ISSN 2616-5775 Vol. 6, Issue 2: 109-119, DOI: 10.25236/AJCIS.2023.060215

# MSFF: Multi-Scale feature fusion for fine-grained image classification

**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:*** *Fine-grained image classification is a sub-category classification problem with a common superior category. Aiming at the characteristics of large intra-class differences and small inter-class differences in fine-grained images, this paper proposes a fine-grained image classification method based on multi-scale feature fusion. The method constructs a three-branch network model. The attention module and local extraction module are used to obtain the image of the target object and the image of the parts with strong distinguishing detail features. The depth metric learning is used to shorten the distance from the same data by using misclassification information to improve the classification accuracy; secondly, without using the image bounding box/partial annotation information, the image information of different scales is fused through a parallel network structure; finally, the entire network is optimized by combining the loss functions of the three-branch networks. This method performs end-to-end training collaboratively in a multi-branch network to enhance the ability to express information, thereby improving the accuracy of image classification. To evaluate the effectiveness of our method, fine-grained classification experiments were conducted on three datasets. The experimental results show that the algorithm has higher classification accuracy than other fine-grained classification algorithms.*

***Keywords:*** *Fine-grained visual classification, Multi-scale feature fusion, Multi-branch network, Deep metric learning*

## 1. Introduction

With the rapid development of artificial intelligence, image classification technology is also constantly improving, and traditional image classification is composed of semantic-level and instance-level images. In today's era, intelligent requirements are constantly put forward, and the shortcomings of traditional image classification have gradually revealed that they cannot meet people's needs. Therefore, fine-grained image classification [1-3]has become a scalding research direction in the fields of computer vision, pattern recognition, image processing, and so on[4-6]. The current research ideas are mainly divided into two types: one is to directly learn better visual representations from the original images. The second is to use an attention-based approach to obtain discriminative domains in images. Although these methods can improve the accuracy to a certain extent, the extraction of key area features includes irrelevant background or object areas, so that the recognized objects cannot correspond and affect the classification effect.

To sum up, based on the recurrent attention convolutional neural network[7], this paper proposes a network framework on the basis of multi-scale feature fusion, constructs an attention module and a local extraction module to extract high-quality semantic information, and combines the global image of the multi-branch network performs feature information fusion to improve the accuracy of image classification. The specific implementation is achieved through the following branches: (1) After inputting the original images in the global branch network, the feature maps of the global images is obtained, the attention module obtains the target images through the action of the spatial transformation layer, which is used as the input of the target branch network to make it learn object information; (2)According to the feature maps of the object images, a local extraction module is introduced to extract the component area with sufficient information as the input of the detail branch network to obtain the area with high recognition degree; (3)The multi-scale information obtained from different branch networks is fused by multi-branch networks, and the feature information on objects of different scales and different parts is fully utilized to improve the recognition ability of classification targets; (4)

In the process of feature embedding, deep metric learning is used to optimize the model, which improves the accuracy of localization of discriminative regions and further improves the classification accuracy.

The main contributions of this paper are as follows:

1) This study proposes a fine-grained image classification network that fuses multi-scale features. Image features of different scales are fused through a multi-branch network to provide the model with different levels of image information.

2) By building an attention module, this study uses affine transformation to learn the target regions from the original images, obtains target feature information, and only uses category labels to achieve the accuracy of target localization. Through the local extraction module, the component area of the object is obtained, so that the model can effectively learn fine-grained features of different scales.

3) This study uses deep metric learning N-pair Loss, which can optimize the model, reduce classification errors and obtain good prediction results.

4) This paper conducts comparison and ablation experiments on three classic fine-grained image classification datasets and achieves good results, with better model performance.

## 2. Related Work

The opportunity to promote the rapid development of fine-grained image classification is that deep learning image classification algorithms have better image representation capabilities. The current research directions of fine-grained image classification are divided into three types: strongly-supervised learning[8-10], weakly-supervised learning[11-13], and unsupervised-learning[14]. Strongly-supervised learning faces the problem of labeling with high-quality datasets. When it comes to multi-label classification tasks, the labeling cost will increase to the number of targets. The research on unsupervised image classification on the basis of deep learning is still in the development stage and is a very challenging research topic. Therefore, weakly-supervised learning has become a research hotspot of fine-grained image classification methods.

### *2.1 Weakly-supervised method*

Weakly-supervised learning can better achieve fine-grained image classification using only image category information. [15] designed a two-level attention algorithm, all the characteristics of the output were calculated from the model, the training speed of the model was slow and the calculation was large. [16] proposed Bilinear convolution model. Compared with a linear model, the Bilinear model can obtain higher quality feature representation, but the Bilinear feature had the disadvantage of high dimension, which was not conducive to analysis. [17] designed a low-rank bilinear pooling model, which used a bilinear classifier to capture second-order statistics and captured the relationship between local features between layers. [18] proposed a visual attention network that maps attention features to input space and guides attention mapping to pay better attention to fine-grained features.

