overall	Training Time
CNN	84.41%	68.84%	76.29%	83.69%	83.94%	79.57%	4 h 28 min 03 s
AlexNet-fc6	87.56%	71.52%	78.63%	86.71%	85.17%	82.04%	4 h 43 min 24 s
VGG-16	87.40%	69.85%	78.63%	85.35%	85.32%	81.44%	5 h 54 min 27 s
DenseNet-161	90.38%	76.55%	86.12%	87.31%	87.61%	85.68%	5 h 31 min 16 s
Resnet-50	91.21%	82.58%	82.68%	87.76%	85.32%	85.90%	6 h 15 min 04 s
Basic	91.38%	88.11%	85.96%	91.39%	90.83%	89.55%	6 h 01 min 24 s
Multi-scale	93.37%	89.45%	86.12%	91.84%	91.44%	90.44%	7 h 21 min 11 s
MGFS	94.53%	92.46%	90.80%	93.81%	94.65%	93.26%	5 h 42 min 42 s
MRDA	95.02%	93.30%	91.41%	95.47%	95.26%	94.11%	5 h 51 min 16 s
NTS-Net [26]	90.88%	88.27%	91.11%	91.84%	88.38%	90.12%	8 h 12 min 50 s
DFL-Net [27]	92.70%	90.12%	89.24%	91.09%	91.28%	90.88%	5 h 56 min 43 s
BSNet [28]	93.53%	91.12%	90.95%	91.24%	92.66%	91.89%	6 h 04 min 52 s
MGFS + MRDA (MRDA-MGFSNet)	96.85%	93.63%	93.14%	97.13%	96.48%	95.47%	5 h 58 min 20 s

Class Activation Mapping (CAM) can give a good visual interpretation of the classification results, and can achieve weakly supervised positioning of the target object.

As shown in Figure 9, we used Gradient-weighted Class Activation Mapping (Grad-CAM) to visually explain the performance of MRDA-MGFSNet on butterfly classification. It can be seen that our MRDA-MGFSNet model could locate the area of the butterfly in the image very well.



Figure 9. Visualization of the proposed MRDA-MGFS's Gradient-weighted Class Activation Mapping (Grad-CAM).

b. The latest methods comparison experiment

As a matter of fact, it is one-sided and unconvincing to only rely on classification accuracy to determine whether the model is truly effective. Therefore, we accurately calculated the F_1 value of each network in this paper. In order to verify the effectiveness of the butterfly recognition training model, in this paper, we used two indexes: the recall rate and accuracy as evaluation index. We selected F_1 as one of the evaluation indicators of butterfly recognition results. F_1 is a measurement function of accuracy and recall rate, defined as the following formulas:

$$F_1 = \frac{2PR}{P + R} \quad (7)$$

$$P = \frac{TP}{TP + FP} \quad (8)$$

$$R = \frac{TP}{TP + FN} \quad (9)$$

In Equation (7), P represents precision; R represents recall rate.

TP represents the number of samples that are actually butterflies, and the model predicts that the sample is a butterfly (detecting a positive sample as a positive sample). In Equation (8), FP represents the number of samples that are not actually butterflies, but the model predicts that the sample is a butterfly (tests negative samples as positive samples). In Equation (9), FN represents the number of samples that are actually butterflies, but the model did not predict them as butterflies (no positive samples were detected as positive samples). As shown in Table 3, the experimental results show that the F_1 value of the classification model in this paper reached the expected level of the experiment, proving that what we discussed and analyzed above is correct.

Table 3. F_1 value of each method.

Network	<i>Argynnis hyperbius</i>	<i>Monarch butterfly</i>	<i>Polygonia caureum</i>	<i>Danaus genutia</i>	<i>Papilio machaon</i>	Average F_1 Value
CNN	84.51%	68.10%	76.33%	82.68%	83.46%	79.01%
AlexNet-fc6	87.69%	72.02%	78.34%	85.98%	85.63%	81.93%
VGG-16	87.66%	69.27%	78.94%	86.06%	85.41%	81.47%
DenseNet-161	90.72%	75.88%	86.26%	87.85%	87.14%	85.57%
Resnet-50	91.36%	81.73%	83.06%	87.46%	85.92%	85.91%
Basic	90.86%	87.28%	85.49%	92.67%	89.97%	89.25%
MGFS	92.76%	92.54%	92.68%	93.88%	94.29%	93.23%
MRDA	93.32%	93.45%	93.46%	95.11%	95.04%	94.08%
NTS-Net	88.10%	88.27%	91.25%	92.54%	90.17%	90.07%
DFL-Net	89.44%	90.34%	90.08%	91.85%	92.56%	90.85%
BSNet	90.17%	91.43%	92.32%	91.31%	94.17%	91.88%
MRDA-MGFSNet	95.42%	94.50%	94.92%	95.83%	96.55%	95.44%

