Pattern Recognition 146 (2024) 110042

Contents lists available at ScienceDirect

Pattern Recognition

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

Two-stage fine-grained image classification model based on multi-granularity feature fusion

Yang Xu [a], Shanshan Wu [b], Biqi Wang [a], Ming Yang [a], Zebin Wu [a,*], Yazhou Yao [a], Zhihui Wei [a]

[a] School of Computer Science and Engineer, Nanjing University of Science and Technology, Nanjing, 210094, China
[b] Nanjing Research Institute of Electronics Engineering (NRIEE), Nanjing, 210000, China

ARTICLE INFO

Keywords:
Fine-grained
Transformer
Feature-fusion
Attention mechanism

ABSTRACT

Fine-grained visual classification (FGVC) is a difficult task due to the challenges of discriminative feature learning. Most existing methods directly use the final output of the network which always contains the global feature with high-level semantic information. However, the differences between fine-grained images are reflected in subtle local regions which often appear in the front of the network. When the texture of the background and object are similar or the proportion of the background is too large, the prediction will be greatly affected. In order to solve the above problems, this paper proposes multi-granularity feature fusion module (MGFF) and two-stage classification based on Vision-Transformer (ViT). The former comprehensively represents images by fusing features of different granularities, thus avoiding the limitations of single-scale features. The latter leverages the ViT model to separate the object from the background at a very small cost, thereby improving the accuracy of the prediction. We conduct comprehensive experiments and achieves the best performance in two fine-grained tasks on CUB-200-2011 and NA-Birds.

1. Introduction

Fine-grained visual classification aims at distinguishing different subclasses of the same class, e.g., subcategories of birds [1,2], cars [3], pets [4]. It is a very challenging task due to the small inter-class variations and large intra-class variations. With the rapid development of deep learning, many excellent deep convolutional networks (such as Resnet50 [5] and Densenet169 [6]) have been proposed and research on fine-grained image classification has gradually become a hot topic. The main challenge of fine-grained image classification lies in the learning of discriminative features. Early works [7–9] proposed methods based on object parts localization to achieve excellent results. These methods are mainly divided into two steps: (1) locating discriminative parts by using additional annotation or analyzing the response map; (2) extracting the features of these parts separately and concatenating these features for classification. Compared with object parts localization, some methods based on self-attention [10–12] are more efficient. All the features extracted through the backbone network have the same importance. For fine-grained images, the features of discriminative parts are more important. This is where the self-attention mechanism comes into play. This mechanism makes the model pay more attention to the information of this part by applying higher weight to the discriminative features. However, there are some disadvantages in the above methods. They use a single-scale network output, which makes it easy to ignore subtle features that play a key role in prediction. In addition, when the background occupies too large a proportion or the texture is similar to the object, the final prediction will be severely disturbed.

In order to solve the above problems, this paper proposes a two-stage fine-grained image classification model based on multi-granularity feature fusion. The model consists of two parts: multi-granularity feature fusion module and two-stage classification based on Vision-Transformer. The former mainly solves the limitations of single-scale features. This module extracts features of different scales in the backbone network and fuses them to generate features that can comprehensively represent the images. These features contain both high-semantic global information and low-semantic local information. The latter mainly reduces the background interference on predictions. With the help of the ViT model, the coarse positioning of the object can be realized at a very small cost, so that the object can be separated from the background. Through image processing, the object occupies the main component in the new image while the details can be enlarged, which is more conducive to the final prediction. We conduct extensive experiments on four commonly used fine-grained datasets (CUB-200-2011, NA-Birds, Stanford Cars, Pets) and achieved state-of-the-art performances on two datasets. In summary, our main contributions are as follows:

* Corresponding author.
E-mail addresses: xuyangth90@njust.edu.cn (Y. Xu), wuzb@njust.edu.cn (Z. Wu).

https://doi.org/10.1016/j.patcog.2023.110042
Received 7 April 2023; Received in revised form 28 September 2023; Accepted 10 October 2023
Available online 12 October 2023

1. We propose multi-granularity feature fusion module to solve the limitations of single-scale features. The module extracts features of different scales in the backbone network and fuses them to generate features that can comprehensively represent images.

2. We propose two-stage classification based on Vision-Transformer to reduce background interference on predictions. With the help of the ViT model, the object can be separated from the background and the details can be enlarged, which is more conducive to the final prediction.

3. Extensive experiments and the best performance can prove the superiority of our model. The visualization results illustrate that our two-stage classification can accurately localize objects and facilitate correct predictions.

2. Related work

2.1. Localization FGVC methods

The method based on location recognition is to locate key components firstly, and then perform segmentation according to these discriminative features. Inspired by object detection, the discriminative region is regarded as an object, and the object detection model is used to locate it. Wei et al. [9] proposed a semantic segmentation model, mask-based convolutional networks (Mask-CNN), with the help of additional labels, using fully convolutional networks (FCN) [13] to learn a partial segmentation model. Transforming the problem of component positioning into three types of semantic segmentation problems, and using the results of semantic segmentation to extract features of components can also obtain better classification results. Ge et al. [14] proposed a three-stage classification model, weakly supervised complementary parts models (WSCPM): the first stage uses class activation mapping (CAM) [15] and Conditional Random Fields (CRF) [16] to achieve image segmentation and object positioning; the second stage uses an improved object detection method to locate the target parts; the third stage uses a two-way long short-term memory network (LSTM) [17] to achieve feature fusion for classification. Such methods rely on additional supervised information to obtain better classification results, but such information requires a lot of time and efforts from experts. Some methods can also locate widgets using only class labels. Due to the lack of supervision on key parts, such models need to filter region proposals [8]. Fu et al. [18] proposed a Recurrent Attention Convolutional Neural Network (RA-CNN) to recursively identify discriminative regions at multiple scales [19] and reinforce each other. Each scale consists of a classification network and an attention-based proposal network (APN). Networks of different scales are composed of the same network structure in a stacked form. The input first uses the backbone for feature extraction and classification, then APN locates the key regions based on the feature. Then they crop and enlarge the part from the original image for the second subnetwork. This process is repeated several times. RA-CNN can focus on the most discriminative regions from coarse to fine.