### *2.2 Multi-scale Feature Fusion*

Multi-scale images can enable the network to learn rich semantic features and texture information that are differentiated between different levels. In fine-grained image classification methods, the use of multiple scales can take into account both objects and detailed regions in the image. [19]proposed a recurrent attention convolutional neural network framework, which was divided into three sub-networks with the same network structure framework. [20] proposed a feature pyramid structure to identify the basic components in the system by detecting objects at different scales. [21]created the SK-Net network framework. The module was divided into three parts. Although the accuracy is improved, the addition of modules increased the complexity of the model.

## 3. Approach

In this section, we describe the proposed network architecture in detail, as shown in Figure 1. The key to fine-grained image classification is to locate precise discriminative regions and learn the saliency of fine features for image classification. Inspired by this, we divide the image into three different scales, which contain global information, target object information, and local detailed component information, respectively.

Our goal is to fuse features from informative regions and complete images to allow the network to learn more feature information from images of different scales, enabling information complementarity to achieve better performance. This paper adopts a multi-branch network structure, which can be divided into the global branch, target branch, and detail branch. In the training process, all branch networks are used to train together, and the global branch and the target branches are used to obtain the classification results in the testing phase.

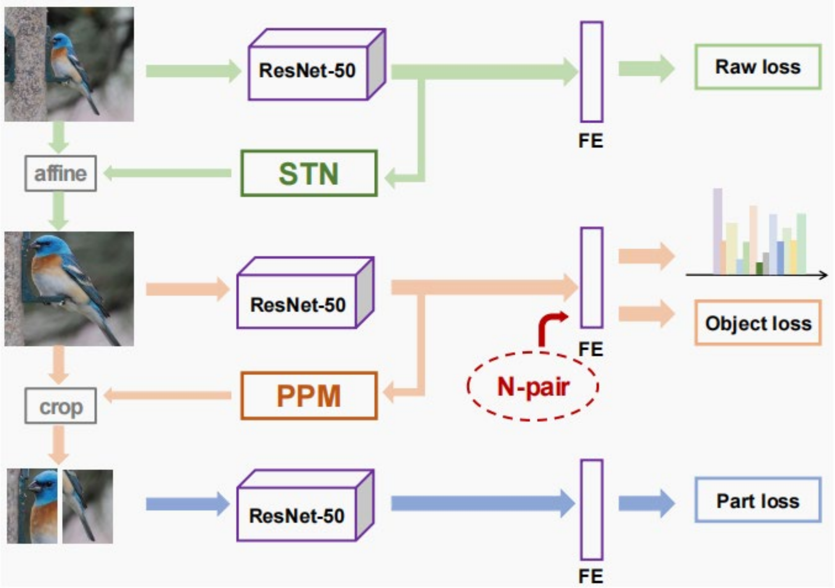

*Figure 1: Overview of the Multi-scale features fusion network's architecture. Global branches are shown in green, target branches are shown in orange, and detail branches are shown in blue. The benchmark network uses ResNet-50 as the feature extraction network and shares parameters with FE (feature embedding). Affine stands for affine transformation*

### 3.1 Attention Module

Fine-grained images are easily affected by the pose of objects, etc. Spatial Transformer Networks [22] is introduced for this problem, which is an attention mechanism network that pays attention to space. A new learnable spatial transformation module is introduced, it can adaptively mine the discriminative area in the image, and can operate in the space of the data in the network so that the model has spatial invariance.

The attention module is composed of three components: a localization network, parameterized sampling grid, and a differentiable sampling mechanism. The working principle is shown in Figure 2. The spatial transformer is introduced through the benchmark network to obtain the attention part of the image from space, and the spatial transformation layer is used to crop and locate the target area to generate the target-level image.

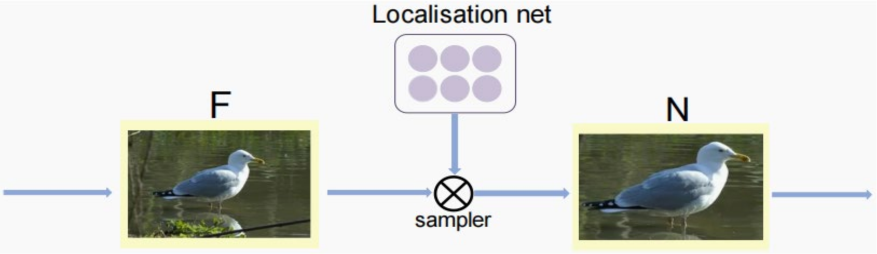

*Figure 2: Schematic diagram of spatial transformation network*

The feature map obtained by inputting the image $X$ through the benchmark network on the last convolutional layer is represented by $F \in \mathbb{R}^{C \times H \times W}$. The output feature graph generated in the $l$ stage is expressed as $F_l$. The function of the localized network is to pass the input feature map $F_l$ through a sub-network to generate the parameters of the spatial transformation and output the transformation matrix of the affine transformation $\phi$.

