International Journal of Frontiers in Engineering Technology
ISSN 2706-655X Vol.5, Issue 2: 31-38, DOI: 10.25236/IJFET.2023.050206

A study on fine-grained image classification algorithm based on ECA-NET and multi-granularity

Yabo Shang[1], Hua Huo[1,*]

[1]*College of Information Engineering, Henan University of Science and Technology, Kaiyuan Avenue 263, Luoyang, 471023, China*
[*]*Corresponding author*

Abstract: *The feature of large intra-class variance in fine-grained image classification is a challenge to the classification task. How to effectively learn the discriminant objects in the graph and find out the small discriminant regions is the key to classification. This paper proposes a weak-supervised fine-grained image classification algorithm based on multi-granularity feature fusion. The ECA module is fused with the classic network ResNet-50 to optimize the residual block to obtain a new basic network to enhance channel attention. Secondly, the local chaos module is introduced into the network to form a new image through random chaos regrouping so that the network can learn local regions with different scales of discrimination and obtain fine-grained feature expressions. The cooperative training of dual network branches makes the overall information and local information complement each other and have better expression. Experimental results on three widely used fine-grained image classification datasets verify that the proposed algorithm improves the accuracy of classification tasks and can effectively identify semantic sensitive features in images.*

Keywords: *fine-grained image classification; dual branch network; linear feature fusion; multi-granularity*

1. Introduction

Based on traditional image classification, the goal of fine-grained image classification is to classify objects belonging to the same category into corresponding subcategories, so fine-grained image classification is also called subcategory image classification [1-3]. Compared with general image classification, there are subtle visual differences between categories with similar appearance features. Fine-grained image classification has the following characteristics: (1) Multiple visually similar subcategories, such as objects of different categories, may have very similar contour shapes, colors, textures, etc. (2) Large intra-class differences, such as different occlusion, background, lighting and many other uncertain factors due to the action form and perspective of the identified object. These characteristics make the appearance difference between images huge and the details between different objects difficult to capture.

To solve this problem, this paper proposes a network model based on ResEca as the benchmark network and integrates multi-granularity features. The image is scrambled and recombined through the local chaos module, and different branch networks learn features of different granularity. The semantic information is enriched by fusing multi-granularity features to improve classification accuracy.

The main contributions of this paper are as follows:

1) A fine-grained image classification method based on dual network branches is proposed. Through the cooperative training and sharing of parameters of the main network and multi-granularity network branches, both the global feature information and the local complementary information between different granularities can be learned

2) The input image is randomly and uniformly divided and recombined through the local chaos module to form a multi-granularity level so that the network can better focus on the discriminant local areas between different granularities.

3) The characteristics of different channels are linearly fused, and information fusion across multiple granularities enhances the linear correlation between channels so that the network performance is better.

ISSN 2706-655X Vol.5, Issue 2: 31-38, DOI: 10.25236/IJFET.2023.050206

2. Related Work

The deep learning algorithm is constantly updated and improved, especially the deep convolution neural network algorithm [4–8, 29], which has brought huge development space for fine-grained image classification. The research of fine-grained image classification methods can be divided into two directions. One is the classification method based on strong supervision information [9]. Due to the characteristics of fine-grained image classification, it is necessary to locate the detailed areas that play a role in distinguishing the image. For these classification objects, the strong supervision method uses additional manual annotation to enhance the information and marks the key parts and regions. Although the accuracy of classification is improved, the labor cost is expensive, the area labeled manually is not necessarily the most suitable area, and the actual utilization rate is not high. The second is the classification method based on weakly supervised information [10,12], which does not rely on manually labeled information, but only relies on image category labels, which greatly reduces the cost of labor, and becomes the mainstream trend in the research of fine-grained image classification methods.

2.1. Strongly-supervised Algorithm

The strong supervised learning method can be seen as the continuation of the artificial design feature method used before the deep learning classification method. It uses the artificial design label box to make the convolution neural network automatically acquire the image features. When training the model, it uses the object label box to intelligently improve the artificial design feature method and overcomes its shortcomings. [11] Use the valve connection function of the aligned subnetwork to improve the positioning effect and the connection effect between the classified subnetworks, and make the network reach a stable state. However, in the process of strict implementation of the system, the cost is high and the cost performance is low.