2.2. Self-attention FGVC methods

The self-attention mechanism [20–22] plays an indispensable role in deep learning whose essence is the reallocation of resources. That is to enhance important information and suppress other information. The self-attention mechanism uses the inherent information inside the feature as much as possible to perform attention interaction. Wang et al. [12] proposed non-local neural networks, which was the first to introduce self-attention into vision tasks. Compared with the traditional convolution kernel with limited receptive field, this model can capture longer distance information dependence. Behera et al. [23] proposed context-aware attentional pooling (CAP), which directly extracts diverse features at different positions on the feature map and tries to use a self-attention mechanism to highlight key parts. Transformer [24] has greatly promoted the research of natural language processing and machine translation. The model makes heavy use of the mechanism. Girdhar et al. [25] utilize a variant of the Transformer to aggregate textual cues related to a specific person in a video. Later, the Transformer model was further extended to other popular computer vision tasks, such as object detection [26], semantic segmentation [27], object tracking [28], etc. Alexey et al. [10] proposed ViT which was the first to apply Transformer directly to the field of general image classification and achieved the best results. He et al. [29] proposed Transformer architecture for fine-grained (TransFG) to apply ViT to fine-grained image classification tasks. TransFG first divides the image into several patches and uses the extracted features as the input of the ViT model, inserts a Part Selection Module (PSM) into ViT to select important features and removes irrelevant features. Finally, they pass the retained features through the rest of the model to get the final classification. In order to solve lacking the local and low-level features that are essential for FGVC. Wang et al. [30] proposed a novel pure transformer-based framework called Feature Fusion Vision Transformer (FFVT). This framework aggregates tokens at different levels in the Transformer to obtain semantic information at multiple levels. This model effectively guides the network to select differentiated tokens without introducing additional parameters.

3. Proposed method

Our proposed model is shown in Fig. 1. The model consists of two parts: the backbone network for feature extraction, the multi-granularity feature fusion module. The model can fuse multi-grained features and enhance discriminative features to generate feature vectors representing the entire image. The model first uses the advanced Swin-Transformer model to extract features from the original image, and extracts four layers of features at different scales. The features are processed through the multi-granularity feature fusion module: firstly, the features are enhanced from the two dimensions of space and channel, then the features are merged, and finally the features are fused using graph convolution to generate a semantically rich feature vector to represent the entire image for classification.

3.1. Feature extraction

Recently, Liu et al. [31] proposed the Swin-Transformer model, which has attracted much attention. This model is improved on the basis of ViT to solve the problem of the single resolution and a huge amount of calculation in the ViT model. The model is designed to be suitable for many types of vision tasks such as image classification, object detection and semantic segmentation. The effect of the model is far better than the traditional pure convolutional network. The image is first segmented into non-overlapping patches, then the patches are encoded by linear connections, and finally the extracted initial feature set is used as the input of Swin Transformer Blocks for deeper feature extraction. To produce hierarchical representations, each set of 2×2 adjacent features is merged by a linear map to reduce the number of features. This is equivalent to downsampling the feature maps by a factor of 2, and the output dimensions are set to double. The process of feature merging and feature transformation is jointly repeated. Each stage produces feature maps with different resolutions, and a total of four feature maps with different resolutions are produced.

3.2. Multi-granularity feature fusion module

After using the Swin-Transformer to extract feature maps in the previous section, it is necessary to transform each layer into the uniform size of each feature (except the last layer). As shown in Fig. 1, a total of three feature maps need to be processed. Their sizes are $48 \times 48 \times 256$, $24 \times 24 \times 512$, and $12 \times 12 \times 1024$. Use convolutions with kernel sizes of 4×4, 2×2, and 1×1. The uniform size of all feature maps is $12 \times 12 \times 1024$. Then for each feature map, self-attention and channel

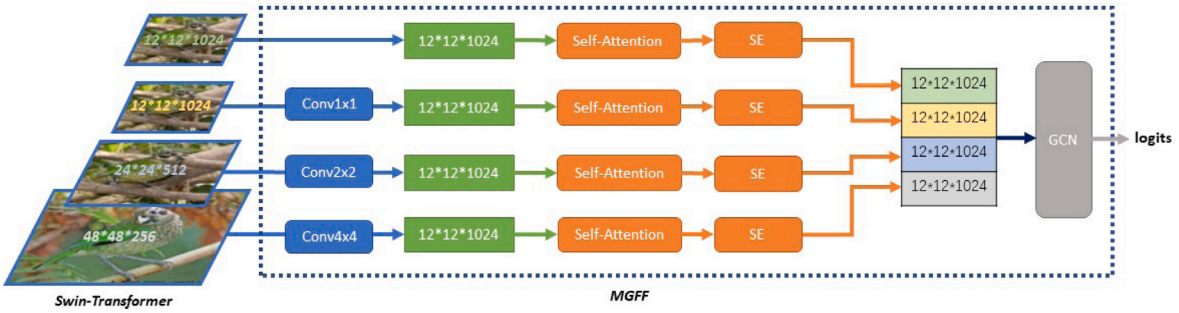

Fig. 1. Framework of the proposed model which consists of two parts: the backbone network for feature extraction, the multi-granularity feature fusion module (MGFF).