$$\phi = f_n\left(F_l\right) \tag{1}$$

$f_n\left(\cdot\right)$ represents a local network function, parameters $\phi$ to be applied to the transformation of the feature map. Second, calculate the position of the sampling point in the original image according to the formula, convert the feature map through the grid generator, and perform coordinate mapping to generate the feature image $N \in \mathbb{R}^{C' \times H' \times W'}$.

$$\begin{pmatrix} x_l^o \\ y_l^o \end{pmatrix} = T_\phi\left(G_l\right) = A_\phi \begin{pmatrix} x_l^p \\ y_l^p \\ 1 \end{pmatrix} \tag{2}$$

The original image coordinates are expressed as $\left(x_l^o, y_l^o\right)$, the pixel position of the target image is expressed as $G_l = \left\{x_l^p, y_l^p\right\}$, means affine relationship. Finally, through the sampler, the bilinear interpolation method is used to collect pixels according to the input feature map and the corresponding affine transformation relationship to generate the final output feature map, that is, the target-level image. $D_{mn}^c$ represents the value at coordinate $\left(m,n\right)$ in the color channel $c$.

$$N_l^c = \sum_m^H \sum_n^W D_{mn}^c \max\left(0,1-\left|x_l^o - n\right|\right)\max\left(0,1-\left|y_l^o - m\right|\right) \tag{3}$$

### *3.2 Part Proposal Module*

The research focus of numerous fine-grained image classification tasks is to solve the problem of large intra-class differences in fine-grained images, so it is essential to localize local regions with rich information. To extract the local areas with rich effective information, this paper introduces the Part Proposal Module (PPM) processes as shown in Figure 3, to enhance the discrimination ability of the model.

Based on the idea of a sliding window in target detection, find the part area with a resolution in the target image.

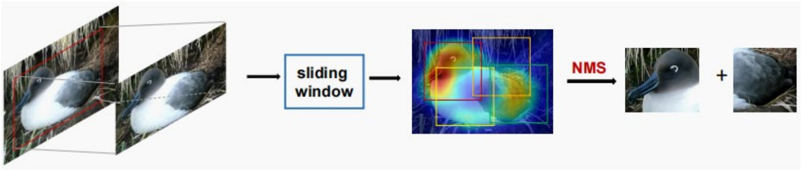

*Figure 3: Local extraction module process diagram*

First, the feature maps under different sliding windows of the previous branch in the network are aggregated to obtain the corresponding feature maps. Under the channel dimension of the feature map, the activation feature map $P_w$ under each sliding window is obtained by aggregation. The square means is used to represent the activation value, which is better than the arithmetic means for semantically rich regions. The activation value of each window can be expressed as the following:

$$\overline{p}_w = \sqrt{\frac{\sum_{x=0}^{W_w-1}\sum_{y=0}^{H_w-1}\left[P_w(x,y)\right]^2}{H_w \times W_w}} \tag{4}$$

According to the activation value of each window, regions with rich semantic information have high activation values, which represent key regions, which are conducive to capturing discriminative features in fine-grained images; regions lacking semantic information, with small activation values, represent irrelevant regions. The window of the feature map is sorted by the size of the score. Due the large number of candidate regions, many redundant regions will be doped. For this problem, the non-maximum suppression method is used to reduce the number of candidate regions and reduce the computational cost. According to whether the activation value of each window is greater than the threshold value, the clipped mask image is generated. The detail image is obtained by clipping the target-level image, and the detail image is input into the detail branch network to optimize the local area information.

$$\widetilde{M}_W\left(x,y\right)=\begin{cases}1, P_w\left(x,y\right)>\overline{p_w}\\ 0, otherwise\end{cases} \tag{5}$$

$P_w\left(x,y\right)$ represents the activation value of the image obtained through the network, $W_w$ indicates the width of the window graph, $H_w$ indicates the height of the window graph.

### *3.3 Deep Metric Learning*

The embedding representation of fine-grained datasets is significantly diverse, the intra-class distance can be greater than the inter-class distance, the differences between classes are more subtle, and fine-grained images are susceptible to appearance and other reasons.

N-pair Loss can solve this problem. The N-pair Loss [23] function belongs to one of the deep metric learning methods. Its working principle is to use deep neural networks to learn highly abstract nonlinear features and the similar relationship between data, as shown in Figure 4.