2.2. Weakly-supervised Algorithm

The strong supervision algorithm relies heavily on image label information and has high requirements for data sets, which makes the weak supervision algorithm relying only on image label information prevail.[13] a multiscale deep reinforcement learning method (M2DRL) is proposed. The two-stage deep reinforcement learning method adaptively determines the discrimination region and localizes the regions at different scales. [14] Using end-to-end feature coding, the region with a specific kind of discrimination block is determined by learning the convolution filter library so that the feature learning is strengthened in the network structure. [15] By constructing a three-line attention model, the feature is transformed into an attention map, and the overall and local features are obtained according to the methods of sampling and knowledge distillation. However, due to the multiple training involved, the training cost was increased. [16] A progressive multi-granularity framework is constructed to find the difference of multi-granularity image information in a phased way and encourage the network to learn different granularity.

3. Methods

Fine-grained image classification has the characteristics of large intra-class changes and small inter-class changes, which makes the classification task pay more attention to local details. To solve this problem, this paper proposes a weak supervised fine-grained image classification method that integrates different granularity features, as shown in Figure 1. ResEca is used as the basic network to extract the feature map of the image. The network is divided into two branches: the main network and the multi-granularity network. The main network focuses on the overall information of the image features; the Multi-granular network obtains discriminative regions of different sizes through the local chaos module and pays attention to local region information of different sizes with recognition. The two networks jointly train and share parameters, learn complementary information between different granularities, and enhance expression ability.

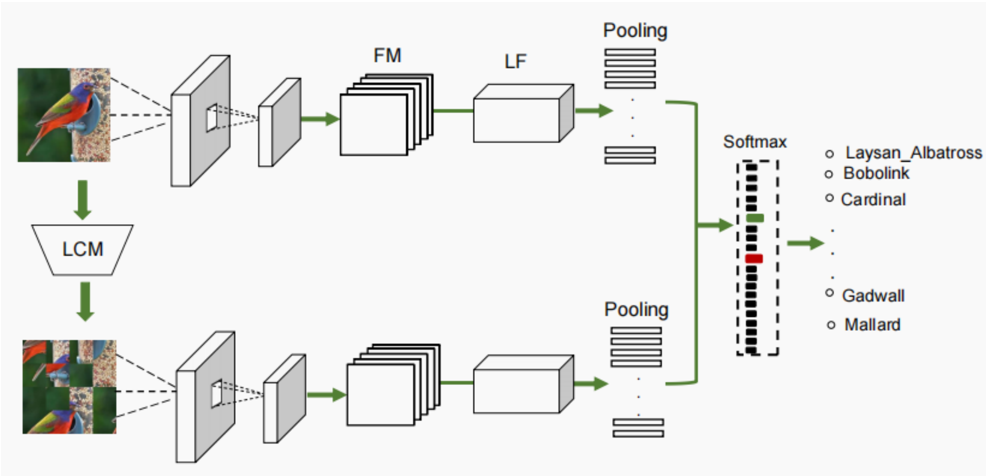

Figure 1: Structure diagram of the dual-branch network framework, FM represents the feature map, LF represents linear fusion and LCM represents the local chaos module

3.1. Main Network

ResEca is selected as the basic convolutional neural network for the main network branch. ResEca network is based on ResNet-50, adding the ECA Net channel attention module to form ResEca basic network. After the ECA module is added, it is easier to extract the resolution features of the image on the basis of the channel dimension, making the overall model easier to train. The ResEca network structure is shown in Figure 2.

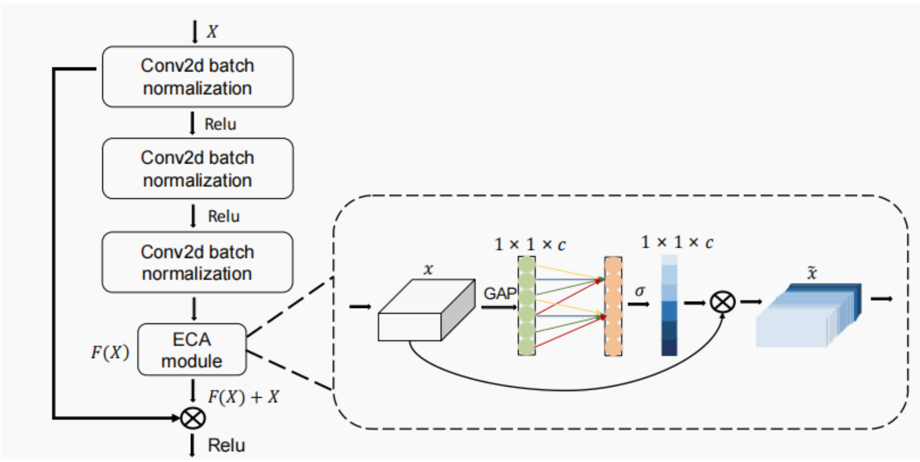

Figure 2: Modification of residual block

The global image features are extracted through the basic network, and the final image representation is obtained through the external product of the output of two feature extractors. Use the X representation input picture, F represent the feature map.

