Journal of Systems Engineering and Electronics
Vol. PP, No. 99, February 2023, pp.1 – 20

A survey of fine-grained visual categorization based on deep learning

XIE Yuxiang[1,*], GONG Quanzhi[1,†], LUAN Xidao[2], YAN Jie[1], and ZHANG Jiahui[1]

1. College of System Engineering, National University of Defense Technology, Changsha 410000, China;
2. College of Computer Engineering and Applied Mathematics, Changsha University, Changsha 410003, China

Abstract: Deep learning has achieved excellent results in various tasks in the field of computer vision, especially in fine-grained visual categorization. It aims to distinguish the subordinate categories of the label-level categories. Due to high intra-class variances and high inter-class similarity, the fine-grained visual categorization is extremely challenging. This paper first briefly introduces and analyzes the related public datasets. After that, some of the latest methods are reviewed. Based on the feature types, the feature processing methods, and the overall structure used in the model, we divide them into three types of methods: methods based on general convolutional neural network (CNN) and strong supervision of parts, methods based on single feature processing, and methods based on multiple feature processing. Most methods of the first type have a relatively simple structure, which is the result of the initial research. The methods of the other two types include models that have special structures and training processes, which are helpful to obtain discriminative features. We conduct a specific analysis on several methods with high accuracy on public datasets. In addition, we support that the focus of the future research is to solve the demand of existing methods for the large amount of the data and the computing power. In terms of technology, the extraction of the subtle feature information with the burgeoning vision transformer (ViT) network is also an important research direction.

Keywords: deep learning, fine-grained visual categorization, convolutional neural network (CNN), visual attention.

DOI: 10.23919/JSEE.2022.000155

1. Introduction

Deep learning has recently achieved excellent results in many computer vision (CV) tasks, such as object detection, action recognition, and semantic segmentation. The core of deep learning is that the large-scale parameters of the model are not set manually, but are learned automatically with a large amount of data. There are many deep networks in the field of CV, the most important of which is the convolutional neural network (CNN) method.

As a multilayer feedforward neural network, CNN is popular in visual categorization problems. Due to the different category granularity of images, image classification tasks can be divided into ordinary image classification and fine-grained visual categorization. The general granularity has a coarser granularity, whose purpose is to distinguish the label-level categories of objects in the images. The fine-grained visual categorization is more refined, and it aims to distinguish the subcategories of some coarse-grained categories. The research on fine-grained visual categorization can be traced back to the subordinate classification task, which was proposed in [1] about Caltech UCSD Birds (CUB) 200 dataset. Since then, fine-grained visual categorization has become the focus of research in CV. Specifically, the most popular definition of fine-grained visual categorization was proposed by Yao et al. [2]: fine-grained categorization refers to the task of classifying objects that belong to the same basic-level class and share similar shape or visual appearances. Due to the prominent characteristics of large intra-class gaps and high inter-class similarity, fine-grained visual categorization is more difficult for both humans and computers.

Specifically, the problems of fine-grained visual categorization are as follows: (i) A small amount of data. Due to the requirement for the granularity of the images,

Manuscript received January 10, 2022.
†Co-first authors.
*Corresponding author.
This work was supported by the National Natural Science Foundation of China (61571453; 61806218).

the data collection is so difficult that there are no large-scale data that can be used for model learning. (ii) The large intra-class gaps and high inter-class similarity. Therefore, the discriminative features need to be extracted, which is not an easy task. (iii) The complex image interference. The background in the image sometimes interferes with the classification, and the occlusion relationship formed by the object itself and the environment in the images may also influence the classification efficiency.

Since fine-grained visual categorization was proposed, researchers have developed various models. By observing the mainstream methods in different periods, it is easy to find that with the continuous improvement of deep network performance, the design ideas of the model are obviously different.

In 2014, the mainstream method was the strong supervision model, which emphasized the detection of image blocks that may contain effective features before feature extraction. The iconic method is part-based region-CNN (PB R-CNN)[3], which has the advantage of strong interpretability. The boundary box annotation can intuitively guide the network to extract the discriminative features of objects. However, the disadvantages are also obvious. Complex manual annotation information is more expensive, and the running time cost of the network is large. Since 2015, weak supervision network has become the focus of research, and the bilinear CNN (B-CNN) [4] was popular at that time. It encodes different features from two networks with independent parameters to enhance the expression ability of the representation. Although it multiplies the parameters of the network, it does not complicate the overall structure. After 2017, the method of using features from multi-scale images has become mainstream. A more representative network is recurrent attention CNN (RA-CNN) [5], which enriches feature representations by extracting both overall features and part-level features. The advantage of RA-CNN is to improve the mining degree of image key information, but the network structure is obviously much more complex. Recently, the models based on the self-attention network, vision tranformer (ViT) [6] are popular. They have simple structures and excellent classification accuracy, but they are defective because of their long calculation time and large number of parameters. In addition, in each stage of the above research, there are researchers trying to develop methods by combining models, which can always improve the performance compared with the single model method, but with doubled parameters and complex network structure.

In summary, fine-grained visual categorization is a challenging task and has received substantial attention in recent years. With the development of deep neural networks, research in this field has also produced many surprising results. This paper outlines the latest developments in various models based on deep neural networks in the field of fine-grained visual categorization and divides them into three types of methods, according to the feature types, feature processing methods, and the overall structure used. They are methods based on general CNN and strong supervision of parts, methods based on single feature processing, and methods based on multiple feature processing.

The contents of this paper are arranged as follows: Section 2 describes the commonly used fine-grained visual categorization public datasets. Section 3 summarizes the existing fine-grained visual categorization models based on deep learning. Section 4 compares the performance of them in common datasets. Finally, future development of this field is analyzed in Section 5.

2. Datasets

Because of the requirements for the vast amounts of fine-grained images and the detailed annotation, the data collection and labeling are difficult. There are some commonly used public datasets for verifying the effectiveness of fine-grained visual categorization models, which can be roughly divided into two categories: datasets with a single label-level category and datasets with different label-level categories. Table 1 shows the statistics of all datasets, where “BBox” refers to the bounding box label; “Part” refers to the location label of the object key part; “HL” refers to hierarchical labeling; “AL” refers to attribute labeling; “Texts” refers to the text description and annotation of the image.

It should be noted that the images in the CUB-200-2011 dataset and the Stanford Dogs dataset overlap with the images in the ImageNet dataset. Therefore, more attention needs to be given when using the pretrained model with the ImageNet dataset on the above two datasets.

Table 1 Common public datasets for fine-grained visual categorization

Dataset	Number	Type	Year	Theme	Annotation
Oxford Flower [7]	8 189	103	2008	Flowers	Texts
CUB-200-2011 [8]	11 788	200	2011	Birds	BBox/Part/AL/Texts
Stanford Dogs [9]	20 580	120	2011	Dogs	BBox
Stanford Cars [10]	16 185	196	2013	Cars	BBox
FGVC Aircraft [11]	10 000	100	2013	Airplanes	BBox/HL
VegFru [12]	160 731	292	2017	Vegetable/fruit	HL
iNat2017 [13]	857 877	5 089	2017	Animals/plants	BBox/HL
RPC [14]	83 739	200	2019	Products	BBox/HL

2.1 Datasets with a single label-level category

In the collected fine-grained visual categorization datasets, the ones with a single label-level category are relatively simple. Each fine-grained class belongs to the same label-level category, which makes it possible to obtain features from the fixed parts of the images to reduce the difficulty of classification.

(i) Oxford Flower dataset [7]: It contains 8 189 images in 103 categories, which are divided into 1 030 images in the training set and validation set and 6 129 images in the test set. Among them, the number of images of each type is not completely balanced, and each type of flower contains 40−250 images.

(ii) CUB-200-2011 dataset [8]: It is a 200-category bird fine-grained visual categorization dataset. The total number of images has reached 11 788, corresponding to 5 994 images in the training set and 5 794 images in the test set. In terms of labeling, in addition to the bounding box, it also contains attribute data and part-related information, specifically, the pixel position and visibility information of a total of 15 parts.

(iii) Stanford Dogs dataset [9]: It contains 20 580 images of 120 types of dogs. The training set contains 12 000 images. Its annotations include labels and bounding boxes. Notably, the backgrounds of the images change greatly, which makes the problem more complicated.

(iv) Stanford Cars dataset [10]: It provides 16 185 images in 196 categories, of which 8 144 images are in the training set and 8 041 images are in the test set. The differences in vehicle types in this dataset are reflected in the brand, model, and announced year. The annotations include labels and bounding boxes.

(v) Fine-grained visual classification of Aircraft (FGVC Aircraft) dataset [11]: It provides 10 000 images covering 100 types of aircraft, in which the numbers of images in the training set, validation set and test set are equally divided. Hierarchical classification is performed according to different granularities. The annotation only includes bounding box information.

2.2 Datasets with multiple label-level category

Datasets with multiple label-level categories are more difficult to classify because the fine-grained categories do not belong to the same label-level category, which causes the deep neural network to encounter more difficulty in the model learning process and increases the complexity of the annotation labeling.