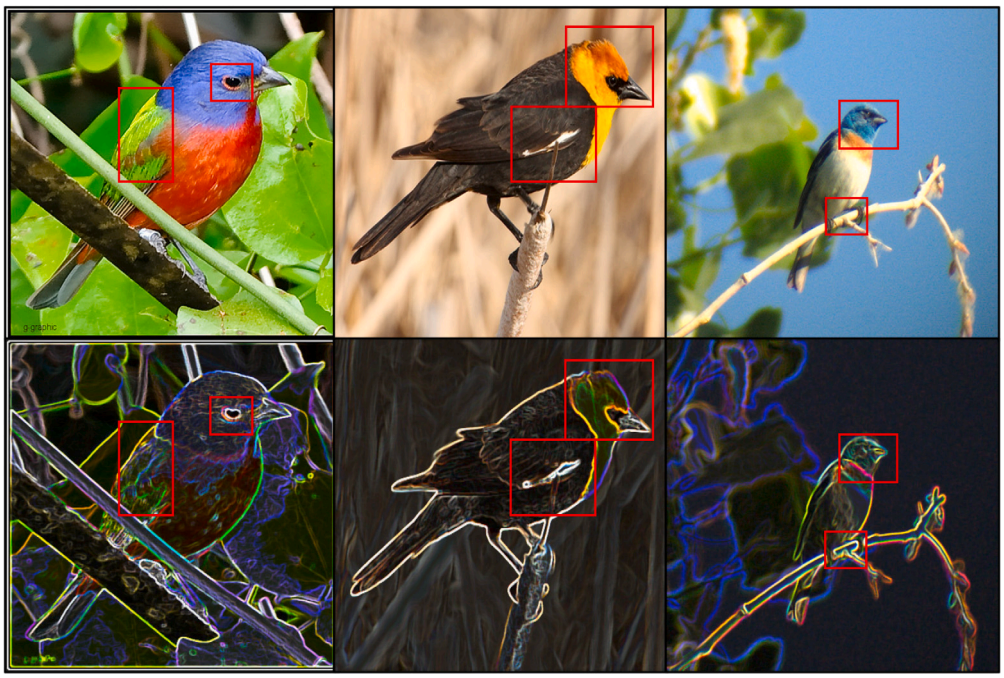

Fig. 2. The image is displayed as the information part that the model focuses on after using self-attention. The first row is the RGB image, and the second row displays heatmaps of images.

attention are used to enhance important information. Finally, the four feature maps are concatenated in the channel dimension and as input of graph convolution network to generate a semantically rich feature that finally represents the entire image. This module is not limited to a single backbone network, it can be combined with the current popular Transformer networks and traditional CNN networks. Thus, it can be treated as a plug-and-play module.

3.2.1. Self-attention

After using the backbone network to extract features, all features have the same importance. However, for fine-grained image classification, not all features are equally important. Some discriminative features should have higher weight, which is beneficial to the final prediction of the model. How to adaptively find these features and enhance them while suppressing other features is where attention comes into play. The self-attention mechanism connects features at different locations, captures long-distance dependencies, and enhances important features. Formulate the above process as follows:

First, the features are projected to the new space with full connections, and then we can get three dense matrixes the query matrix (Q), the key matrix (K) and the value matrix (V):

$$K = W_k \cdot F \tag{1}$$

$$Q = W_q \cdot F \tag{2}$$

$$V = W_v \cdot F \tag{3}$$

where W_k, W_q and W_v are the weight matrixes corresponding to the full connection, F is the feature set, and K, Q, V are the feature sets in the new feature space.

Then, the self-attention mechanism uses the dot product similarity to describe the correlation between any two features, it obtains the score matrix by multiplying the query matrix with the transposed key matrix, softmax normalization is performed to obtain the attention map:

$$A' = Softmax\left(\frac{Q \cdot K^T}{\sqrt{d_k}} + E_{pos}\right) \tag{4}$$

where d_k is the dimension of K, $\sqrt{d_k}$ can prevent gradient instability during direction propagation, E_{pos} is the relative position embedding which can preserve the position information of the feature.

Finally, a new feature set I is obtained by calculating the attention maps A' and V:

$$I = A' \cdot V \tag{5}$$

As shown in Fig. 2, after the image is processed using the self-attention mechanism, the discriminative parts of the object are significantly enhanced, while other areas are obviously suppressed.

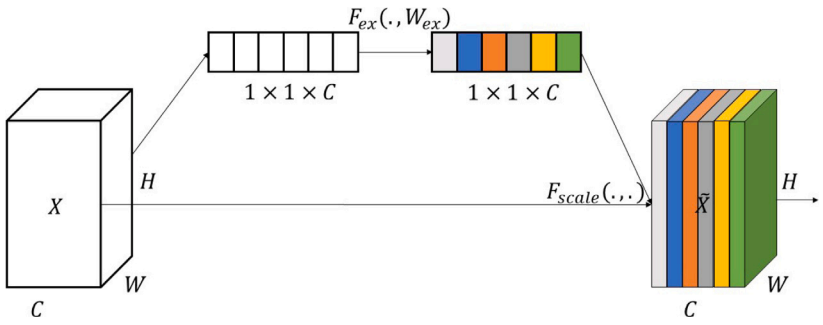

Fig. 3. As shown in the figure, the channel dimension of the feature map is enhanced by using the channel attention mechanism.

3.2.2. Channel-attention

To perform adaptive enhancement of features from the channel dimension [32], a Squeeze-and-Excitation (SE) module is introduced. This module enhances features by capturing the interdependence between channels, capable of learning global information, selectively emphasizing informative features, and suppressing less useful features.

The structure of the SE module is shown in Fig. 3. For any given input feature $X \in R^{H\times W\times C}$, the global spatial information is compressed into channel descriptors by using global mean pooling. Formally, a statistic $Z \in R^{1\times1\times C}$ is produced by compressing the spatial dimensions of X. For channel elements Z_c, the generation method is as follows:

$$Z_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} x_c(i,j) \tag{6}$$

where Z_c represents the cth element of Z, $x_c(i,j)$ represents the element of the cth channel of the input. In order to utilize aggregated information, a second operation is required to fully capture the channel-related dependencies. To achieve this, the function must meet two criteria: first, it must be flexible; second, it must allow multiple channels to be enhanced.

$$S = F_{ex}(Z,W_{ex}) = \sigma(W_2\delta(W_1 z)) \tag{7}$$

where σ is the Sigmoid function, δ is the ReLU function, F_{ex} represents the channel attention computed by Z, $W_1 \in R^{\frac{C}{r}\times C}$ and $W_2 \in R^{C\times\left(\frac{C}{r}\right)}$ are matrix weight, and r controls the complexity of the module. Use the channel attention map to scale the original features to get the final output . The corresponding specific channel can be obtained by the following formula:

$$\tilde{X}_c = F_{scale}\left(X_c, S_c\right) = S_c X_c \tag{8}$$

where $\tilde{X} = [\tilde{X}_1,\ \tilde{X}_2,\ \cdots,\ \tilde{X}_c]$, the formula represents the product of the scalar S_c and the original feature $X_c \in R^{H\times W}$.

3.2.3. Graph convolution network

Graph Convolution Networks (GCNs) [33–35] are a family of neural networks that can naturally operate on graph-structured data. By extracting and utilizing features from the underlying graph, GCNs can make more informed predictions about these associated entities than models that consider individual entities in isolation. All features on feature map are regarded as a graph structure, where nodes represent features at different spatial locations and scales. As shown in Fig. 4, the feature map is input into GCN, and the network can learn the relationship between different nodes. Then, the feature points are aggregated into multiple super nodes through the pooling layer. Finally, the features of these super nodes are averaged, and a linear classifier is used to complete the prediction. The advantage of this method is that the features of each point can be integrated more efficiently without corrupting the results output by the backbone model. Therefore, graph convolution network is finally used as a feature fusion mechanism.