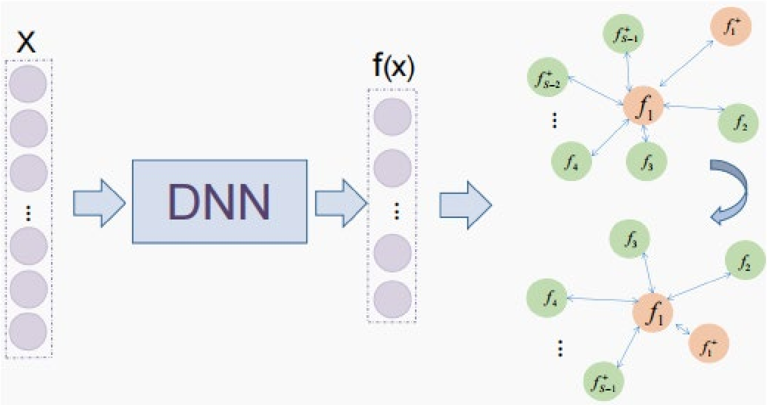

*Figure 4: N-Pair Loss schematic*

Denote the input image by $X$, $f(X)$ stands for deep feature embedding. The goal of deep embedding learning is to learn the deep feature embedding of the input image as a feature vector so that the similarity of the vectors can achieve higher scores when they belong to the same class. High scores belong to the same category, and low scores belong to different categories. Positive dots indicate that the samples are from the same class, and negative dots indicate different classes.

$$L_{N-pair}=\frac{1}{S}\sum_{a=1}^{S}\log\left(1+\sum_{a\neq b}\exp\left(f_a{}^{\top}f_b^{+}-f_a{}^{\top}f_a{}^{+}\right)\right) \tag{6}$$

### *3.4 Loss Function*

The network framework proposes in this paper needs to deal with coarse grained features and fine-grained features. The three-branch networks can obtain target-level and component-level images from original images through different modules and uses multiple scales to learn feature information to improve classification accuracy. The classification loss function of the three-branch network uses the cross-entropy loss function to measure the fitting ability of the neural network model to fine-grained image data. The loss function under each branch is recorded as the following:

$$\begin{aligned} L_g &= H\left(p_g, q_g\right) = -\sum_{i=1}^{n} p_g\left(x_i\right)\log\left(q_g\left(x_i\right)\right) \\ L_o &= H\left(p_o, q_o\right) \\ L_d &= \sum_{n=0}^{N-1} H\left(p_{d(l)}, q_{d_{(l)}}\right) \end{aligned} \tag{7}$$

$L_g$, $L_o$, $L_d$ represents the three-branch global network, target network, and detail network loss function. The proposed loss function improves the generalization ability of the model and optimizes the model parameters to fit the network training. The overall loss function of the three-branch network is defined as:

$$L_t = L_g + L_o + L_d + L_{N-pair} \tag{8}$$

## 4. Experiments

### *4.1 Datasets*

This experiment uses three public experimental datasets in the field of fine-grained image classification: the CUB-200-2011 dataset[24] contains a total of 11,788 images of 200 species of birds, each image in the dataset is annotated with a bounding box, part location, and attribute labels. The FGVC-Aircraft dataset[25] contains a total of 100 models, with 10,000 images, the main aircraft in each image is annotated with a bounding box and a hierarchical aircraft model label. The Stanford Cars dataset[26] contains a total of 16,185 images of 196 vehicle types, car images in the dataset are taken from multiple angles and categorized by year of production and model. The specific information of each dataset, the number of categories including the division of the training set and the test set in the experiment is shown in Table 1. Use the default-split training and test sets during the experiment. Only image-level labels were used during the experiments without any additional manual annotations.

*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 |

### *4.2 Implementation Details*

Experimental parameter settings: The benchmarking network of this experimental model uses the residual network Resnet-50 as the image feature extractor. Training parameter settings: the number of batch samples is set to 6, and the learning rate is set to 0.001. The experiment adopts the stochastic gradient descent method and uses the batch normalization method as the regularization term to train the model, and the weight decay is set to 1e-4. During the experiment, the global image and target images are processed to 448*448 size, and the part image is processed to 224*224 size.

### *4.3 Ablation Experiments*

In order to verify whether each module proposed in this paper can effectively improve the network performance, two groups of ablation experiments are performed on three datasets.

### *4.3.1 Ablation Experiments 1*

In ablation experiment 1, the benchmark network ResNet-50 was compared with the network with various modules added, and the influence of the three modules on the overall model classification effect was compared. The experimental results of the model on the test set are shown in Table 2. It can be seen from the table that the network after adding each module improves the classification accuracy of the dataset to varying degrees. The accuracy of CUB data was increased by 1.14, 2.89, 2.64, and 3.07 percentage points, respectively. The classification accuracy of the FGVC dataset increased by 1.52, 2.27, 2.36, and 3.74 percentage points, respectively. In the Stanford Cars dataset, the classification accuracy increased by 0.67, 1.26, 1.11, and 1.81 percentage points, respectively.