(i) VegFru dataset [12]: It can be divided into two major categories: vegetables and fruits, covering a total of 25 upper categories and 292 subcategories including 160 731 images. Among them, the training set has 43 800 images, and the test set has 116 931 images. It should be noted that although this dataset has overlapping images with the ImageNet dataset, their labels may not be exactly the same. In addition, the images of the two datasets are different in terms of classification tendency.

(ii) iNaturalist 2017 (iNat2017) dataset [13]: It is a fine-grained visual categorization dataset covering 13 major categories, 5 089 subcategories, and 857 877 images of animals and plants. Its training and validation set contains 675 170 images, while the test set has 182 707 images. Regarding the annotation, this dataset provides bounding boxes, but some inconveniently labeled images are not labeled, such as bushes and kelp.

(iii) Retail Product Checkout (RPC) dataset [14]: It mainly involves retail products and consists of 83 739 images, which are divided into 200 subcategories and 17 major categories. The training set contains 59 739 images, and the test set contains 24 000 images. In terms of labeling, it contains the category labels, the corresponding instance counts of all objects in the images, the center position of the instances, and the bounding box of the instances.

3. Fine-grained visual categorization based on deep learning networks

In the task of fine-grained visual categorization, the deep learning network has a satisfying performance. There are many existing methods for fine-grained visual categorization based on deep learning. In the existing methods, the use of annotation information and the

supervision of classification methods are diverse. Therefore, we divide these methods into the following three categories according to the feature types, the feature processing methods, and the overall structure used in the network: (i) methods based on general CNN and strong supervision of parts; (ii) methods based on single feature processing; (iii) methods based on multiple feature processing. The classification situation is shown in Fig. 1. In this way, the methods can be better compared with the networks with similar structural complexity, and it is also easier for the innovations of different methods to be understood.

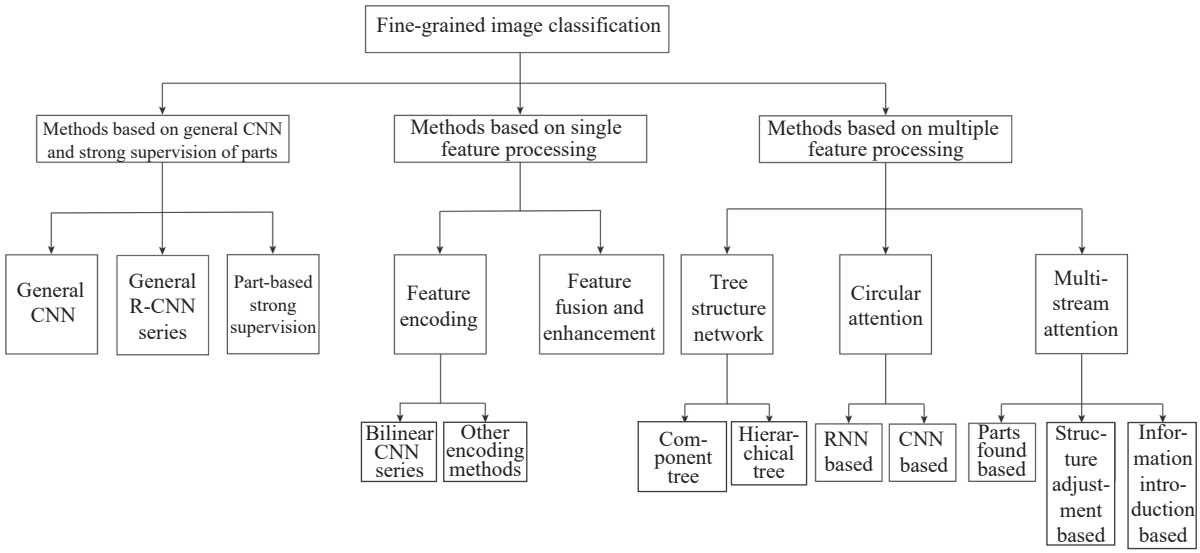


Fig. 1 Overview of fine-grained visual categorization methods

3.1 Methods based on general CNN and strong supervision of parts

Ever since the application of deep learning technology in the field of CV, CNNs have been popular and applied in many fields. The R-CNN series networks are widely used. These highly versatile CNNs can be directly applied to the fine-grained visual categorization task. After the advent of these R-CNN networks, some part-based strong supervision methods inherited similar ideas and promoted the development of this field. The common characteristics of these methods are that they are not specific to the problem of fine-grained visual categorization. And the application of the models is relatively straightforward, which belongs to the achievements in the early stage of exploration in this field. Fig. 2 presents a schematic diagram of these methods.

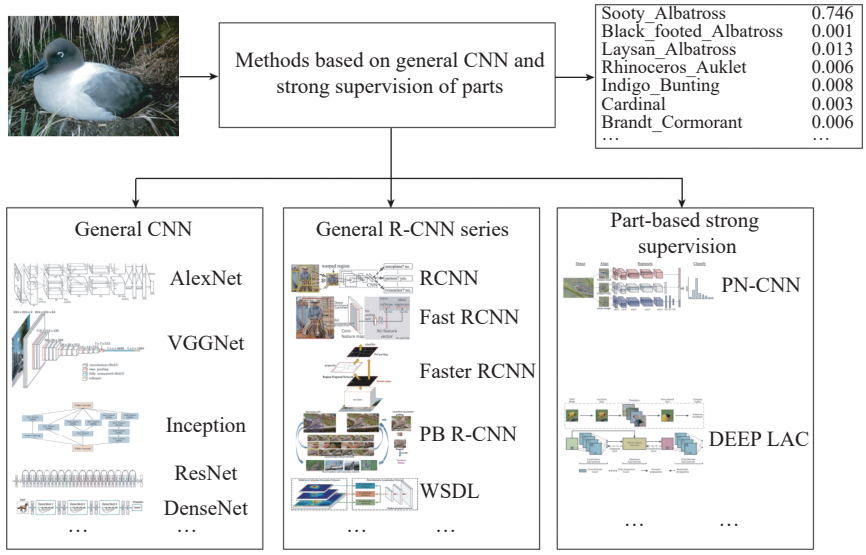


Fig. 2 Methods based on general CNN and strong supervision of parts

The earliest CNN was established by LeCun et al. [15] and was called LeNet-5. It is a very basic network, containing the basic modules of deep learning: convolutional layers, pooling layers, and fully connected layers. Since then, AlexNet [16], Visual Geometry Group networks (VGGNet) [17] and GoogleNet [18] appeared one after another, introducing the ReLU structure, the dropout method and the small convolution kernel. ResNet created the concept of residual learning [19], which effectively prevented the performance degradation caused by an increase in the depth of the deep network. Then, DenseNet researched feature reuse, which improved the efficiency of feature use [20]. In addition, networks such as squeeze-and-excitation networks (SENet) [21], Highway Network [22] and FractalNet [23] also achieved good results in the field of CV. At the beginning, researchers directly applied these networks to fine-grained visual categorization, which are still used as backbone networks to extract the fine-grained features.

In addition to the above mentioned general CNN, the more commonly used networks in object recognition and image classification tasks are the R-CNN series of networks (R-CNN [24], fast R-CNN [25], faster R-CNN [26]), whose idea is that they extract a large number of regional blocks from the image to filter and merge to obtain candidate regions and to classify the images with the features of the candidate regions. The advantage of these methods is that they eliminate the influence of the background in images, so that the feature expression is more concentrated on the discriminative parts. In the initial research, the application of the R-CNN was based on strong supervision. PB R-CNN [3] uses an R-CNN and spatial constraint methods to search for image parts and uses the features of a strong resolution of components to classify fine-grained images. Afterward, the researchers explored weak supervision to overcome the difficulty of label acquisition encountered by strong supervision. Weakly supervised discriminative localization (WSDL) uses an attention extraction network to generate bounding box information to guide faster R-CNN to complete classification, which reduces the need for bounding box labeling and achieves an accuracy of 85.71% on the same dataset [27].

It should be noted that the single shot multibox detector (SSD) network [28] and the You Only Look Once (YOLO) networks [29−32] were proposed as effective object detection methods after R-CNN, which have obvious advantages in positioning accuracy and the speed of detection. Because the object detection task mostly aims to detect the images with low classification difficulty, the models are developed and optimized for localization rather than classification. Therefore, SSD network and YOLO networks have no outstanding improvement in the extraction and analysis of fine-grained features, which makes these methods can hardly get competitive results in fine-grained visual categorization. The latest relevant research is the model combining the YOLOv3 with the bilinear features [33]. The method in [31] used YOLOv3 to roughly obtain the position of the object in the image, and extracted features with the background suppression method and the improved bilinear convolution network, whose accuracy was 86.3% on the CUB-200-2011 dataset. Unfortunately, this method with complex structure does not have obvious advantage in accuracy compared with other models in the same period. However, from the perspective of model fusion, the combination of the object detection algorithm and the fine-grained visual categorization method is worth studying.