In the field of machine learning, the confusion matrix is also called the possibility table or error matrix. It is a specific matrix used to visualize the performance of the algorithm, usually supervised learning (unsupervised learning, usually matching matrix). Each column represents the predicted value, and each row represents the actual category. This is very important, because in the real-world classification, the TP value and the FP value are the most direct indicators that ultimately determine whether the classification is correct, and the F_1 value is a comprehensive manifestation of these two indicators.

As shown in Figure 10, we compared the experimental results of the MRDA-MGFSNet-based model for each category with the experimental results of some of the state-of-the-art models such as NTS-Net [26], DFL-Net [27] and BSNet [28].

Argynnis hyperbius, the *Monarch butterfly* and *Polygonia caureum* have very similar patterns, spots and shapes. Similarly, *Danaus genutia* and *Argynnis hyperbius*, *Papilio machaon* and the *Monarch butterfly* also have similar patterns, shapes, colors and other features. Therefore, it can be clearly seen from the confusion matrix of each method that the network still had a great chance of misclassifying them.

NTS-Net proposes a novel training paradigm, and enables the navigator to detect the area with the largest amount of information under the guidance of the teaching device. However, this self-supervised learning mode is prone to missing or wrong extraction in the extraction of more subtle butterfly spots and other features. Therefore, it has higher FN values on the recognition of the three categories of *Argynnis hyperbius*, *Monarch butterfly*, *Polygonia caureum*, of butterflies with very similar patterns, spots and shapes.

DFL-Net proposes a discriminative mid-level patch, which uses a 1×1 convolution kernel as small “patch detectors” to design an asymmetric, multi-branch structure to utilize patch-level information. Although this kind of classification avoids the trade-off between recognition and positioning, it tends to be more on the classification itself and ignores the recognition and removal of background information. This leads to the result that it can

better identify butterflies with similar features. The FP value of *Danaus genutia* and *Papilio machaon* is low, but the recognition effect of the *Monarch butterfly* and *Polygonia caureum* with a complex background is not good enough.