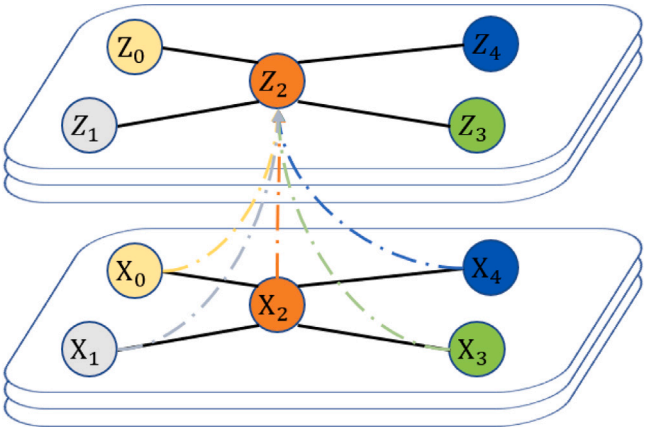

Fig. 4. As shown in the figure, the process of graph convolution network connecting original features to generate new features.

3.3. Two-stage classification based on Vision-Transformer

If the background occupies a large proportion of the image, this can lead to inaccurate predictions of the final result. To solve this problem, we extract the object from the original image at a small cost and enlarge it, so that the new image can be used to classify again. The enlarged image contains less background, while the original smaller details are enlarged, both of which are conducive to the model to capture more useful information. Finally, the results of the two phases are combined as the final result.

3.3.1. Response graph obtained through ViT model

The ViT model divides the original image $X \in R^{H\times W\times3}$ into multiple $patch \in R^{16\times16\times3}$ in a non-overlapping manner, and each patch is linearly mapped into a corresponding $token \in R^{1\times1\times C}$. The Multi-Head Attention (MHA) module uses Eq. (4) to calculate the relationship between different tokens multiple times, as shown in Fig. 5. Each column of the matrix is processed by the softmax function, combined with Eq. (5), it can be seen that the value corresponding to each column represents the proportion of the corresponding new feature. Then this non-negative value is also an important indicator to measure it. Sum the rows of the A' matrix to get the importance of each token. Since ViT is stacked by multiple MHAs, the process involves the calculation of multiple weight matrices. details as follows:

$$W_i = \sum_{l=1}^{L}\sum_{n=1}^{N}(\alpha_{iln}) \tag{9}$$

$$W' = Reshape\,(W) \tag{10}$$

$$\bar{W} = Bilinear\left(W'\right) \tag{11}$$

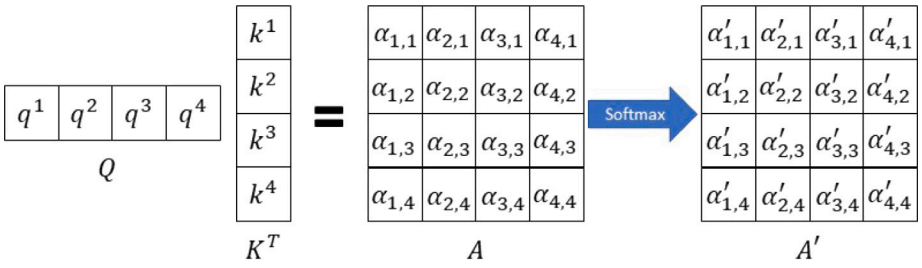

Fig. 5. Attention matrix.

where L is the number of MHAs in ViT, and $N = \frac{H}{16} \times \frac{W}{16}$ is the number of tokens. The function of Reshape is to align W_i with the token at the corresponding position, so that the one-dimensional matrix W becomes a two-dimensional matrix $W' \in R^{\frac{H}{16} \times \frac{W}{16}}$. Upsampling W' by a factor of 16 through bilinear interpolation results in a response map $\bar{W} \in R^{H \times W}$ of the same size as the original image.

3.3.2. Object location based on response graph

The specific data at different positions on the response graph $\bar{W}$ correspond to the importance of the pixel at that position. The high response is concentrated on the part where the object is located, and the low response is mostly the background. Use the following method to filter most of the background pixels, and convert the response map $\bar{W}$ into a binary image Map, as follows:

$$\text{avg} = \frac{1}{H \times W} \sum_{h=0}^{H} \sum_{w=1}^{W} (\bar{W}_{hw}) \tag{12}$$

$$Map_{hw} = \begin{cases} 0 & \text{if } \bar{W}_{hw} < avg \\ 1 & \text{if } \bar{W}_{hw} \geq avg \end{cases} \tag{13}$$

where Eq. (12) calculates the global mean of the response map. Eq. (13) keeps the position if the value is greater than avg, otherwise removes it. The white part in Fig. 6 is the reserved pixels, which can better contain objects. Although scattered points are preserved, and these special points occupy a small part, it does not have much impact on the positioning of the entire object. In order to crop the object from the original image, it is necessary to convert this irregular area into a corresponding rectangular coordinate representation. The specific method is as follows:

$$(x, y, h, w) = Location(Map) \tag{14}$$

$$Img_{crop} = Bilinear(Img[x : x + h, y : y + w]) \tag{15}$$

where Location is a method to obtain object coordinates through binary image Map. Common methods include the traversal method based on image processing, the center diffusion method to detect the largest inscribed rectangle of an object, and so on. We utilize binary opening operation to process the response map and select the largest area as the target location. For each image, we only locate one target region. Therefore, our method is only applicable to single-object fine-grained detection. Crop the original image $Img \in R^{H \times W}$ according to the coordinates, and then use bilinear interpolation to obtain a new image $Img_{crop} \in R^{H \times W}$.

3.3.3. Final two-stage classification process

As shown in Fig. 7, it shows a two-stage classification model. Use the above multi-granularity classification model to classify the original image and the cropped image respectively, and weight the classification results to get the final prediction. The formulation process is as follows:

$$Logits = Logits_S + \lambda Logits_C \tag{16}$$

where $Logits_S$ and $Logits_C$ represent the predicted probability of the original image and the cropped image respectively, and λ is the hyper-parameter (usually set to 0.8).

Fig. 6. Response graph visualization.