Based on the idea about strong supervision of parts [3], pose-normalized CNN (PN-CNN) [34] proposed the method based on posture normalization, which reduces the misjudgment rate caused by different postures of similar parts. On this basis, deep locatization alignment and classification (DEEP LAC) [35] integrates the positioning, alignment, and classification of parts into a large framework and improves the accuracy to 80.26% on the CUB-200-2011 dataset.

3.2 Methods based on single feature processing

To solve the problem of fine-grained visual categorization, it is important to extract the discriminative features and construct appropriate expressions. Methods based on single feature processing refer to the methods in which the network focuses on processing the same level of features and constructs a single feature expression in the entire model framework. Such methods are mainly divided into feature coding methods and feature fusion and enhancement methods. The classification of these methods is shown in Fig. 3.

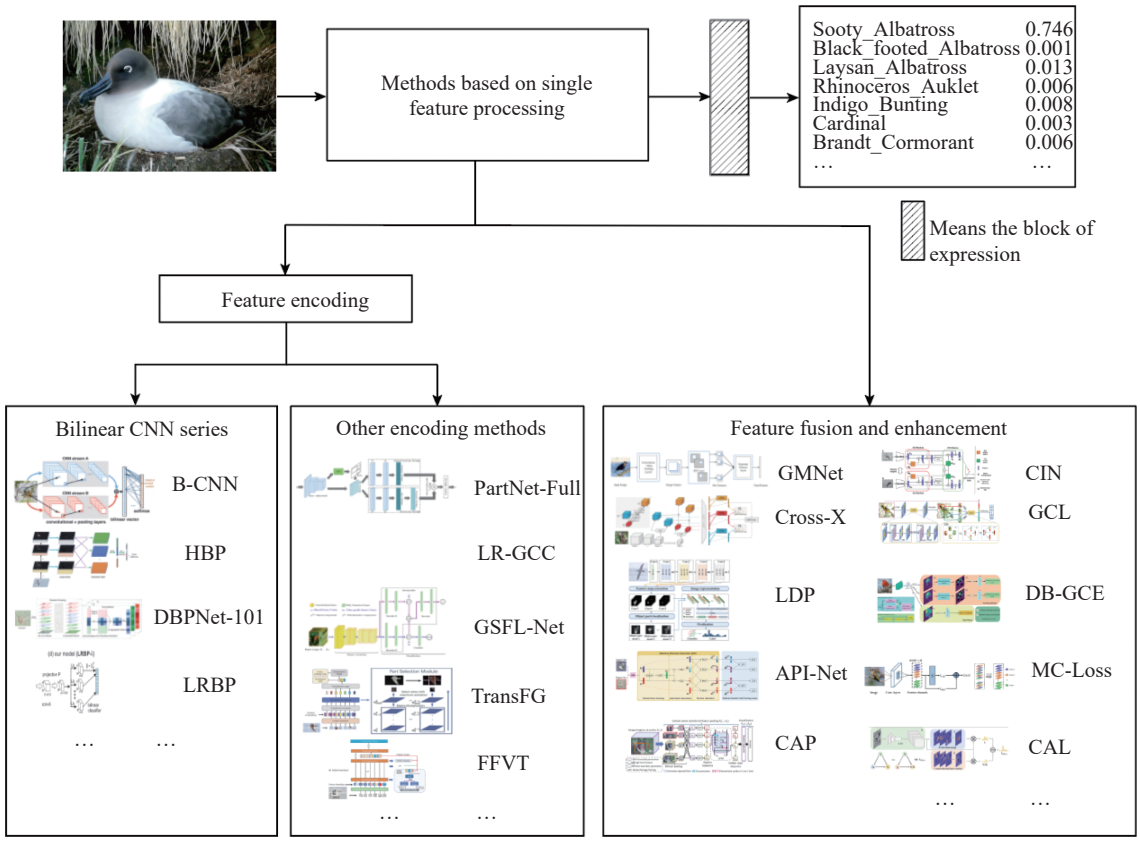


Fig. 3 Methods based on single feature processing

3.2.1 Feature coding methods

Feature encoding methods mainly refer to the methods to extract features and form expressions by processing the images. Among them, the most common and effective method is to use the high order features with specific design, which is a CNN feature fusion representation and includes most of the discriminative information in the images. In practical application, the most commonly used method of this type is the bilinear CNN (B-CNN).

A more classic technique is B-CNN [4], which uses a dual-stream network to obtain the object location and the corresponding location features, and then calculates the outer product of the two feature maps and performs a series of transformations. The final feature representation is obtained and input into a classifier, such as a support vector machine (SVM). A schematic diagram of the structure is shown in Fig. 4. This method achieved an 84.1% accuracy on the CUB-200-2011 dataset. B-CNN has two pipelines in structure, but the feature-stream and location-stream constitute the same feature expression, which extract feature information at the same scale. Therefore, the B-CNN is divided into the scope of single feature processing methods in our classification system.

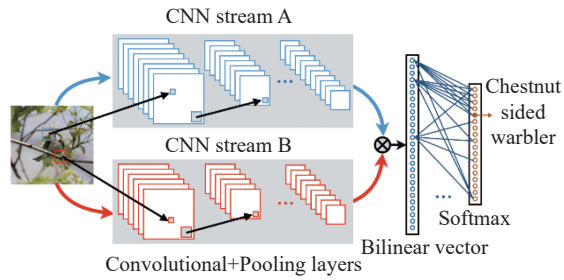


Fig. 4 Schematic diagram of bilinear CNN for image classification

A series of networks with this basic structure have been put forward one after another through continuous improvement and innovation by researchers. Considering the complementarity between features, hierachical bilinear pooling (HBP) [36] proposed an HBP framework based on the mutual promotion of the features of each layer, which integrates multiple cross-layer bilinear features. DBPNet-10 [37] proposed a deep bilinear transformation (DBT) module, which can group feature channels according to semantics and perform bilinear operations within the group, which enhances the expressiveness of the features. In addition to improving the accuracy, low-rank bilinear pooling (LRBP) [38] generated a low-rank approximation through singular value decomposition in order to reduce the degree of freedom of the classifier

parameters. The co-decomposition module was designed to separate the shared structure of the classifier to further reduce the parameters. On the CUB dataset, the accuracy of this method is similar to that of the original B-CNN, but the parameters are reduced from 200 M to 0.8 M.

In addition to the B-CNN series coding methods, there are many other feature coding methods. Similar to the idea of combining the position and feature information with the B-CNN, PartNet-Full [39] proposesd to use a two-stream structure to analyze the part type and effectiveness in the feature map. Based on low-order sparse coding technology, low-rank sparse coding with general and class-specific codebooks (LR-GCC) [40] designed general codebooks and class-specific codebooks that have learned images and it can combine the spatial and structural information when jointly encoding the local features. Based on the technology of autoencoders, group based deep shared feature learning networks (GSFL-Net) [41] decomposed image features into shared parts and distinguishable parts that are conducive to classification and uses the latter to construct the feature expression of images with strong classification ability. Instead of using the traditional CNN, transformer architecture for fine-grained recogition (TransFG) [42] extracted the features through the ViT network [6], and selected the discriminative patch for feature expression by using extra attention module. Similar to the traditional method of extracting the image features of different convolution layers in CNN, feature fusion tranformer (FFVT) [43] encoded the expression by integrating the features of each encoder layer based on the ViT network.

3.2.2 Feature fusion and enhancement methods

In terms of feature fusion and enhancement, graded-feature multilabel-learning network (GMNet) [44] enhanced the semantics by clustering the part features and used the Gaussian mixture method to fit a multimodal distribution with a linear combination of feature cluster centers, in order to obtain a feature expression with key information. In contrast, Cross-X [45] proposed a cross-layer fusion method based on the information correlation of different convolutional layers, and used a one-squeeze multi-excitation (OSME) module to generate multi-attention feature and a cross-semantic regularizer to group maps and fuse attention block features with similar semantics. Among them, the OSME module is a module proposed by [46] to capture different positions or different attention feature maps and is used in the network of multiline feature processing. Similarly, considering the diversity of feature map information in different convolutional layers, localizing object parts (LOP) [47] performed spectral clustering on the information of multiple convolutional layers to enhance the features and obtains more accurate and effective objects. Based on image feature interaction, context-aware attentional pooling (CAP) [48] proposed to calculate the influence weight between different features to enhance the expression ability of the individual feature. In addition to the technology of processing a single image, attentive pairwise interaction network (API-Net) [49] proposed to train the network in units of image pairs and use the pairwise interaction of attention to strengthen the distinction of features. Channel interation networks (CIN) [50] considered the relationship between the image self-channels and the differences between different image features and the proposed two modules, i.e., the self-channel interaction module and the contrast channel interaction module, to merge the image features. Graph-propagation based correlation learning (GCL) [51] designed vertical and horizontal cross graph propagation modules to aggregate feature maps into several nodes with a large amount of effective information and introduces graph convolutional network for feature interaction and enhancement. In addition to studying how to integrate the information more effectively, diversification block and gradient-boosting cross entropy (DB-GCE) loss [52] also proposed to train the network by continuously suppressing the activation area of the feature map so that the model can obtain more discriminative information for fusion. In order to understand the ability of the network to extract key features intuitively, counterfactual attention learning (CAL) [53] proposed to use counterfactual attention to quantify the quality of the attention, which effectively improves the expression effect of the features.