*Table 2: Results of ablation experiments on datasets*

| | ResNet-50 | Attention | Part | Accuracy |
|---|---|---|---|---|
| CUB | ✓ | | | 86.00 |
| | ✓ | ✓ | | 87.89 |
| | ✓ | ✓ | ✓ | **89.07** |
| FGVC | ✓ | | | 90.10 |
| | ✓ | ✓ | | 92.37 |
| | ✓ | ✓ | ✓ | **93.84** |
| Stanford Cars | ✓ | | | 93.00 |
| | ✓ | ✓ | ✓ | 94.26 |
| | ✓ | ✓ | | **94.81** |

### *4.4 Contrast Experiment*

In the experiments in this section, in order to verify the superiority of this method, this method is compared with the current mainstream strong and weakly supervised classification methods.

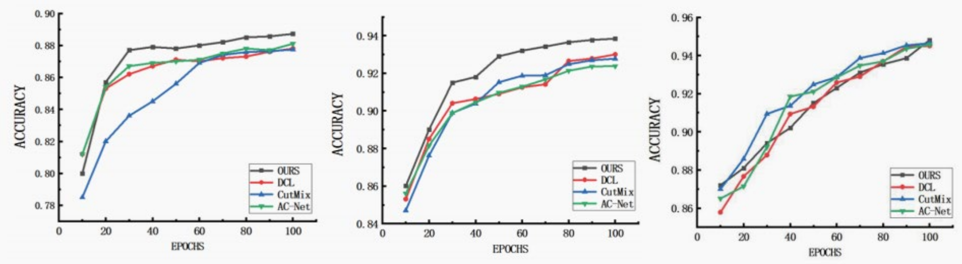

*Figure 5: Testing process of OURS, DCL, CutMix, and AC-Net on the Caltech-UCSD Birds Dataset, FGVC-Aircraft, and Stanford Cars*

On three challenging datasets, our model outperforms the baseline model RA-CNN significantly by 3.67%, 3.94%, and 1.71%, respectively. Table 3 reports the accuracy comparison with current mainstream methods. Compared with other weakly supervised algorithms, this method achieves better classification on all three data sets. The results show that the generalization of the network structure is verified. Figure 5 visualizes the testing process with other algorithms, and the superiority of this method can be seen in the figure.

*Table 3: Comparison of classification performance on three datasets*

| Module | CUB | FGVC | Stanford |
|---|---|---|---|
| DB[27] | 88.60 | 93.50 | 94.70 |
| CIN[28] | 87.50 | 92.60 | 94.10 |
| DFL-CNN[29] | 87.40 | 91.40 | 94.50 |
| NTS-Ne[30] | 87.50 | 93.00 | 94.70 |
| Cross-X[31] | 87.70 | 92.60 | 94.60 |
| HBP[32] | 87.10 | 90.30 | 93.70 |
| TASN[33] | 87.90 | —— | 93.80 |
| MCL[34] | 87.30 | 92.60 | 93.70 |
| MSEC[35] | 88.30 | 93.40 | —— |
| | 88.10 | 92.40 | 94.60 |
| AC-Net[36] | **89.07** | **93.84** | **94.81** |
| OURS | | | |

Figure 6 shows the attention module for target object localization in theform of feature map. We can see from the figure that compared with the baseline network architecture RA-CNN, the method in this paper can better focuson the area of the target object and obtain more useful feature information, ignoring the environmental information.

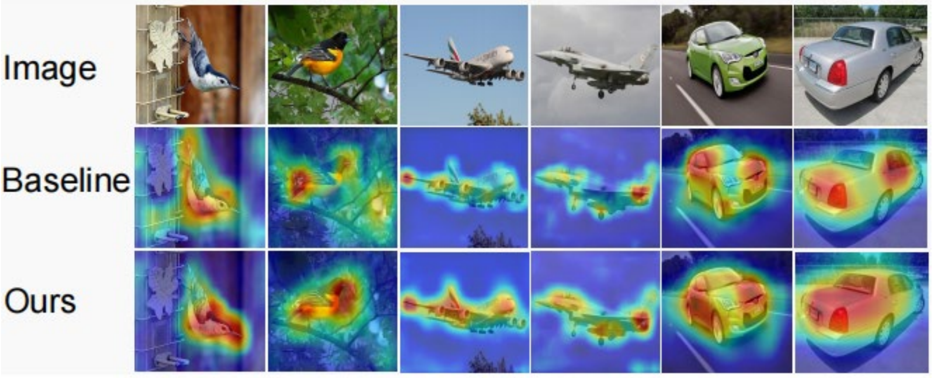

*Figure 6: Attention module comparison*

Figure 7 show the localization results of the local extraction module, using red, orange, yellow, and green to denote important regions in the image. For the anchor design, this study uses two scales of 48 and 96 sizes, and the two scales are 1:1 and 2:3, respectively. It is found from the map that the local extraction module can well locate multiple detailed parts of the target object and find areas with rich information.