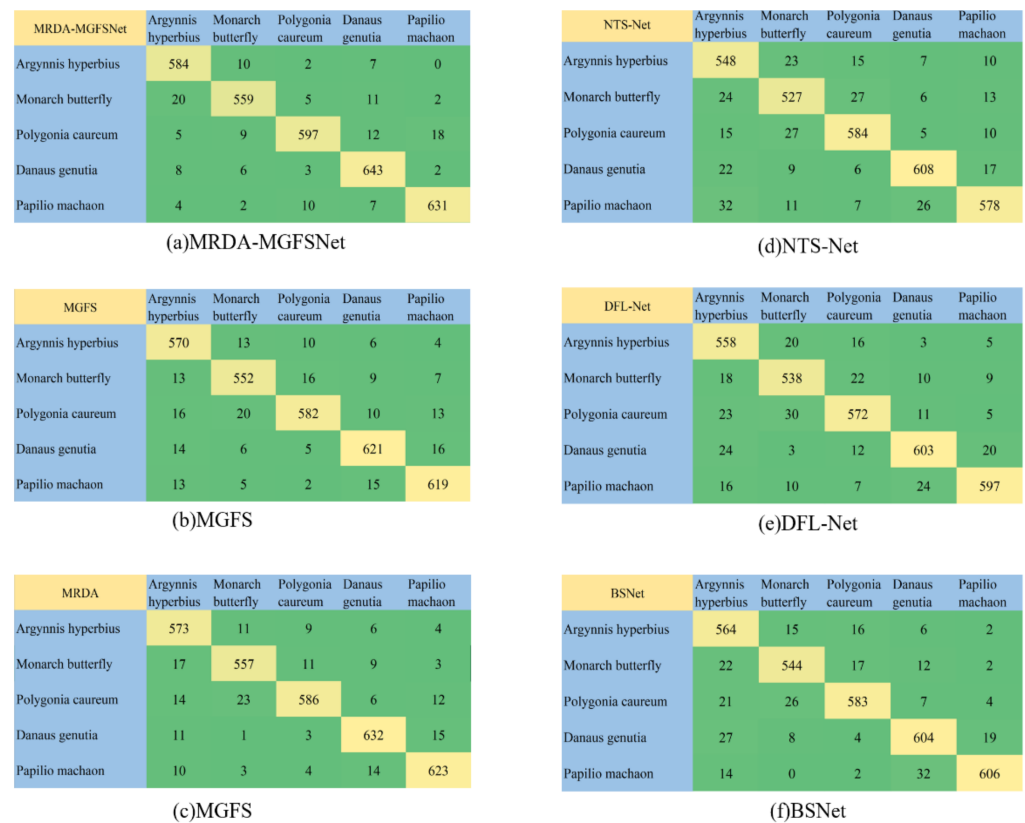


Figure 10. Confusion matrix of partial network.

BSNet is composed of an optical band attention module (BAM), optical band weighting (BRW) and reconstruction network (RecNet). BS-Net-Conv improves the utilization of spectrum-space information in HSI. After these three modules, the background information of the image can be better filtered out to obtain a good classification effect. This network can also achieve better results in the classification of butterfly images. It can be seen from the confusion matrix that, except for *Argynnis hyperbius*, it has low FN values for other species of butterflies. However, this is still not enough in real butterfly recognition.

It can be seen from the confusion matrix that the MRDA-MGFSNet proposed in this paper had a better butterfly classification effect. Compared with NTS-Net, DFL-Net and BSNet, the FP value (FN value in the same column) of the MRDA network for the *Monarch butterfly* and *Polygonia caureum* with a large proportion of complex and similar backgrounds in their images was significantly reduced. This is because the MRDA algorithm is more inclined to retain and pay attention to the butterfly characteristics and discard useless background information. The decrease in FP value increases the recall rate of these two categories of butterflies. According to Equation (7), with the same accuracy, the increase in recall rate eventually increases the F_1 value. Similarly, compared to the NTS-Net, DFL-Net and BSNet, the MGFS network had an effective inhibitory effect on the FP values of *Argynnis hyperbius*, *Polygonia caureum*, and *Papilio machaon*, which have spots as important information and have fewer patterns. This is because the MGFS algorithm has a more scale-feature-sharing mechanism that can retain more subtle features such as butterfly patterns and spots. In the end, the suppression of the FP value can achieve the effect of increasing the F_1 value. The network that combines MRDA and MGFS is more powerful in terms of overall performance, and has a good effect on improving the F_1 value. In addition, as shown in Table 4, our model achieved an mAP0.5 value of 96.64%, which proves that our

model had a good effect in butterfly classification once again. Experiments have proved that under the same experimental environment, our algorithm is more suitable for butterfly classification than some other latest fine-grained classification models.

Table 4. mAP0.5(%) of each method.

Network	mAP(%)
CNN	78.45
AlexNet-fc6	81.24
VGG	82.03
DenseNet-161	84.38
Resnet-50	86.55
NTS-Net	90.23
DFL-Net	89.27
BSNet	91.09
MRDA-MGFSNet (ours)	96.64

c. Noise processing capability experiment

Of course, in most cases, butterfly images taken in nature are always affected by many factors, such as noise. A stable model should also achieve a good precision and F_1 value for images with noise. Therefore, we designed a noise processing capability experiment.

First, as shown in Figure 11, we added different degrees of noise to the 3157 butterfly images in the test set to obtain a new dataset contaminated by noise.

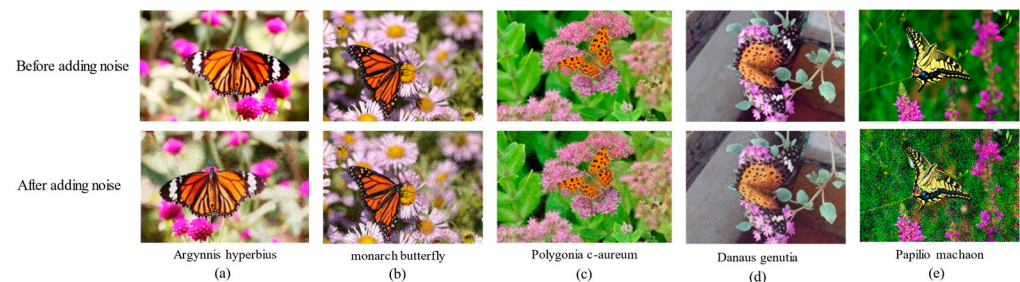


Figure 11. Noise pollution dataset.

Then, we used the trained model to perform classification tests on the above two datasets, respectively, and obtained Table 5.

Table 5. Accuracy and F_1 value of two circumstances (%).