$$F = MainNet(X) \quad (1)$$

m represents the feature matrix obtained from the feature graph, F through the feature extractor, $m \in R^{c \times k}$, Where c represents the number of channels in the feature map, and k represents the number of features contained in the feature map. The feature maps obtained by two feature extractors are aggregated by the outer product to obtain the linear relationship between the feature map channels to obtain the image representation:

$$A = \frac{1}{ab} m_1 \cdot m_2 \quad (2)$$

ISSN 2706-655X Vol.5, Issue 2: 31-38, DOI: 10.25236/IJFET.2023.050206

Where A represents aggregation matrix, $A \in R^{a \times b}$, a and b represent the number of channels of the two characteristic graphs. Secondly, the feature is normalized, and the matrix is expanded into a vector. According to the signed square root regularization and normalization, the feature matrix has a good distribution.

$$I = sign(Vec(A))\sqrt{|Vec(A)|} \tag{3}$$

$$i_{main} = I/\|I\|$$

3.2. Multi-granularity Network

Different fine-grained categories may have similar global structures. How to find discriminative local details becomes the key to fine-grained image classification. Inspired by the DCL method [17], this paper proposes the Locally chaotic module (LCM), which divides the image into multiple local areas for random chaotic reorganization to obtain local area images without relevance, as shown in Figure 3.

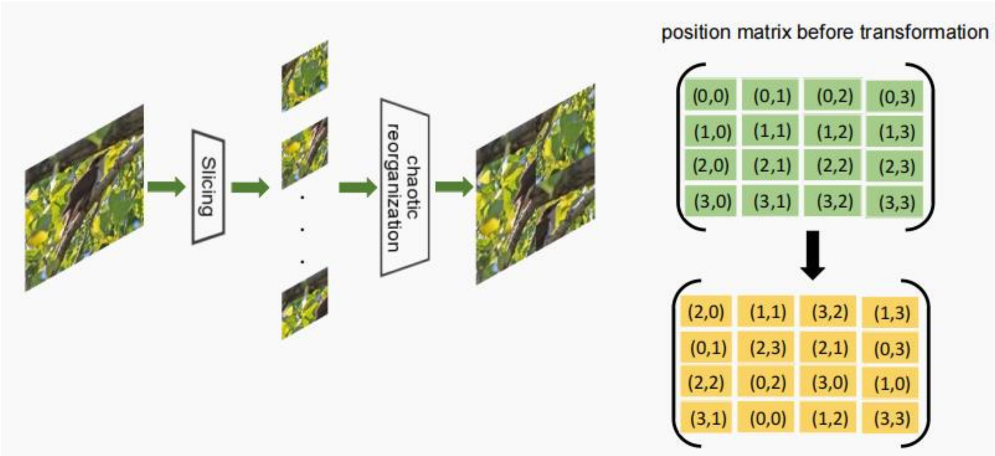

Figure 3: Schematic diagram of partial chaos module

Normalize the size of the input image X (the three-dimensional vector of the image) into a square, and set the division granularity to N that is, divide the image into $N \times N$ sub- regions. Each area of the picture is marked with $R(k)$, and k represents the one-dimensional sequence number of the molecular area, $1 \leq k \leq N^2$.

Randomize the divided sub-region to generate a random vector V with length N^2, and the value of the mth element is f, $1 \leq m \leq N^2$, $f \sim U\left(1, N^2\right)$ uniform distribution. The expression of random vector a is:

$$v[m] = f \tag{4}$$

The random vector V is used to rearrange the chaotic regions, and the reconstructed image $\overline{X}$ is obtained by stitching and combining them according to the new sub-region order.

$$\overline{x} = R\left(v\left[f\right]\right) \tag{5}$$

ISSN 2706-655X Vol.5, Issue 2: 31-38, DOI: 10.25236/IJFET.2023.050206

3.3. Loss Function

In this paper, the cross entropy function is used to calculate the loss between the real label and the predicted probability distribution, and the backpropagation update coefficient is used. The loss function of the main network is combined with the loss function of the multi-granularity network to form the total loss function, and the network parameters are optimized through collaborative training.

$$Loss_{total}(x,i) = Loss_{main}(x,i_{main}) + Loss_{multi}\left(x,i_{mutil}\right)$$

4. Experimental Analysis

This paper selects three classic fine-grained image datasets CUB-200-2011[18], FGVC-Aircraft[19], and Stanford Cars[20]. During the experiment, the image category label of the dataset is checked without additional label information such as boundary box and position key. The usage of datasets is shown in Table 1.