The above methods all directly group the extracted features, while multi-channel loss (MC-Loss) [54] proposed a channel loss so that the different feature channels extracted by CNN have the ability to distinguish different classes. This channel loss is composed of discriminative parts and diversity parts. The former constrains the feature channels to focus on specific categories and requires features to be sufficiently discriminative, while the latter requires the feature channels of the same category to focus on different areas of the image. With this channel loss, many networks improve the accuracy. The best result is the method of using this loss function on B-CNN, which achieves the accuracy of 94.4% on the Stanford Cars dataset.

3.3 Methods based on multiple feature processing

In contrast to the single feature processing networks, methods based on multiple feature processing are composed of multiple individual subnetworks. The network always processes multi-level features and constructs multiple feature expressions for separate classification. Generally, these methods are complicated but effective. According to different structures and feature processing

methods, it can be divided into the tree structure network methods, the cyclic attention methods, and the multi-stream network methods. Fig. 5 shows the classification of the first two methods.

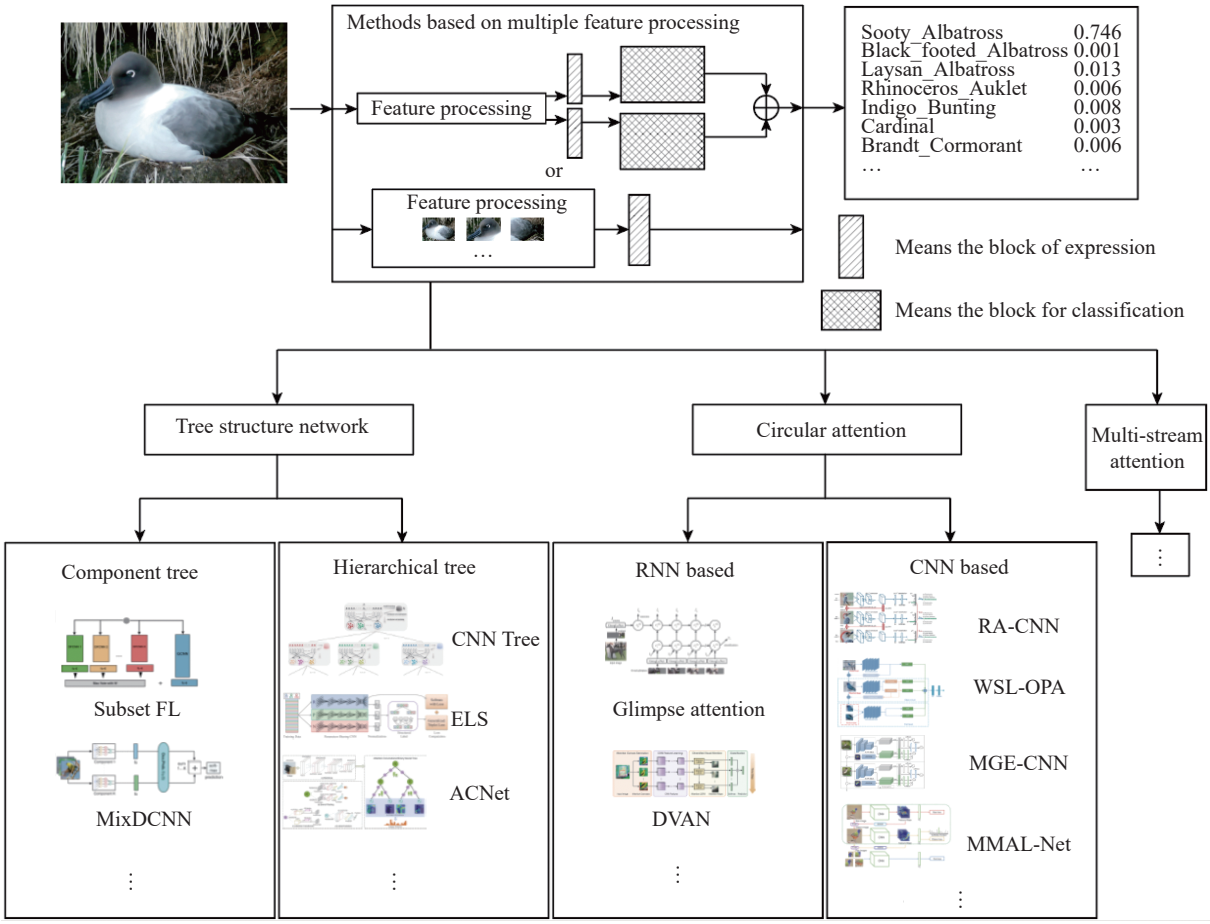


Fig. 5 Methods based on multiple feature processing

3.3.1 Tree structure network methods

The tree structure network methods refer to that the networks use a tree-like structure, which can be roughly divided into two subcategories: the component tree and the hierarchical tree.

The component tree networks are composed of multiple subnetworks, where each subnetwork is a component to extract and process one or several fine-grained features. A typical method is subset feature learning (Subset FL) proposed in [55]. The author separately trains the corresponding CNN to classify each group of images obtained by pre-training clustering to obtain features with the ability to distinguish similar images. In addition, this method also trains a subset selector to select the corresponding component CNN for the image. The advantage of this type of method is that the classification of similar images is more targeted, but the overall network is more complex and the accuracy is not sufficient. Based on this method, mixture of deep CNN (MixDCNN) [56] only uses component features and additionally designs the occupancy probability to obtain the probability weight of each component CNN feature to construct the final feature representation.

The hierarchical tree refers to the network filtering and processing the data layer by layer according to the tree structure while processing the data. CNN Tree [57] proposed a tree-shaped hierarchical structure to process the image layer by layer, where each intermediate node is a specific CNN of the confusion set obtained when the basic CNN is used for tentative classification. By continuously classifying images in the confusion set, images can be classified with a high accuracy in a targeted manner. The tree structure can then bring a more targeted discrimination ability and a lower training difficulty, which is meaningful for fine-grained visual categorization. Embedding label structures (ELS) [58] proposed an embedded tag structure, which divides tags into four-element categories in a tree-like hierarchy based on semantics or domain knowledge, in order to search for classification boundaries. In addition, attention convolutional binary neural tree network (ACNet) [59] proposed an attention convolutional binary neural tree model, which extracts different attention from the image layer by layer

and obtains the final result by counting the prediction results of all the leaf nodes and weights stored in the pathfinding module prediction.

3.3.2 Circular attention methods

Among the multiple feature processing methods, circular attention methods are also popular. This means that the entire framework uses the same or similar network structure cyclically with weakly supervised attention technology to extract the features of patches in different regions or different scales in the images. Among them, the most substantial difference between each method lies in the structure of the backbone network and the method to select image patches of different regions or scales. Specifically, it can be divided into two types: the networks using recurrent neural network (RNN) and the networks not using RNN.

Among the methods of the network using the RNN, the first is Glimpse attention, a method of attention RNN proposed in [60]. This method inputs the attention coordinates and image block features of different scales based on the current coordinates in each layer of the network, outputs the next attention coordinates, and additionally outputs the classification results in the last layer. This is one of the few models that performed unsupervised learning in 2015 and achieved 76.8% results on the Stanford Dogs dataset. This type of method cleverly uses the characteristics of information accumulation and intermediate output of the RNN and enhances the model's ability to understand images. After that Zhao et al. proposed a diversified visual attention network (DVAN) [61], which inputs the features of image blocks of different scales into each layer of attention long short-term memory (LSTM), obtains the local features that should be noted in each feature map, and then performs the process according to each feature classification.

Among these networks, some networks do not use RNN but they are with a cyclic structure. One of the earlier typical methods is RA-CNN proposed by [5], which is stepwise detailed. This method constructs an attention proposal network module, takes the feature map as the input, and outputs the coordinate and length data to indicate the next image block of interest. The structure of the network is shown in Fig. 6. Through the cyclic action of this special module, the entire model can gradually increase the intensity of the attention to the input image. Therefore, more details are discovered for fine-grained visual categorization. To solve the number limitation of concerned parts, weakly supervised learning of object-part attention (WSL-OPA) model [62] proposed to use the feature map channel summation method. Through the effect of the squeeze-and-excitation (SE) module and spatial part constraints, this model analyzes multiple different part blocks to obtain the features. Following the idea of gradual learning, mixture of granularity-specific experts CNN (MGE-CNN) [63] proposed to add constraints based on upper-level network knowledge to make the attention parts different and increase the effect of network learning. The latest achievement of the similar structure is multi-branch and multi-scale attention learning network (MMAL-Net) [64], which designs an object attention positioning module and a part attention proposal module to acquire the whole object patch and the multiple part patches.