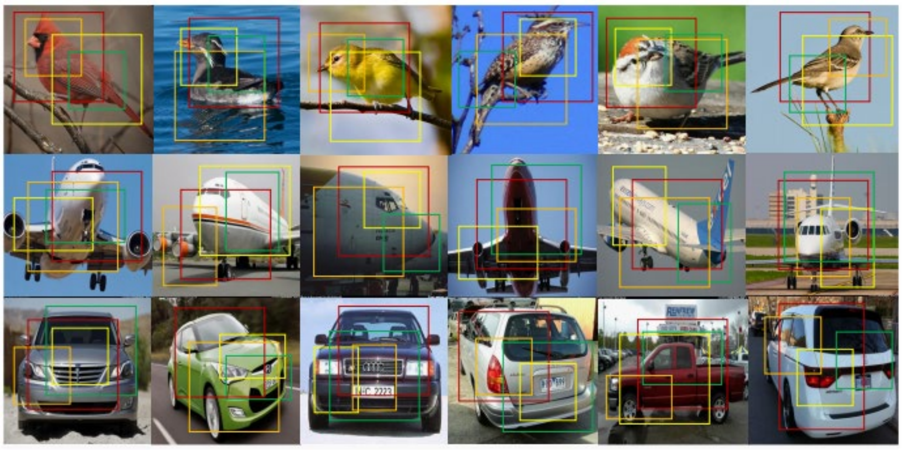

*Figure 7: Visualization of parts area*

Figure 8, 9, 10 visualizes the selection of detail regions, which can well locate multiple different detail regions of the target object, and select regions that obtain more local information.