Table 1: Fine-grained image classification dataset

Datasets	Category	Training	Testing
CUB-200-2011	200	5994	5794
FGVC-Aircraft	100	6667	3333
Stanford Cars	196	8144	8041

4.1. Experimental Design

During the experiment, the input image size is cut to 448 × 448 to standardize the image, and the random rotation and horizontal rotation images are used for data enhancement to improve the network training effect and avoid over-fitting. The random gradient descent method is selected as the model optimizer. The momentum parameter is set to 0.9, the number of batch samples is set to 12, the learning rate is set to 0.01, the weight attenuation is set to 0.00001, and the maximum number of training iterations is set to 150.

4.2. Ablation Experiment

Ablation experiments were carried out on three datasets to verify the contribution of each module and its combination to the accuracy of model classification. It can be seen from Table 2 that each module can effectively improve the accuracy of model classification. The classification algorithm consists of three parts: basic network, multi-granularity network, and linear fusion. It can be seen that the accuracy of network classification is greatly improved after adding the local chaos module, and the linear fusion-aided model training achieves the highest accuracy of the model.

Table 2: The contribution rate of each module combination.(%)

Module	Accuracy
ResEca	88.05
ResEca+LF	88.41
ResEca+LCM	88.72
ResEca+LF+LCM	89.63

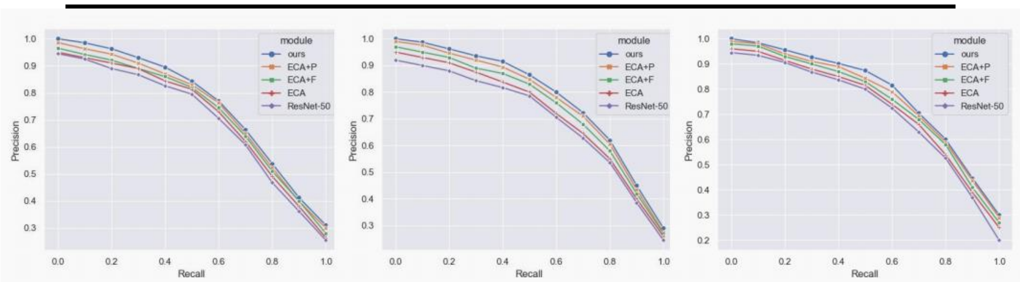

Figure 4: PR curves of different modules on three datasets

As shown in Figure 4 of the precision-recall curve, we can intuitively feel the effectiveness of each module, showing the superior performance of this method. The experimentalresults show that the PR curve after adding the ECA module is significantly better than the baseline residual network, indicating that the channel attention mechanism is helpful to improve the performance of the model; At the same time, adding linear feature fusion and local chaos module has improved the performance of the model to varying degrees. Several modules work together better than a single module.

4.3. Contrast Experiment

In the experiment shown in Table 3, our method compares the results of other advanced algorithms in the same period on three classic datasets CUB-200-2011, FGVC-Aircraft, and Stanford Cars. The PMG [34] algorithm uses smaller-scale information than the method in this paper, and the maximum granularity is 8. Therefore, small-scale objects have stronger robustness, but the information expression ability is insufficient; The algorithm in thispaper fuses features linearly and complements each other. The accuracy rate is 0.33%, 1.92%, and 0.84% higher than that of its method on three datasets. The method in this paper has shown excellent performance on three data sets, reaching the accuracy rate respectively, which proves to be effective and feasible, and improves the classification effect.