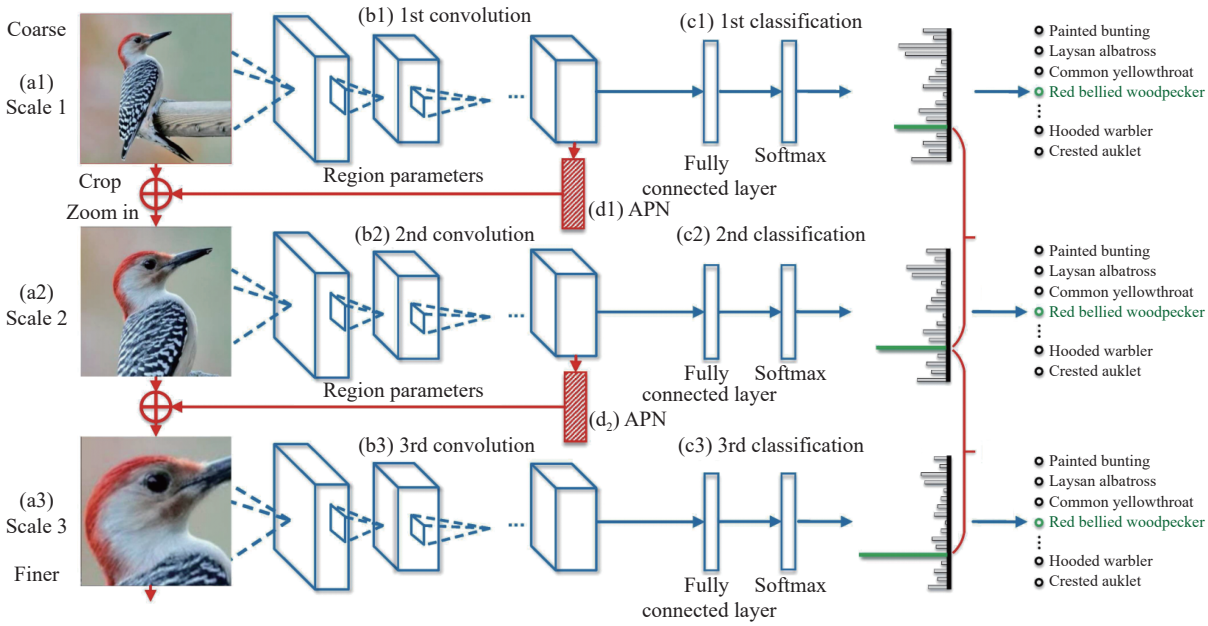


Fig. 6 Schematic diagram of the structure of the RA-CNN network

3.3.3 Multi-stream attention

The multi-stream attention methods emphasize the use of multiple pipelines in the network to process the multi-level features. The overall structures of these methods are similar. Essentially, after the preprocessing of the images, the CNN is used to extract global features. Then multiple attention is used to obtain multi-level features, and finally the classification results are obtained according to the fusion feature expressions. Under this basic framework, researchers' innovative work can be classified into three categories: part discovery, structural adjustment, and information introduction. The classification is shown in Fig. 7.

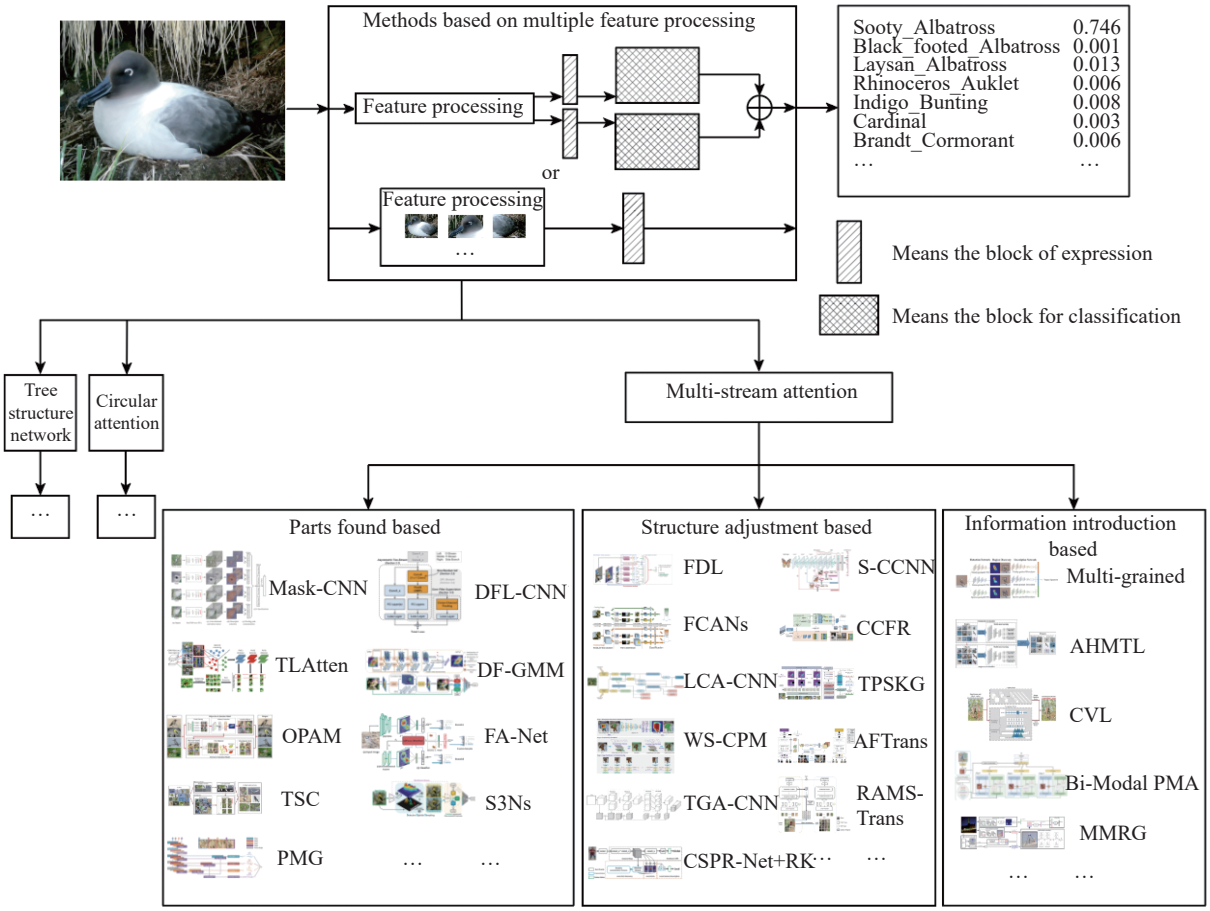


Fig. 7 Multi-stream attention methods based on multiple feature processing methods

(i) Methods based on parts found. In the work of part discovery, there have been many methods, from strong supervision to weak supervision. Among the strong supervision methods, the most representative one is the method proposed by Mask-CNN [65], which uses the part annotation point information of the image in the dataset to train a full convolutional network in order to obtain the part mask, obtain the object mask through the combination mask, and finally classify the image by extracting the features of the image block corresponding to the mask. The effect of the fully convolutional network is better in a strong supervision environment, but with the development of technology, weak supervision methods have received greater attention from researchers. The earlier method is the spectral clustering method proposed by two-level attention model (TL Atten) [66], which uses the hidden layer information of a trained global classification network to perform spectral clustering in order to obtain part blocks for the analysis of the second stream network. In contrast, object-part attention model (OPAM) [67] uses class activation map (CAM) technology to extract feature maps to obtain part boxes and object boxes and classifies them through object-level and part-level second-rate features. Compared with this method, the two spatical constraints (TSC) [68] method of the same paper also uses the saliency extraction method to obtain the part position, but the saliency value is obtained through a specially trained classification CNN network to reversely

calculate the category score derivative. After that, discriminative filter learning CNN (DFL-CNN) [69] proposed a method of using a filter library to obtain a heat map to obtain a discriminative image block. The structure of this method is shown in Fig. 5. In the process of using convolutional networks to obtain feature maps to locate parts, discriminative feature-oriented Gaussian mixture model (DF-GMM) [70] proposed a GMM to process feature maps with a low-rank expression mechanism in order to locate the parts more accurately. In addition to the technical research on attention discovery, there are also related studies on the distinction and restraint of attention. Reference [46] proposed using OSME to learn multiple attention features of an image and used a multi-attention and multi-class constraint (MAMC) module to guide and distinguish the attention features. Focus attention deep networks (FA-Net) [71] proposed the attention focus module, which obtains different parts by the method of gradual elimination and innovates from the perspective of multi-attention differentiation so that the model can distinguish the front and back scenes of the image while effectively preventing the regions that the model pays attention to coincide. In addition, selective sparse sampling networks (S3Ns) [72] distinguishes features into discriminative and complementary features by learning sparse attention and enhance their functions and characteristics. In terms of the granularity effectiveness of the parts, progressive multi-granularity (PMG) training of Jigsaw patche [73] forced the model to learn multi-granularity information through the form of multi-granularity image block puzzles. While classifying, it also obtained the details of each part that are truly effective for classification.