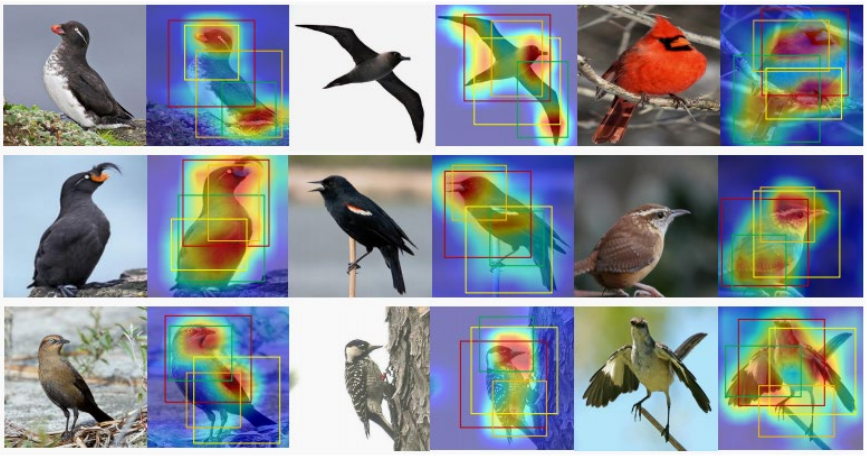

*Figure 8: Compare on CUB-200-211*

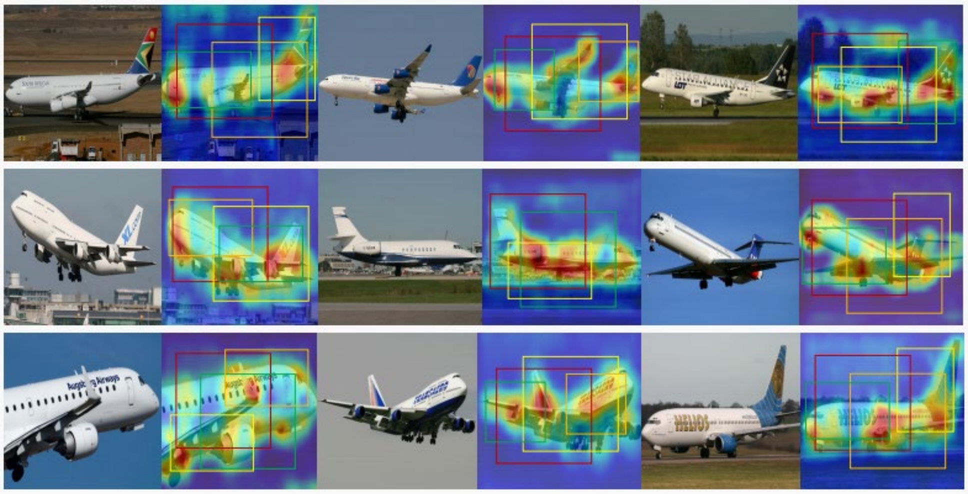

*Figure 9: Compare on FGVC-Aircraft*

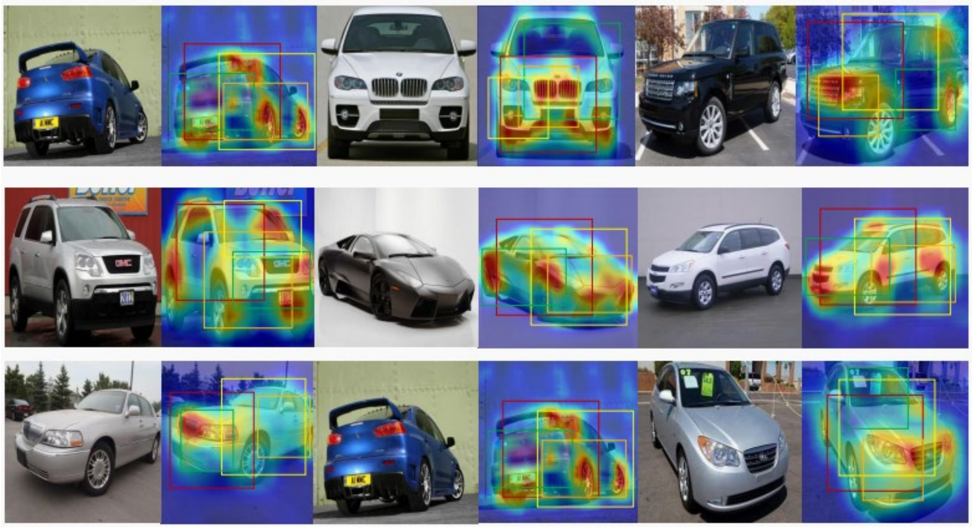

*Figure 10: Compare on Stanford Cars*

## 5. Conclusion

Aiming to how the fine-grained image classification can effectively locatethe target area and extract the discriminative details, this paper proposes a multi-scale feature fusion fine-grained image classification method. The method in this paper can enable end-to-end training to make the attention module, local extraction module, and deep metric learning cooperate, learn the feature information of different scales of the image through different branch networks and perform feature fusion to complement the information to improve the network classification performance. The experimental results show that the classification accuracy of our method on the CUB-200-2011, FGVC-Aircraft, and Stanford Cars datasets reaches 89.07%, 93.84%, and 94.81%. This paper only uses image class labels to achieve better classification results than other methods under weak supervision. Future research will be carried out in two directions. First, how to reduce the redundant information of features based on maintaining network performance, to reduce the amount of model calculation; secondly, to further screen the extracted fine-grained features to deal with subclasses that are easily confused.

## References

*[1] Wei, X, Wang, P, Liu, L, Shen, C, Wu, J. Piecewise classifier mappings: Learning fine-grained learners for novel categories with few examples. IEEE Transactions on Image Processing 2019; 28(12): 6116–6125. 52doi:10.1109/TIP.2019.2924811.*

*[2] Wei, XS, Luo, JH, Wu, J, Zhou, ZH. Selective convolutional descriptor aggregation for fine-grained*

*image retrieval. IEEE transactions on image processing: a publication of the IEEE Signal Processing Society 56 2017; 26(6):2868. doi:10.1109/TIP.2017.2688133.*

*[3] Bo Zhao, Xiao Wu, Jiashi Feng, et al. Diversified visual attention networks for fine-grained object classification. IEEE Transactions on Multimedia 2017;doi:10.1109/TMM.2017.2648498.*

*[4] Deng, A, Wu, Y, Zhang, P, Lu, Z, Li, W, Su, Z. A weakly supervised framework for real-world point cloud classification. Computers Graphics 2022; 102:78–88.*

*[5] Xiang, J, Zhang, N, Pan, R, Gao, W. Efficient fine-texture image retrieval using deep multi-view hashing. Computers & Graphics 2021; 101:93–105.*

*[6] Bibissi, DL, Yang, J, Quan, S, Zhang, Y. Dual spin-image: A bidirectional spin-image variant using multi-scale radii for 3d local shape description. Computers & Graphics 2022; 103: 180–191. doi: 10. 1016/ 69j.cag.2022.02.010.*

*[7] Fu, J, Zheng, H, Tao, M. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: IEEE Conference on Computer Vision Pattern Recognition. 2017, 4438–4446.*

*[8] Ning, Z, Farrell, R, Darrell, T. Pose pooling kernels for sub-category recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2012, p. 3665–3672.*