Network Accuracy/ F_1	<i>Argynnis hyperbius</i>	<i>Monarch butterfly</i>	<i>Polygonia caureum</i>	<i>Danaus genutia</i>	<i>Papilio machaon</i>	Overall
Before adding noise	96.19/97.22	93.80/94.52	92.98/93.35	97.58/97.41	97.25/96.57	95.60/95.81
After adding noise	94.69/94.88	92.46/93.03	90.80/91.28	94.11/93.84	95.72/95.16	93.57/93.64

It can be seen from the experimental results that the model had a certain decrease in accuracy and F_1 value after adding noise. The overall recognition accuracy dropped by 2.03%, and the average F_1 value dropped by 2.17%, which shows that noise can indeed have a certain impact on the recognition of the model, but both the overall recognition accuracy and F_1 value still exceeded 93%. It can be seen that the model has good robustness, and has the potential to deal with noisy images to a certain extent.

5. Conclusions

Aiming at solving the problems of high background complexity in some butterfly images and difficulty in identifying them caused by their small inter-class variance, we pro-

posed a new fine-grained butterfly classification architecture in this paper which achieved good performance in identifying butterfly species. The discussion is as follows:

- a. Ablation experiments showed that the MRDA had better results (+5.19%, +5.45%) for butterflies which have more complex backgrounds that are similar to their own features; the MGFS had a good recognition effect (+4.84%, +4.43%, +3.64%) for the three categories of butterflies whose spots are important information and patterns are few; under the same experimental conditions, compared with the multi-scale network, the training time of the MGFS module (with depthwise separable convolution module) was reduced by 1 h 38 min 29 s. The above results show that the two architectures proposed in this paper achieved the expected experimental results, and can effectively solve the problems of complex backgrounds and small inter-class variance between butterflies.
- b. Compared with some of the current state-of-the-art fine-grained classification methods, our mAP reached 96.64%, and the average F_1 value reached 95.44%. The designed butterfly fine-grained classification method can achieve better performance. This method had good effects and obvious advantages in identifying different patterns and spots in different butterfly images and removing complex interference information in the background. After the noise processing capability experiment, our model had an accuracy of 93.57% and an F_1 value of 93.64%, which is only 2.03% lower than the accuracy before noise was added, and the F_1 value was 2.17% lower, showing that our model has good potential to deal with noisy images. It can be well applied to the butterfly recognition to better protect the important butterflies for ecological protection in the future.

The butterfly recognition model proposed in this paper greatly improved the effect of fine-grained butterfly classification in a complex background. However, considering that the dataset contained few butterfly categories, it will be expanded in the future to improve the generalization ability of the recognition model. In addition, our model was still an early research prototype, and the number of butterfly images in the dataset was still insufficient. In the future, we need to collect more datasets to improve the recognition accuracy and further improve the performance of the model, so that our model can play a more important role in the field of butterfly protection and ecosystem protection.

Author Contributions: Conceptualization, M.L. (Maopeng Li) and W.C.; Methodology, M.L. (Maopeng Li); Software, M.L. (Maopeng Li); Writing—Original draft preparation, M.L. (Maopeng Li); Project administration, G.Z.; Supervision, G.Z. and L.L.; Funding acquisition, G.Z.; Investigation, J.L.; Validation, M.L. (Mingxuan Li); Formal analysis, M.H.; Resources, Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Changsha Municipal Natural Science Foundation (Grant No. kq2014160), in part by the National Natural Science Foundation in China (Grant No. 61703441), in part by the key projects of the Department of Education Hunan Province (Grant No. 19A511), and in part by the Hunan Key Laboratory of Intelligent Logistics Technology (2019TP1015).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to partial authors' disagreement.

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

References

1. Lowe, D.G. Object recognition from local scale-invariant features. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 20–27 September 1999; Volume 2, pp. 1150–1157. [[CrossRef](#)]
2. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision & Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.