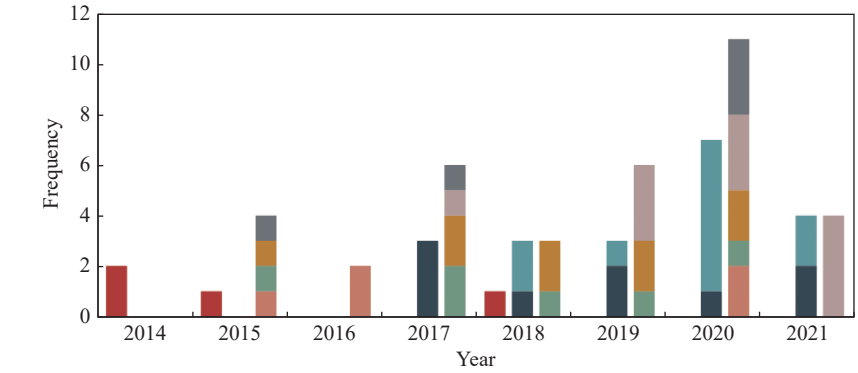
(ii) Methods based on structural adjustment. The work of structural adjustment mainly focuses on the overall planning of the network, and through the design of a more effective network structure, the features extracted from the image can be used more efficiently. In recent research, filtration and distillation learning (FDL) framework [74] added a region proposal filtering module and a knowledge transfer and distillation module to the traditional multi-stream network and made innovations in the optimization of region selection and the integration of feature semantics through structural adjustments. Fully convolutional attention networks (FCANs) [75] proposed stacking positioning parts with the convolutional layers and regarded the classification problem as a Markov decision process during the training process in order to supervise the network learning in more detail. However, learning cascade attention CNN (LCA-CNN) [76] considered the merging of pre-training models of different datasets, and proposed a cascaded attention learning model, which combines the pre-training information on the two datasets of ImageNet and iNaturalist, and designed a spatially constrained attention module, network fusion module and cross-network attention module to complete the classification task. Weakly supervised complernetary parts models (WS-CPM) [77] first designed object detection and instance segmentation tasks to process the images, then used Mask R-CNN and part constraint models to improve the quality of acquiring local image blocks, and then also used LSTM for feature analysis and classification. The concentrated statistical-positional-relational networks with reranking (CSPR-Net+RK) [78] adjusts the structure of the feature processing module. While acquiring global features, the design generates special statistics-location-relation descriptors to better describe the local features of parts and to generate special fine-grained part representation. In the overall structure, the hierachical CNN denoting skip-connections CNN (S-CCNN) [79] directly uses the feature maps of the first three levels of the convolutional layers and the last layer to combine and introduces a special module of “skip connection” for adjustment and classification. Two-level attentions and grouping attention CNN (TGA-CNN) [80] created a two-level attention model and a group attention model, which fuses the two-stream features of pixel-level and object-level attention and merged the features with high similarity after channel feature processing to express the semantics better. Coarse classification and fine re-ranking (CCFR) [81] designed a two stream network, in which the coarse stream uses global features to construct the top-N candidate result set, and the fine-grained stream uses discriminative regional features to fuse multi-scale information for the re-rank of the top-N results. Transformer with peak suppression and knowledge guidance (TPSKG) [82] proposed the peak suppression module and the knowledge guidance module based on ViT network. The former suppresses the most discriminative regions to pay more attention to the other informative parts for diverse expressions, and the latter guides the classification by constructing a learnable knowledge set. Adaptive attention multi-scale fusin transformer (AFTrans) [83] combined local and global features in multi-stream ViT network, and proposed a selective attention collection module to extract discriminative regions. Similarly, recurrent attention multi scale transformer (RAMS-Trans) [84] constructed a dynamic patch proposal module to adjust the image scale, merged the attention weight, and obtained the appropriate binary mask with the adaptive threshold for the fusion of multi-scale image features.

(iii) Methods based on information introduction. In the use of information, many methods use different granularities or different types of information. In terms of information granularity, multiple granalarity descriptors (Multi-grained) [85] proposed the concept of multi-grained descriptors as early as 2015 and designed three different granular CNN networks to extract image blocks with different degrees of attention to obtain diversified information. Correspondingly, attribute hierarchy based multi-task learning (AHMTL) [86] also proposed a model based on the attribute hierarchy. Based on the two attention levels of the original image and the object image, this is further divided into three classification tasks: coarse-grained, fine-grained, and attribute-based ultra-fine-grained to analyze the images from multiple angles. It is worth mentioning that the image blocks and features of different granularities can enhance the analysis and expression capabilities of the model and help improve the classification accuracy. In terms of the information types, the two-stream model combining vision and language (CVL) [87] constructed two information streams of vision and language for classification, where language information is a visual description learned from images using a text encoding model that contains a CNN-RNN structure. Additionally, using language flow, Bi-model progressive mask attention (Bi-Modal PMA) [88] proposed training a bimodal progressive mask attention model with the introduction of image description text and using knowledge distillation technology to apply the model to a monomodal environment. In addition, multi-model reasonging (MMR) [89] proposed a fine-grained classification method that combines text information in the images using image text recognition, bounding box feature coding and graph convolutional network reasoning and other technologies.

4. Method performance comparison and analysis

Various methods have tried to solve the problem of fine-grained visual categorization from different angles. On public datasets such as the CUB-200-2011 dataset, Stanford Cars dataset, and Aircraft dataset, these models all show the performance in different application contexts. Through the statistics and summary of the relevant data of the outstanding methods since 2014, we discover some laws of technological development in this field.

Fig. 8 shows the publication times of the methods under the classification of this paper. It can be found that with an increase of time, research in the entire field is increasingly based on multiple feature processing methods. According to the analysis, these methods generally have a greater advantage in multi-level features, which always use various fusion modules to synthesize multiple feature expressions, accompanied by complex structures. Common shortcomings include the large demand for computing power and the long training time. At the same time, the methods based on single feature processing are also important in recent years, which often have unique innovations in the use of feature information. By designing special modules to excavate the discriminative features of the image, these methods still have the competitive accuracy.



: Methods based on general network and strong supervision of components;
: Feature coding methods based on single feature processing;
: Feature fusion and enhancement methods based on single feature processing;
: Tree structure network methods based on multiple feature processing;
: Circular attention methods based on multiple feature processing;
: Parts found based methods of multi-stream attention based on multiple feature processing;
: Structure adjustment based methods of multi-stream attention based on multiple feature processing;
: Information introduction based methods of multi-stream attention based on multiple feature processing.

Fig. 8 Publication status of outstanding methods each year under the classification of this paper

Correspondingly, Fig. 9 illustrates the accuracy changes of the methods that perform the best on the CUB dataset for various models in the corresponding years. It can be observed that the earliest method is based on the general network and strong supervision of parts, but the recognition accuracy is not satisfactory. The methods based on multiple feature processing has maintained a high recognition accuracy since its appearance and has advantages in most years. The recognition accuracy of the methods based on single feature processing has been rising fiercely, and there is a hidden trend to catch up with the methods based on multiple feature processing. These two technologies actually promote one another. For example, when processing a certain stream of multi-stream features, the framework can be designed to use the model with the single feature processing technology. Some new feature extraction and processing methods are also inspired from the fusion expressions in the methods based on multiple feature processing. Therefore, from a development perspective the two kinds of methods in this field will continue to advance in a mutual stimulation and promote the resolution of fine-grained visual categorization problems.

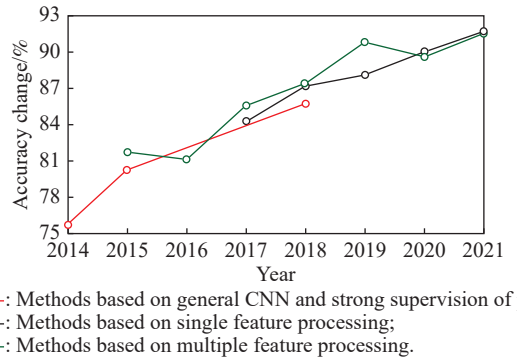


Fig. 9 The highest accuracy changes of different types of models in the CUB-200-2011 dataset

The current fine-grained visual categorization methods can already obtain high-precision results. From Fig. 9, it can be determined that the recognition accuracies of some methods in the CUB dataset exceeds 91.7%. Therefore, we show in Table 2 the backbone network and recognition accuracy of some typical networks with high recognition accuracy in this field. The "/" in the labeling situation means that labeling information other than the label is not required, while the "–" in the accuracy means that there is no experimental result on the corresponding dataset. Among them, the classification accuracy is defined as the average recognition accuracy.