Table 3: This is a table caption. Tables should be placed in the main text near to the first time they are cited. (%)

Method	Network	CUB	FGVC	Stanford Cars
DCL[17]	ResNet-50	87.80	93.00	94.50
Attentive[21]	ResNet-101	87.10	92.80	94.20
Bi-Modal[22]	ResNet-50	88.70	90.80	93.10
LIO[23]	ResNet-50	88.00	92.70	94.50
SnapMix[24]	ResNet-50	88.70	93.24	95.00
SEF[25]	ResNet-50	87.30	92.10	94.00
AP-CNN[26]	ResNet-50	88.40	94.10	95.40
SSSNET[27]	ResNet-50	89.00	93.30	95.00
GHNS[28]	ResNet-50	89.06	94.40	95.68
MC Loss[30]	ResNet-50	87.30	92.90	93.70
AE-AN[31]	ResNet-50	87.80	91.30	93.80
MPCF[32]	ResNet-50	89.10	93.60	95.00
FBSD[33]	ResNet-50	88.90	92.70	94.40
PMG[34]	ResNet-50	89.60	93.40	95.00
OURS	ResNet-50	89.25	94.25	95.38
OURS	ResEca	89.63	94.72	95.74

4.4. Visualization

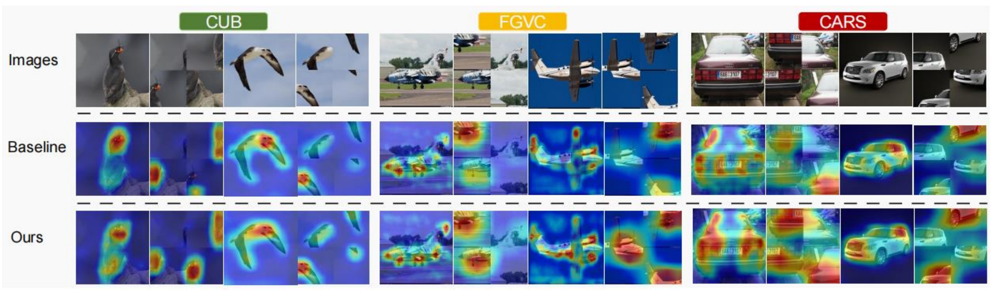

Figure 5: Comparison with baseline method attention map

To better verify the effectiveness of the method, use Grad-CAM to visually compare the baseline and the proposed attention area. As shown in Figure 5. Through comparison, it can be seen that the network model pays more attention to the feature area and focuses on the more discriminative area and more details for fine-grained image classification.

5. Conclusions

Aiming at the problem of fine-grained image classification, a weakly supervised fine-grained image classification method based on multi-granularity feature fusion is proposed. Aiming at distinguishing

ISSN 2706-655X Vol.5, Issue 2: 31-38, DOI: 10.25236/IJFET.2023.050206

discriminant regions of fine-grained map image categories, this algorithm designs a network model trained by three components: basic network, local chaos module, and linear fusion. The effectiveness of the proposed method was verified by comparison of experimental results, and the accuracy of 89.63%, 94.72%, and 95.74% was achieved on the three datasets, respectively. Compared with other algorithms, it has better results, which improves the accuracy of fine-grained image classification. There are two main directions for future work: (1) the general method will deepen the network model to improve the classification accuracy, and the next step is to focus on compressing the network model based on high accuracy (2) for the chunked images generated by the local chaos module, it is possible to produce redundant data, and the next step is to focus on how to improve the utilization rate of such images.

References

[1] Zhang C, Chao L, Liang L, et al. Fine-Grained Image Classification via Low-Rank Sparse Coding With General and Class-Specific Codebooks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(7):1-10.

[2] Wu L, Wang Y, Li X, et al. Deep Attention-Based Spatially Recursive Networks for Fine-Grained Visual Recognition [J]. IEEE Transactions on Cybernetics, 2018:1791-1802.

[3] Dubey A, Gupta O, Raskar R, et al. Maximum-Entropy Fine-Grained Classification[C]// 2018.

[4] Krizhevsky A, Sutskever I, Hinton G. ImageNet Classification with Deep Convolutional Neural Networks [J]. Advances in neural information processing systems, 2012, 25(2).

[5] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition [J]. Computer Science, 2014.

[6] Szegedy C, Wei L, Jia Y, et al. Going deeper with convolutions[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015.

[7] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016.

[8] Howard A G, Zhu M, Chen B, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications [J]. 2017.

[9] Wang Y, Wang Z. A survey of recent work on fine-grained image classification techniques [J]. Journal of Visual Communication and Image Representation, 2019.

[10] Wang D, Shen Z, Shao J, et al. Multiple Granularity Descriptors for Fine-Grained Categorization. 2015.