Table 1
Statistics of datasets.

Dataset	Train	Test	Classes
CUB-200–2011	5994	5794	200
Oxford-IIIT pets	3680	3669	37
Stanford cars	8144	8041	196
NA-Birds	23 929	24 633	555

4. Experiments

In this section, we demonstrate the detailed setup, including the datasets and comparative experiment, quantitative analysis, ablation studies, and qualitative analysis.

4.1. Experiments setup

Datasets: Experiments are evaluated on four widely used fine-grained datasets: CUB-200-2011 [2], Stanford Cars [3], Oxford-IIIT Pets [4], NA-Birds [1]. This method only uses class labels, so no comparisons are made with methods that use additional information. Table 1 counts the data division and number of categories of the dataset.

Implementation details: In all experiments, the input image is resized to 512×512, and then randomly cropped to 384×384. Since the Swin-Transformer and ViT models of different specifications have different parameters and effects, the most commonly used model of B specification is used in this experiment. This method uses an optimizer (Stochastic Gradient Descent, SGD) with a momentum of 0.99 to optimize the model. The initial learning rate is $2e^{-4}$, we use the

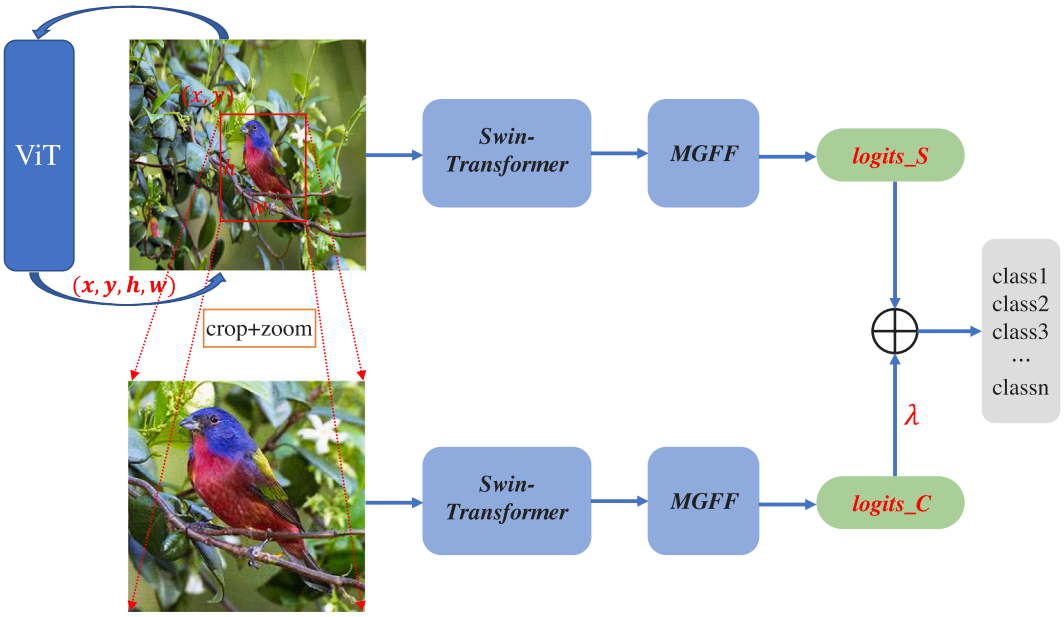

Fig. 7. The figure shows a two-stage classification process. With the assistance of the ViT model, the object is positioned, cropped, and enlarged to obtain a new image, and the model recognizes it again. Combine the results of the two stages as the final output.

Table 2
Comparing the four datasets mentioned with the most recent top five SOTA methods.

CUB-200–2011		NA-Birds	
Method	Acc (%)	Method	Acc (%)
WSCPM [14]	90.40	AP-Net [37]	88.10
FFVT [30]	91.60	MGE-CNN	88.60
TransFG [29]	91.70	FixSENet-154 [38]	89.20
SWAG [39]	91.70	TransFG [29]	90.80
CAP [23]	91.80	CAP [23]	91.00
Proposed	**92.66**	Proposed	**92.08**
Stanford cars		Oxford-IIIT pets	
Method	Acc (%)	Method	Acc (%)
ResMLP-24 [40]	89.50	EfficientNet-B7 [41]	95.40
Deit-B [42]	93.30	ALIGN [43]	96.19
AutoFormer-S [44]	93.40	BIT-L [45]	96.62
TransFG [29]	94.80	EffNet-L2 [46]	97.10
CAP [23]	95.70	CAP [23]	97.30
Proposed	93.33	Proposed	95.34

function $cos(x)$ to adjust the learning rate for each round, and train for 50 rounds. All experiments are performed using PyTorch [36] and NVIDIA RTX 3090 GPU (24 GB).

4.2. Quantitative analysis

Table 2 shows the comparison results between our method and state-of-the-art methods on the above datasets. Overall, our method outperforms the state-of-the-art methods on two datasets, and performs similarly to state-of-the-art methods on other datasets. Specifically, as shown in Table 2 (1), our method can achieve 92.60% on the CUB-200-2011 dataset, which is the most commonly used dataset in fine-grained image classification. Compared with the current best accuracy CAP, it has an improvement of 0.86%. Compared with TransFG and FFVT, which also use the powerful Transformer model as the backbone network, this method still achieves an accuracy improvement of 0.96%. TransFG is the first attempt to apply the Transformer model to fine-grained image classification. Both it and our method try to filter out important partial features for final classification. However, TransFG only selects features from the same scale and discards other features, while our method selects features from multiple scales and enhances the selected features while suppressing other features. NA-Birds is a large bird dataset with more images and categories, which further challenges fine-grained visual classification. Many models perform well on small datasets but not well on large datasets. In Table 2, this method can obtain 92.08%, far exceeding other methods. It demonstrates the robust ability of our method to build models that can effectively identify subcategories without using additional datasets or auxiliary networks. On the other two datasets Stanford Cars and Oxford-IIIT Pets, although our method does not achieve the best accuracy, it still has great advantages compared to the Deit-B, TransFG and other networks that also use Transformer. To sum up, our proposed method can preserve the advances and effectiveness of many datasets.