Table 2 Recognition accuracy of some excellent methods on public datasets %

Method	Year	Backbone	Annotation	Accuracy		
				CUB	Cars	Air
S3Ns [72]	2019	Resnet-50	/	88.5	94.7	92.8
LCA-CNN [76]	2019	Inception V3	/	90.8	–	92.1
DB-GCE [52]	2020	ResNet-50	/	88.6	94.9	93.5
GCL [51]	2020	ResNet-50	/	88.3	94	93.2
Bi-Modal PMA [88]	2020	ResNet-50	Text	88.7	93.1	90.8
DF-GMM [70]	2020	ResNet-50	/	88.8	94.8	93.8
LDOP [47]	2020	ResNet-50	/	88.9	94.2	92.3
FDL [74]	2020	DenseNet161	/	89.1	94.0	91.3
MMAL-Net [64]	2020	ResNet-50	/	89.6	95	94.7
PMG [73]	2020	ResNet-50	/	89.6	95.1	93.4
API-Net [49]	2020	DenseNet-161	/	90	95.3	93.9
CAP [48]	2021	ResNet-50	/	–	94.9	94.9
CAL [53]	2021	ResNet-101	/	90.6	95.5	94.2
CCFR [81]	2021	ResNet-50	/	91.1	95.5	94.1
AFTrans [83]	2021	ViT	/	91.5	95.0	–
TransFG [42]	2021	ViT	/	91.7	94.8	–

Table 2 shows that the accuracy rankings of these methods on three different datasets are not consistent. For example, the recognition results of the LCA-CNN network with the highest recognition accuracy on the CUB dataset on the Air dataset are not ideal. This shows that the performance of many models in this field fluctuates due to the characteristics of the dataset. In addition, it is easy to find that the backbone networks used by the models in the table have almost all become the ResNet, DenseNet, and ViT series of networks. These deep learning networks that use a new generation structure have stronger feature extraction capabilities and the ability to capture image details, which can then meet the high demand of fine-grained task for feature information. Another noteworthy point is that most of the models in the table do not rely on the additional annotation information of the dataset. The research focus of the entire field has completed the transition from strong supervision to weak supervision, which also shows that the technology in this field has made a leap forward in practical use and production. We analyze some typical methods with high recognition accuracy in the following.

LCA-CNN [76] is a multiple feature processing method proposed in 2019, whose structure is complicated. The general idea of this method is to design a merged network based on the preprocessing parameters of the two datasets and combine the two pre-training backbone networks to reinforce one another to complete the fine-grained classification task. This model intro-

duces a variety of obfuscated attention modules and losses for feature fusion and processing, which combines the advantages of the two reference network streams. It has strong versatility from two datasets. On the CUB-200-2011 dataset, LCA-CNN has a classification accuracy of 90.8%. About the shortcomings, this model has a too large network structure with long running time.

API-Net [49] is an effective model based on single feature processing. It does not divide image features into different granularities for the analysis and does not obtain the final result through multiple feature distributions. The network enhances the feature extraction for single image by capturing the semantic differences and contrast cues of the paired images during the training process, which makes the network more sensitive to the detailed features when distinguishing between fine-grained images. This kind of macro-contrast processing makes full use of the learning ability of the network itself, and there is no need to artificially design the network learning process by adding a large number of network modules. As a method proposed in 2020, it has a 95.3% classification result on the Stanford Cars dataset.

CAP [48] is a single feature processing method proposed in 2021. The main innovation is the context-aware attention pooling module. Firstly, the module extracts image feature blocks of different scales to complete comprehensive context awareness. Then an attention module is designed to make all feature blocks interact with each other by calculating the weight by product, which can enrich the expression of features. Finally, the step-by-step input characteristics of LSTM network are used to complete feature fusion instead of simple connection. This resnet-50 based network achieves the best result of 94.9% on the aircraft dataset.

As a multi-stream method based on multiple feature processing, CCFR [81] is not limited to using local and global features at the same time, but creatively proposes the classification mode of top-N re-ranking. Specifically, the network constructs a hierarchical classification system, which extracts global features from multiple perspectives of superclasses and subclasses for coarse-grained classification, and constructs a top-N result set containing correct classification results. After that, the model obtains fine-grained image information through multi-scale feature extraction and fusion technology to guide re-rank. This method achieves 95.5% results on the Stanford Cars dataset.

The TransFG [42] network with the best result of 91.7% on the CUB-200-2011 dataset is a single feature processing method, which used the most popular ViT as the backbone network for feature extraction. The ViT has a good effect on the extraction of global information, and also makes the features of each patch contain the information of global context. Therefore, TransFG network selects the most effective component for classification through the method of maximum activation to construct the final feature expression, classification token. In addition, the comparative feature loss calculated by the similarity between different visual categorization tokens also further improves the performance of the model.

By analyzing the above methods, it can be determined that these methods have a good performance in terms of the acquisition and utilization of the effective discriminative information in the images. The used modules of these methods can innovatively solve the problems that other methods do not care, such as part complementarity, image pair differentiation, and information granularity. In general, the current fine-grained visual categorization research has achieved good results in a continuous progress, and the development trend is more obviously inclined to the method of multi-feature processing. On the current commonly used public datasets, the research direction of feature discovery based on image pairs and part differences is more popular than others. At the same time, models with complex structures still have a great advantage in classification accuracy. In this research field, the technology circulation mode will not change that the new technology module is first studied in the single feature processing method, and then combined in the multiple feature processing network. In terms of information introduction, the main research direction is still in the bimodal fusion of language and vision. How to make good use of the advantages of language processing methods to assist in solving the semantic problem of fine-grained visual categorization remains to be solved.

In terms of backbone network, it is worth mentioning that the self-attention network ViT, as an emerging and popular network, has an excellent effect in feature extraction and analysis, which was used frequently in 2021. In a large number of experiments, researchers have proved that the ViT has the ability to replace CNN in visual tasks, which also promotes the wide application of ViT. It can be predicted that in the future research, the models using ViT will continue to increase.

In fact, the core of the research in this field is how to efficiently extract and use the information in the image for fine-grained classification. Therefore, there are still many problems that can be studied in the field of fine-grained visual classification, such as the extraction of multi-scale and multiple granularity feature information or the analysis of multiple pattern information.

5. Future research ideas

5.1 Future research direction

In recent years, fine-grained visual classification technology has developed rapidly. As a visual task close to practical application, it will still be one of the research focuses in the future. The following contents include some views and research directions, which may inspire new models.

5.1.1 Improvement of self-attention network ViT

After the self-attention network achieved outstanding results in the field of natural language processing (NLP), a similar model in CV, ViT, has become a more popular backbone network than CNN. When it was initially applied in fine-grained visual categorization task, ViT achieved excellent results, which was similar to the results of the advanced CNN model. Therefore, it is easy to find that ViT will become an important direction of future research. Furthermore, there are many aspects that can be studied around ViT, including three main ideas. One is to build a similar ViT model to the network structure based on CNN in fine-grained visual categorization. The second is to build a new fine-grained feature extraction structure with ViT based on its own self-attention mechanism. In addition, due to the complementary relationship between CNN and ViT networks in the extraction of local features and global information, it can be speculated that developing a fusion network with both advantages is also a valuable research direction.

5.1.2 Discriminative features and attention mechanism

In the field of fine-grained visual categorization, deep learning networks mostly encode the key information in the whole image into a representation in the process of feature extraction. From the perspective of feature granularity, this representation is generally the fusion of two parts of feature information: the label-level and the fine-grained level. The latter often plays a more important role in the classification process, which is called discriminative features. In recent studies, some researchers use the attention mechanism to strengthen the discriminative features in the overall features, and obtain competitive results. This shows the effectiveness of attention mechanism in this field. Recently, there are many research in this field, including channel attention, spatial attention, mixed attention, and self-attention. However, by observing the publication frequency of relevant research results, it can be found that a variety of attention mechanisms are still being developed, which means that the research in this field is not mature and saturated. Therefore, mining discriminative features by using attention mechanism will still be a hot research direction in fine-grained visual categorization.

5.1.3 Introduction of background semantics and graph neural network (GNN)

Because the fine-grained images under the same label class do not have obvious differences, the existing technologies focus on extracting the discriminative fine-grained features in the image. However, some researchers proposed that the background semantics of images can also play an auxiliary role in fine-grained visual categorization. From the perspective of information association, the context of the image can provide different feature information outside the appearance of the subject object, which helps to overcome the lack of discriminative features between fine-grained classes with similar appearance and improve the accuracy of classification. Following this idea, it will become a potential research direction to introduce prior knowledge and GNN into the field of fine-grained visual categorization to explicitly model image background semantics and subject objects. In this way, the ability to mine effective classification information of image context can be improved, and the distinctive feature representation can be constructed.

5.2 Summary of problems and possible solutions

Fine-grained visual categorization is a challenging task. The existing methods based on deep learning have achieved many satisfactory results, but some problems are still difficult to solve. Among them, the two most difficult problems are complex datasets and insufficient samples.