[11] Di L, Shen X, Lu C, et al. Deep LAC: Deep localization, alignment and classification for fine-grained recognition[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2015.

[12] Ge W, Lin X, Yu Y. Weakly Supervised Complementary Parts Models for Fine-Grained Image Classification from the Bottom Up[C]// IEEE Conference on Computer Vision & Pattern Recognition. arXiv, 2019.

[13] He X, Peng Y, Zhao J. Which and How Many Regions to Gaze: Focus Discriminative Regions for Fine-Grained Visual Categorization [J]. International Journal of Computer Vision, 2019.

[14] Wang Y, Morariu V I, Davis L S. Learning a Discriminative Filter Bank within a CNN for Fine-grained Recognition [J]. 2016.

[15] Zheng H, Fu J, Zha Z J, et al. Looking for the Devil in the Details: Learning Trilinear Attention Sampling Network for Fine-grained Image Recognition[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019.

[16] Du R, Chang D, Bhunia A K, et al. Fine-Grained Visual Classification via Progressive Multi-Granularity Training of Jigsaw Patches[C]// 2020.

[17] Chen Y, Bai Y, Zhang W, et al. Destruction and construction learning for fine-grained image recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 5157-5166.

[18] Wah C, Branson S, Welinder P, et al. The Caltech-UCSD Birds-200-2011 Dataset [J]. california institute of technology, 2011.

[19] Maji S, Rahtu E, Kannala J, et al. Fine-Grained Visual Classification of Aircraft [J]. HAL - INRIA, 2013.

[20] Krause J, Stark M, Deng J, et al. 3D Object Representations for Fine-Grained Categorization [C]// IEEE International Conference on Computer Vision Workshops. IEEE, 2014.

[21] Guo C, Lin Y, Xu M, et al. Inverse transformation sampling-based attentive cutout for fine-grained visual recognition[J]. The Visual Computer, 2022:1-12.

[22] Song K, Wei X S, Shu X, et al. Bi-Modal Progressive Mask Attention for Fine-Grained Recognition[J]. IEEE Transactions on Image Processing, 2020, PP (99):1-1.

ISSN 2706-655X Vol.5, Issue 2: 31-38, DOI: 10.25236/IJFET.2023.050206

[23] Zhou M, Bai Y, Zhang W, et al. Look-Into-Object: Self-Supervised Structure Modeling for Object Recognition[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[24] Huang S, Wang X, Tao D. SnapMix: Semantically Proportional Mixing for Augmenting Fine-grained Data[C]// National Conference on Artificial Intelligence. 2021.

[25] Luo W, Zhang H, Li J, et al. Learning Semantically Enhanced Feature for Fine-Grained Image Classification [J]. IEEE Signal Processing Letters, 2020, PP (99):1-1.

[26] Ding Y, Ma Z, Wen S, et al. AP-CNN: Weakly Supervised Attention Pyramid Convolutional Neural Network for Fine-Grained Visual Classification [J]. IEEE Transactions on Image Processing, 2021, PP (99)

[27] Rong S, Wang Z, Wang J. Separated Smooth Sampling for Fine-grained Image Classification [J]. Neurocomputing, 2021, 461(5).

[28] Tk A, Kh A, Hb A. The Feature Generator of Hard Negative Samples for Fine-Grained Image Recognition [J]. Neurocomputing, 2020.

[29] Zhang Y, Sun Y, Wang N, et al. MSEC: Multi-Scale Erasure and Confusion for fine-grained image classification [J]. Neurocomputing, 2021, 449: 1-14.

[30] Chang D, Ding Y, Xie J, et al. The Devil is in the Channels: Mutual-Channel Loss for Fine-Grained Image Classification [J]. IEEE Transactions on Image Processing, 2020, PP (99):1-1.

[31] Ji J, Jiang L, Zhang T, et al. Adversarial erasing attention for fine-grained image classification[J]. Multimedia tools and applications, 2021(80-15).

[32] Lei J, Yang X, Yang S. Multiscale Progressive Complementary Fusion Network for Fine-Grained Visual Classification [J]. IEEE Access, 2022, 10: 62800-62810.

[33] Song J, Yang R. Feature Boosting, Suppression, and Diversification for Fine-Grained Visual Classification[C]// 2021.

[34] Du R, Chang D, Bhunia A K, et al. Fine-Grained Visual Classification via Progressive Multi-Granularity Training of Jigsaw Patches[J]. 2020.