4.3. Ablation study

In order to prove the effectiveness of the multi-granularity feature fusion module (MGFF) proposed in this chapter and the ViT-based two-stage classification method, ablation experiments are carried out on the CUB-200-2011 dataset. In order to prove the versatility of this method, five commonly used backbone networks were selected in the experiment: Swin-Transformer, Resnet50, EfficentNet-B2, Densenet169, and Inception-V4. For each network, an inter-module ablation experiment is performed: using a separate original network, adding an MGFF module, and adding a ViT-based two-stage classification method. The entire experimental data is shown in Table 3, showing the accuracy of each model and the corresponding gain after adding modules. For different networks, the gain of the module is different: after adding the MGFF module, the network using Densenet169 increased by 0.90%, while the network using EfficentNet-B2 only increased by 0.19%. For the ViT-based two-stage classification method, the network using Inception-V4 increased by 0.58%, while the network using Swin-Transform only increased by 0.13%. Comparing the gains of the two parts, it is not difficult to find that the network with a larger gain after adding the MGFF module tends to have a smaller improvement after using the ViT two-stage classification. As for the gain of the final network depends on the sum of the two parts, the network using Densenet169 has an improvement of 1.43%, while only 0.60% for the network using EfficentNet-B2. Although the gains of this method for different networks are different, they are all significantly improved, which is

Table 3
Ablation experiments of different backbone networks and different modules.

CUB-200–2011				
Backbone	Base	+MGFF	+Two-stage	Sum
Swin-Transform [31]	91.56%	92.5%(+0.97%)	92.66%(+0.13%)	+1.10%
Resnet50 [5]	85.28%	85.64%(+0.36%)	85.97%(+0.33%)	+0.69%
EfficentNet-B2 [47]	85.26%	85.45%(+0.19%)	85.86%(+0.41%)	+0.60%
Densenet169 [6]	84.59%	85.49%(+0.90%)	86.02%(+0.53%)	+1.43%
Inception-V4 [48]	84.81%	85.04%(+0.23%)	85.62%(+0.58%)	+0.81%

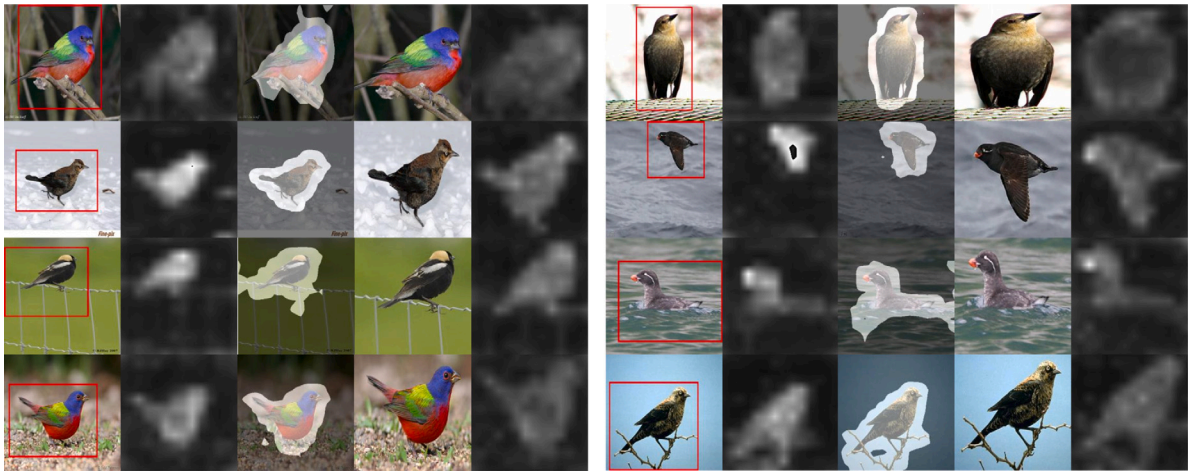

Fig. 8. Visual attention map and corresponding object localization.

Table 4
Compare the results of GCN with No GCN.

	CUB-200–2011	NA-Birds	Stanford cars	Oxford-IIIT pets
No GCN	91.82	90.46	92.54	95.26
GCN	92.66	92.08	93.33	95.34

Table 5
Compare the results of the model with and without the self-attention(SA) mechanism.

	CUB-200–2011	NA-Birds	Stanford cars	Oxford-IIIT pets
No SA	91.96	91.05	93.03	95.28
SA	92.66	92.08	93.33	95.34

enough to prove the effectiveness and robustness of our method and can adapt to most networks and tasks.

Compared to CNN, GCN can extract more representative features by cross-layer feature propagation. It also takes into account the relationships between nodes and their neighboring nodes. This helps to capture global structural information. From Table 4, it can be observed that the model with GCN has better classification performance than the model without GCN. The self-attention mechanism enhances the discriminative features and suppresses other features. The Table 5 demonstrates the effectiveness of self-attention mechanism.

4.4. Qualitative analysis

In order to better understand the ViT-based two-stage classification method, the experiment randomly selected 8 images from the CUB-200-2011 dataset for visualization, as shown in Fig. 8. Columns 1 and 6 in the figure are the original image, and the red box on the figure is the display of the coordinates obtained by formula (14) on the original image. Columns 2 and 7 in the figure are the visualization results of the weight matrix of formula (9). The brighter the pixel on the picture, the larger the value of the position. Columns 3 and 8 in the figure are the visualization results of superimposing the matrix obtained in formula (9) with the original image. After filtering, most pixels in the image appear dark, but the position of the object is still highlighted. Subsequent positioning of specific objects through these highlighted pixels, it can be seen that the objects are all in the highlighted area on the picture, which shows that it is reasonable to use this method to locate objects. Columns 4 and 9 in the figure are the images cropped and enlarged according to the positioning. The proportion of the object in the image and the details become larger, which is conducive to the accurate classification of the model. Columns 5 and 10 in the figure are the response maps of the cropped image. It can be seen from the figure that the highlighted area of the object occupies the majority and the information is more abundant.

5. Conclusion

This paper proposes a fine-grained image classification model based on Transformer's multi-granularity feature fusion. This method uses the currently more advanced Swin-Transformer model to extract features and select feature maps with different resolutions. Through the multi-granularity feature fusion module, the features of different granularities are fused. And the attention mechanism is used to enhance the features in the two dimensions of channel and space. The fused features have both high-semantic global information and low-semantic local information. In addition, the Vision-Transformer is used as an auxiliary model to locate the position of the object in the image at a very small cost. After image processing, the object is separated from the background to the greatest extent so as to reduce the impact on the classification results. The multi-granularity feature fusion module is a plug-and-play module that can be combined with the currently popular Transformer and traditional CNN networks. The whole method is efficient and has high scientific research and application value.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62071233, 61971223, 61976117), the Jiangsu Provincial Natural Science Foundation of China (BK20211570, BK20191409), the Fundamental Research Funds for the Central Universities, China (30919011103,30919011402, 30921011209).