3. Jégou, H.; Douze, M.; Schmid, C.; Pérez, P. Aggregating local descriptors into a compact image representation. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3304–3311. [[CrossRef](#)]
4. Sánchez, J.; Perronnin, F.; Mensink, T.; Verbeek, J. Image Classification with the Fisher Vector: Theory and Practice. *Int. J. Comput. Vis.* **2013**, *105*, 222–245. [[CrossRef](#)]
5. Le Cun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput.* **1989**, *1*, 541–551. [[CrossRef](#)]
6. Xin, D.; Chen, Y.-W.; Li, J. Fine-Grained Butterfly Classification in Ecological Images Using Squeeze-And-Excitation and Spatial Attention Modules. *Appl. Sci.* **2020**, *10*, 1681. [[CrossRef](#)]
7. Tan, A.; Zhou, G.; He, M. Rapid Fine-Grained Classification of Butterflies Based on FCM-KM and Mask R-CNN Fusion. *IEEE Access* **2020**, *8*, 124722–124733. [[CrossRef](#)]
8. Zhang, J.W. *Automatic Identification of Butterflies Based on Computer Vision Technology*; China Agriculture University: Beijing, China, 2006. [[CrossRef](#)]
9. Liu, F. *The Application of Wings' Color Characters in Butterfly Species Automatic Identification*; China Agricultural University: Beijing, China, 2007. [[CrossRef](#)]
10. Kaya, Y.; Kayci, L.; Tek in, R. A computer vision system for the automatic identification of butterfly species via Gabor-filter-based texture features and extreme learning machine: GF+ ELM. *TEM J.* **2013**, *2*, 13–20.
11. Kaya, Y.; Kayci, L. Application of artificial neural network for automatic detection of butterfly species using color and texture features. *Vis. Comput.* **2013**, *30*, 71–79. [[CrossRef](#)]
12. Kang, S.-H.; Cho, J.-H.; Lee, S.-H. Identification of butterfly based on their shapes when viewed from different angles using an artificial neural network. *J. Asia-Pac. Entomol.* **2014**, *17*, 143–149. [[CrossRef](#)]
13. Hernández-Serna, A.; Jiménez-Segura, L.F. Automatic identification of species with neural networks. *PeerJ* **2014**, *2*, e563. [[CrossRef](#)] [[PubMed](#)]
14. Zhou, A.M.; Ma, P.P.; Xi, T.Y.; Jiang-Ning, W.; Jin, F.; Ze-Zhong, S.; Yu-Lei, T.; Qing, Y. Automatic identification of butterfly specimen images at the family level based on deep learning method. *Acta Entomol. Sin.* **2017**, *60*, 1339–1348. [[CrossRef](#)]
15. Juan-Ying, X.; Qi, H.; Ying-Huan, S.; Peng, L.; Jing, L.; Zhuang, F.; Zhang, J.; Tang, X.; Xu, S. The automatic identification of butterfly species. *J. Comput. Res. Dev.* **2018**, *55*, 1609–1618. [[CrossRef](#)]
16. Tan, A.; Zhou, G.; He, M. Surface defect identification of Citrus based on KF-2D-Renyi and ABC-SVM. *Multimed. Tools Appl.* **2021**, *80*, 9109–9136. [[CrossRef](#)]
17. Chen, X.; Zhou, G.; Chen, A.; Yi, J.; Zhang, W.; Hu, Y. Identification of tomato leaf diseases based on combination of ABCK-BWTR and B-ARNet. *Comput. Electron. Agric.* **2020**, *178*, 105730. [[CrossRef](#)]
18. Huang, S.; Zhou, G.; He, M.; Chen, A.; Zhang, W.; Hu, Y. Detection of Peach Disease Image Based on Asymptotic Non-Local Means and PCNN-IPELM. *IEEE Access* **2020**, *8*, 136421–136433. [[CrossRef](#)]
19. Zhu, X.; Cheng, D.; Zhang, Z.; Lin, S.; Dai, J. An Empirical Study of Spatial Attention Mechanisms in Deep Networks. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Institute of Electrical and Electronics Engineers (IEEE), Seoul, Korea, 27 October–2 November 2019; pp. 6687–6696.
20. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [[CrossRef](#)]
21. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807. [[CrossRef](#)]
22. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural. Inf. Process. Syst.* **2012**, *25*, 1097–1105. [[CrossRef](#)]
23. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
24. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 26 July 2017; pp. 4700–4708.
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
26. Yang, Z.; Luo, T.; Wang, D.; Wang, D.; Hu, Z.; Gao, J.; Wang, L. Learning to Navigate for Fine-Grained Classification. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2018.
27. Wang, Y.; Morariu, V.I.; Davis, L.S. Learning a Discriminative Filter Bank within a CNN for Fine-grained Recognition. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4148–4157.
28. Li, X.; Wu, J.; Sun, Z.; Ma, Z.; Cao, J.; Xue, J.-H. BSNet: Bi-Similarity Network for Few-shot Fine-grained Image Classification. *IEEE Trans. Image Process.* **2021**, *30*, 1318–1331. [[CrossRef](#)]