*[9] Xie, L, Qi, T, Hong, R, Yan, S, Bo, Z. Hierarchical part matching for fine-grained visual categorization. In: IEEE International Conference on Computer Vision. 2014, p. 1641–1648.*

*[10] Zhang, N, Donahue, J, Girshick, R, Darrell, T. Part-based r-cnns for fine-grained category detection. European Conference on Computer Vision 2014; 834–849doi: 10. 1007/ 978- 3- 319- 10590-1_ 54.*

*[11] Girshick, R, Donahue, J, Darrell, T, Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. IEEE Computer Society 2013; 580– 587doi: 10. 1109/ CVPR. 2014. 81.*

*[12] Tao, H, Qi, H. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. arXiv 2019;doi:10.48550/arXiv.1901.09891.*

*[13] Shu, X, Tang, J, Qi, GJ, Li, Z, Jiang, YG, Yan, S. Image classification with tailored fine-grained dictionaries. IEEE Transactions on Circuits and Systems for Video Technology 2018; 28(2):454–467. doi: 10.1109/TCSVT.2016.2607345.*

*[14] Wang, D, Shen, Z, Jie, S, Wei, Z, Zheng, Z. Multiple granularity descriptors for fine-grained categorization. In: 2015 IEEE International Conference on Computer Vision (ICCV). 2015, p. 2399–2406.*

*[15] Xiao, T, Xu, Y, Yang, K, Zhang, J, Peng, Y, Zhang, Z. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. IEEE 2014; 842–850doi: 10.1109/ CVPR.2015.7298685.*

*[16] Lin, TY, Roychowdhury, A, Maji, S. Bilinear cnn models for fine-grained visual recognition. Proceedings of the IEEE international conference on computer vision 2015; 1449–1457doi: 10. 48550/ arXiv. 1504. 07889.*

*[17] Kong, S, Fowlkes, C. Low-rank bilinear pooling for fine-grained classification. IEEE Computer Society 2017; 7025–7034.*

*[18] Eshratifar, AE, Eigen, D, Gormish, M, Pedram, M. Coarse2fine: a two-stage training method for fine-grained visual classification. Machine Vision and Applications 2021; 32(2):1–9. doi:10.1007/ s00138- 021-01180-y.*

*[19] Fu, J, Zheng, H, Tao, M. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: IEEE Conference on Computer Vision Pattern Recognition. 2017, p. 4438–4446.*

*[20] Lin, TY, Dollar, P, Girshick, R, He, K, Hariharan, B, Belongie, S. Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision Pattern Recognition (CVPR). 2017, p. 2117–2125.*

*[21] Wu, W, Zhang, Y, Wang, D, Lei, Y. Sk-net: Deep learning on point cloud via end-to-end discovery of spatial keypoints. Proceedings of the AAAI Conference on Artificial Intelligence 2020; 34(04): 6422– 6429.*

*[22] Jaderberg, M, Simonyan, K, Zisserman, A, et al. Spatial transformer networks. Advances in neural information processing systems 2015; 28.*

*[23] Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. Advances in neural information processing systems 2016; 29.*

*[24] Wah, C, Branson, S, Welinder, P, Perona, P, Belongie, S. The caltech- ucsd birds-200-2011 dataset. california institute of technology 2011;.*

*[25] Maji, S, Rahtu, E, Kannala, J, Blaschko, M, Vedaldi, A. Fine-grained visual classification of aircraft. HAL - INRIA 2013;*

*[26] Krause, J, Stark, M, Deng, J, Li, FF. 3d object representations for fine-grained categorization. In: IEEE International Conference on Computer Vision Workshops. 2014, p. 554–561.*

*[27] Sun G, Cholakkal H, Khan S, et al. Fine-grained recognition: Accounting for subtle differences between similar classes[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(07): 12047-12054.*

*[28] Gao Y, Han X, Wang X, et al. Channel interaction networks for fine-grained image categorization[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(07): 10818-10825.*

*[29] Wang Y, Morariu V I, Davis L S. Learning a Discriminative Filter Bank Within a CNN for Fine-Grained Recognition[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018：4145-4154.*

*[30] Yang Z, Luo T, Dong W, et al. Learning to Navigate for Fine-grained Classification[J]. Springer, Cham, 2018:438-454.*

*[31] Luo W, Yang X, Mo X, et al. Cross-X Learning for Fine-Grained Visual Categorization [J]. 2019: 8241-8250.*

*[32] Yu C, Zhao X, Zheng Q, et al. Hierarchical Bilinear Pooling for Fine-Grained Visual Recognition [J]. 2018:595-610.*

*[33] 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 [J]. IEEE, 2019:5012-5021.*

*[34] 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.*

*[35] 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.*

*[36] Ji R, Wen L, Zhang L, et al. Attention Convolutional Binary Neural Tree for Fine-Grained Visual Categorization[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020:10468-10477.*