References

[1] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, Serge J. Belongie, Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, IEEE Computer Society, 2015, pp. 595–604.

[2] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, Serge Belongie, The Caltech-UCSD Birds-200–2011 Dataset, California Institute of Technology, 2011.

[3] Jonathan Krause, Michael Stark, Jia Deng, Li Fei-Fei, 3D object representations for fine-grained categorization, in: 2013 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013, IEEE Computer Society, 2013, pp. 554–561.

[4] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, C.V. Jawahar, Cats and dogs, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012, IEEE Computer Society, 2012, pp. 3498–3505.

[5] Abhishek Verma, Hussam Qassim, David Feinzimer, Residual squeeze CNDS deep learning CNN model for very large scale places image recognition, in: 8th IEEE Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, UEMCON 2017, New York City, NY, USA, October 19-21, 2017, IEEE, 2017, pp. 463–469.

[6] Dilbag Singh, Vijay Kumar, Manjit Kaur, Densely connected convolutional networks-based COVID-19 screening model, Appl. Intell. 51 (5) (2021) 3044–3051.

[7] Bikash Santra, Avishek Kumar Shaw, Dipti Prasad Mukherjee, Part-based annotation-free fine-grained classification of images of retail products, Pattern Recognit. 121 (2022) 108257.

[8] Xiao Ke, Yuhang Cai, Baitao Chen, Hao Liu, Wenzhong Guo, Granularity-aware distillation and structure modeling region proposal network for fine-grained image classification, Pattern Recognit. 137 (2023) 109305.

[9] Xiu-Shen Wei, Chen-Wei Xie, Jianxin Wu, Chunhua Shen, Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization, Pattern Recognit. 76 (2018) 704–714.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net, 2021.

[11] Qi Wang, JianJun Wang, Hongyu Deng, Xue Wu, Yazhou Wang, Gefei Hao, AA-trans: Core attention aggregating transformer with information entropy selector for fine-grained visual classification, Pattern Recognit. 140 (2023) 109547.

[12] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, Kaiming He, Non-local neural networks, 2017, CoRR, abs/1711.07971.

[13] Jonathan Long, Evan Shelhamer, Trevor Darrell, Fully convolutional networks for semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, IEEE Computer Society, 2015, pp. 3431–3440.

[14] Weifeng Ge, Xiangru Lin, Yizhou Yu, Weakly supervised complementary parts models for fine-grained image classification from the bottom up, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 3034–3043.

[15] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, Antonio Torralba, Learning deep features for discriminative localization, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society, 2016, pp. 2921–2929.

[16] Charles Sutton, Andrew McCallum, An introduction to conditional random fields, Found. Trends Mach. Learn. 4 (4) (2012) 267–373.

[17] Klaus Greff, Rupesh Kumar Srivastava, Jan Koutník, Bas R. Steunebrink, Jürgen Schmidhuber, LSTM: A search space odyssey, IEEE Trans. Neural Netw. Learn. Syst. 28 (10) (2017) 2222–2232.

[18] Jianlong Fu, Heliang Zheng, Tao Mei, Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, 2017, pp. 4476–4484.

[19] Zhao Pei, Zhiyang Wan, Yanning Zhang, Miao Wang, Chengcai Leng, Yee-Hong Yang, Multi-scale attention-based pseudo-3D convolution neural network for Alzheimer's disease diagnosis using structural MRI, Pattern Recognit. 131 (2022) 108825.

[20] Wenqian Zhu, Zhongyuan Wang, Xiaochen Wang, Ruimin Hu, Huikai Liu, Cheng Liu, Chao Wang, Dengshi Li, A dual self-attention mechanism for vehicle re-identification, Pattern Recognit. 137 (2023) 109258.

[21] Xinjian Gao, Zhao Zhang, Tingting Mu, Xudong Zhang, Chaoran Cui, Meng Wang, Self-attention driven adversarial similarity learning network, Pattern Recognit. 105 (2020) 107331.

[22] Mingyang Zhang, Hanhong Zheng, Maoguo Gong, Yue Wu, Hao Li, Xiangming Jiang, Self-structured pyramid network with parallel spatial-channel attention for change detection in VHR remote sensed imagery, Pattern Recognit. 138 (2023) 109354.

[23] Ardhendu Behera, Zachary Wharton, Pradeep R.P.G. Hewage, Asish Bera, Context-aware attentional pooling (CAP) for fine-grained visual classification, in: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, AAAI Press, 2021, pp. 929–937.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, Attention is all you need, in: Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S.V.N. Vishwanathan, Roman Garnett (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 5998–6008.

[25] Rohit Girdhar, João Carreira, Carl Doersch, Andrew Zisserman, Video action transformer network, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 244–253.

[26] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, Sergey Zagoruyko, End-to-end object detection with transformers, in: Andrea Vedaldi, Horst Bischof, Thomas Brox, Jan-Michael Frahm (Eds.), Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I, in: Lecture Notes in Computer Science, vol. 12346, Springer, 2020, pp. 213–229.

[27] Enze Xie, Wenjia Wang, Wenhai Wang, Peize Sun, Hang Xu, Ding Liang, Ping Luo, Trans2Seg: Transparent object segmentation with transformer, 2021, CoRR, abs/2101.08461.

[28] Peize Sun, Yi Jiang, Rufeng Zhang, Enze Xie, Jinkun Cao, Xinting Hu, Tao Kong, Zehuan Yuan, Changhu Wang, Ping Luo, TransTrack: Multiple-object tracking with transformer, 2020, CoRR, abs/2012.15460.

[29] Ju He, Jieneng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, Changhu Wang, Alan L. Yuille, TransFG: A transformer architecture for fine-grained recognition, 2021, CoRR, abs/2103.07976.

[30] Jun Wang, Xiaohan Yu, Yongsheng Gao, Feature fusion vision transformer for fine-grained visual categorization, 2021, CoRR, abs/2107.02341.

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, IEEE, 2021, pp. 9992–10002.

[32] Xin Jin, Yanping Xie, Xiu-Shen Wei, Bo-Rui Zhao, Zhao-Min Chen, Xiaoyang Tan, Delving deep into spatial pooling for squeeze-and-excitation networks, Pattern Recognit. 121 (2022) 108159.

[33] Zi Ye, Yogan Jaya Kumar, Goh Ong Sing, Fengyan Song, Junsong Wang, A comprehensive survey of graph neural networks for knowledge graphs, IEEE Access 10 (2022) 75729–75741.

[34] Fei Wu, Shuaishuai Li, Guangwei Gao, Yimu Ji, Xiao-Yuan Jing, Zhiguo Wan, Semi-supervised cross-modal hashing via modality-specific and cross-modal graph convolutional networks, Pattern Recognit. 136 (2023) 109211.

[35] Li Zhang, Heda Song, Nikolaos Aletras, Haiping Lu, Node-feature convolution for graph convolutional networks, Pattern Recognit. 128 (2022) 108661.

[36] Antonio Carta, Lorenzo Pellegrini, Andrea Cossu, Hamed Hemati, Vincenzo Lomonaco, Avalanche: A PyTorch library for deep continual learning, 2023, CoRR, abs/2302.01766.

[37] Peiqin Zhuang, Yali Wang, Yu Qiao, Learning attentive pairwise interaction for fine-grained classification, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press, 2020, pp. 13130–13137.

[38] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, Hervé Jégou, Fixing the train-test resolution discrepancy: FixEfficientNet, 2020, CoRR, abs/2003.08237.

[39] Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross B. Girshick, Piotr Dollár, Laurens van der Maaten, Revisiting weakly supervised pre-training of visual perception models, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, IEEE, 2022, pp. 794–804.

[40] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, Hervé Jégou, ResMLP: Feedforward networks for image classification with data-efficient training, 2021, CoRR, abs/2105.03404.

[41] Mingxing Tan, Quoc V. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in: Kamalika Chaudhuri, Ruslan Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, in: Proceedings of Machine Learning Research, vol. 97, PMLR, 2019, pp. 6105–6114.

[42] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, Hervé Jégou, Training data-efficient image transformers & distillation through attention, in: Marina Meila, Tong Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, in: Proceedings of Machine Learning Research, vol. 139, PMLR, 2021, pp. 10347–10357.

[43] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, Tom Duerig, Scaling up visual and vision-language representation learning with noisy text supervision, in: Marina Meila, Tong Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, in: Proceedings of Machine Learning Research, vol. 139, PMLR, 2021, pp. 4904–4916.

[44] Minghao Chen, Houwen Peng, Jianlong Fu, Haibin Ling, AutoFormer: Searching transformers for visual recognition, 2021, CoRR, abs/2107.00651.

[45] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, Neil Houlsby, Big transfer (BiT): General visual representation learning, in: Andrea Vedaldi, Horst Bischof, Thomas Brox, Jan-Michael Frahm (Eds.), Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V, in: Lecture Notes in Computer Science, vol. 12350, Springer, 2020, pp. 491–507.

[46] Pierre Foret, Ariel Kleiner, Hossein Mobahi, Behnam Neyshabur, Sharpness-aware minimization for efficiently improving generalization, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net, 2021.

[47] Mingxing Tan, Quoc V. Le, EfficientNetV2: Smaller models and faster training, in: Marina Meila, Tong Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, in: Proceedings of Machine Learning Research, vol. 139, PMLR, 2021, pp. 10096–10106.

[48] Feng Chen, Jiangshu Wei, Bing Xue, Mengjie Zhang, Feature fusion and kernel selective in Inception-v4 network, Appl. Soft Comput. 119 (2022) 108582.

Yang Xu received the B.Sc. degree in applied mathematics and the Ph.D. degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology (NUST), Nanjing, China, in 2011 and 2016, respectively. He is currently an Associate Professor with the School of Computer Science and Engineering, NUST. His research interests include hyperspectral image classification, hyperspectral detection, image processing, and machine learning.

Shanshan Wu received the B.Sc. degree in Computer Science and Technology from Nanjing University of Science and Technology, Nanjing, China. Currently, she is an advanced researcher with the Nanjing Research Institute of Electronics Engineering. Her research interests include image processing, and intelligent computing.

Biqi Wang received the B.B.A. degree in Information Management and Information System from the School of Information Engineering, Nanjing University of Finance and Economics, Nanjing, China, in 2019. She is currently pursuing the Ph.D. degree with the Nanjing University of Science and Technology, Nanjing, China. Her research interests are transfer learning, hyperspectral images processing and deep learning.

Ming Yang was born in Anhui, China, in 1998. He received the B.Sc. degree in computer science and technology from the School of Management Science and Engineering, Anhui University of Finance and Economics in 2020. He is currently pursuing the M.Sc. degree with Nanjing University of Science and Technology. His research interests are fine-grained classification, and deep learning.

Zebin Wu was born in Zhejiang, China, in 1981. He received the B.Sc. and Ph.D. degrees in computer science and technology from the Nanjing University of Science and Technology, Nanjing, China, in 2003 and 2007, respectively. He was a Visiting Scholar with the Department of Mathematics, University of California at Los Angeles, Los Angeles, CA, USA, from August 2016 to September 2016 and from July 2017 to August 2017. He was a Visiting Scholar with the Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, Escuela Politécnica, University of Extremadura, Cáceres, Spain, from June 2014 to June 2015. He is currently a Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests include hyperspectral image processing, high-performance computing, and computer simulation.

Yazhou Yao is currently a Professor at the School of Computer Science and Engineering, Nanjing University of Science and Technology. He obtained the Ph.D. degree in 2018 from Global Big Data Technologies Center (GBDTC), University of Technology Sydney, Australia. From July 2018 to July 2019, he worked as a Research Scientist at Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. His research interests include multimedia, computer vision, and machine learning.

Zhihui Wei was born in Jiangsu, China, in 1963. He received the B.Sc. and M.Sc. degrees in applied mathematics and the Ph.D. degree in communication and information system from South East University, Nanjing, China, in 1983, 1986, and 2003, respectively. He is currently a Professor and a Doctoral Supervisor with the Nanjing University of Science and Technology (NUST), Nanjing. His research interests include partial differential equations, mathematical image processing, multiscale analysis, sparse representation, and compressive sensing.