5.2.1 Complex dataset

The existing fine-grained image datasets generally contain two types: datasets with a single label-level category and datasets with different label-level categories. For these datasets, the common problem is that the gap between classes of fine-grained images is small and the difference within classes is large. This requires that the network designed by researchers can effectively extract discriminative features. In addition, in the datasets with different label-level categories, the network needs to complete the visual categorization in the complex category domain including different levels of labels, which greatly improves the difficulty of classification.

For the extraction of discriminative features, most of the existing methods use attention mechanism to solve the problem, which helps to enhance the effective features of classification and make the representation

obtained by coding easy to be classified. The commonly used attention mechanism is mixed attention, which is the integration of channel and spatial attention. For the datasets with different label-level categories, one of the popular solutions in recent years is to use multi-scale features. Multi-scale features can be extracted from feature maps of different scales. In terms of the feature granularity, this integrates coarse-grained and fine-grained features. Therefore, the model can not only use the coarse-grained features to distinguish the label classes, but also use the fine-grained features to classify the subcategories.

5.2.2 Insufficient samples

Because it is more difficult to capture the fine-grained images than obtaining other visual data, and the data annotation needs much professional knowledge, most datasets of fine-grained visual categorization task have the problem of insufficient samples. This is particularly deadly for data-driven deep learning. In the case of insufficient training data, the network cannot be fully trained, so it is easy for the model to be over fitting and losing generalization.

For the problem of insufficient samples, the common method is data augmentation. In order to increase the amount of data, the conventional technology uses preprocessing methods such as random clipping, discoloration and stretching. In addition, some studies propose to use the generative adversarial network (GAN) to generate new image sample data to expand the original datasets. Considering from the perspective of network structure, introducing the Siamese network structure in the few-shot image classification task to fine-grained visual categorization task is also a research direction to solve the problem. In addition, using comparative learning to pretrain the model in a large dataset in advance can obtain model parameters with good generalization and portability. On this basis, the generalization of the network can be maintained even if only a few samples are used for fine-tuning.

5.3 Competition and application

5.3.1 Competition

With the development of fine-grained visual categorization technology, many competitions in this field also appear gradually. The most popular competition in recent years is FGVC series competitions [90−95]. In 2018, FGVC5 competitions were held with the main support of CVPR2018, including iMaterialist Challenge (Fashion)，iMaterialist Challenge (Furniture), and iNaturalist Challenge. In 2019, the number of competitions held in FGVC6 increased significantly, with a total of more than 10 competitions, including clothing, retail goods, animals, plants, food, collectibles, butterflies, moths, and cassava leaf diseases. As for FGVC7 in 2020, the research scope of the competition was extended to a wide range of plant pathology and semi-supervised labeled animal and plant datasets. In 2021, FGVC8 launched the Hotel-ID to Combat Human Trafficking competition, which requires to identify hotel IDs through the detailed style differences in the scene images inside the hotel.

From the results of the plant pathology competition of FGVC7 in 2020 and FGVC8 in 2021, the two teams that won the first place used the method based on ResNet. The former uses the se-resnext-50 network and achieves 98.4% on the plant pathology 2020 dataset with 3 651 images. And the latter combines ResNet and ResNext50 networks and achieves 88.3% accuracy on the plant pathology 2021 dataset with approximately 23 000 high quality RGB images. In addition to fine-tuning the network parameters, the contestants also make many attempts in data augmentation, which effectively improves the effect of network training.

Objectively speaking, FGVC competitions can show the characteristics of various network models in practical application, which helps researchers in this field to consider more practical factors in research, and also promotes the expansion of fine-grained visual categorization technology in practical application.

5.3.2 Application

Fine-grained visual categorization is a valuable technology, which can effectively improve the application mode of existing visual categorization technology. In civil use, fine-grained visual categorization technology can help users to obtain the object information in the image in real time, and provide more detailed content than ordinary visual categorization technology. For example, in the case of animals and plants, fine-grained classification technology can obtain accurate information of the objects specific to families, genera, and even species from the images, which can improve the practicability and the user experience. Because the military field has high requirements for the richness and accuracy of information, the advantages of fine-grained visual categorization technology are more obvious. Compared with manual judgment and ordinary visual categorization algorithms, fine-grained visual categorization technology can provide results with faster speed and more detailed classification degree. Compared with manual judgment and ordinary visual categorization algorithms, fine-grained visual categorization technology can provide results with faster speed and more detailed classification degree, which

makes it have important application value both in personnel classification task in complex battlefield environment and acquisition task of enemy equipment intelligence. For example, the fine-grained visual categorization technology can classify a large number of enemy ships of images accurately when the maritime conflict occurs. It is conducive to the combatants to quickly obtain the ship type, weapons, and other detailed information, improve the army's information superiority on the battlefield and enhance the combat effectiveness.

6. Conclusions

This paper summarizes some progress in the field of fine-grained visual categorization in recent years and innovatively classifies these methods according to the types of features used in the model, the differences about the feature processing, and the differences in the structure of the networks. According to this principle, we divide these methods into three categories: methods based on general CNN and strong supervision of parts, methods based on single feature processing, and methods based on multiple feature processing. After analyzing the networks with outstanding recognition accuracy, the reasons for their success can be attributed to the two aspects of the effective information extraction of the image and the rational network structure. Finally, we propose to increase the research on various feature processing methods with the ViT, and improve the robustness for more complex datasets in the future research.

References

[1] WELINDER P, BRANSON S, MITA T, et al. Caltech-UCSD Birds 200. Pasadena: California Institute of Technology, 2010: CNS–TR-2010-001.

[2] YAO B P, BRADSKI G, LI F F. A codebook-free and annotation-free approach for fine-grained image categorization. Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2012: 3466–3473.

[3] ZHANG N, DONAHUE J, GIRSHICK R, et al. Part-based R-CNNs for fine-grained category detection. Proc. of the European Conference on Computer Vision, 2014: 834–849.

[4] LIN T Y, ROYCHOWDHURY A, MAJI S. Bilinear convolutional neural networks for fine-grained visual recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2018, 40(6): 1309–1322.

[5] FU J L, ZHENG H L, MEI T. Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4476–4484.

[6] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale. https://arxiv.org/abs/2010.11929.

[7] NILSBACK M E, ZISSERMAN A. Automated flower classification over a large number of classes. Proc. of the 6th Indian Conference on Computer Vision, Graphics & Image Processing, 2008: 722–729.

[8] WAH C , BRANSON S, WELINDER P, et al. The Caltech-UCSD Birds-200-2011 Dataset. Pasadena: California Institute of Technology, 2011: CNS–TR-2011-001.

[9] KHOSLA A, JAYADEVAPRAKASH N, YAO B P, et al. Novel dataset for fine-grained image categorization. https://people.csail.mit.edu/khosla/papers/fgvc2011.pdf.

[10] KRAUSE J, STARK M, DENG J, et al. 3D Object representations for fine-grained categorization. Proc. of the 4th International IEEE Workshop on 3D Representation and Recognition, 2013: 554–561.

[11] MAJI S, RAHTU E, KANNALA J, et al. Fine-grained visual classification of aircraft. https://arxiv.org/abs/1306.5151.

[12] HOU S H, FENG Y S, WANG Z L. VegFru: a domain-specific dataset for fine-grained visual categorization. Proc. of the IEEE International Conference on Computer Vision, 2017: 541–549.

[13] HORN G, AODHA O, SONG Y, et al. The iNaturalist species classification and detection dataset. Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 8769–8778.

[14] WEI X S, CUI Q, YANG L, et al. RPC: a large-scale retail product checkout dataset. https://arxiv.org/abs/1901.07249.

[15] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278–2324.

[16] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84–90.

[17] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition. https://arxiv.org/abs/1409.1556.

[18] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2818–2826.

[19] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770–778.

[20] HUANG G, LIU Z, MAATEN L, et al. Densely connected convolutional networks. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2261–2269.

[21] HU J, SHEN L, SUN G. Squeeze-and-excitation networks. Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 7132–7141.

[22] SRIVASTAVA R K, GREFF K, SCHMIDHUBER J. Highway networks. https://arxiv.org/abs/1505.00387.

[23] LARSSON G, MAIRE M, SHAKHNAROVICH G. FractalNet: ultra-deep neural networks without residuals. https://arxiv.org/abs/1605.07618.

[24] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 580–587.

[25] GIRSHICK R. Fast R-CNN. Proc. of the IEEE International Conference on Computer Vision, 2015: 1440–1448.

[26] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. on Pattern Analysis and Machine Intelli-

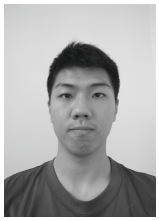
gence, 2017, 39(6): 1137–1149.
[27] HE X T, PENG Y X, ZHAO J J. Fast fine-grained image classification via weakly supervised discriminative localization. IEEE Trans. on Circuits and Systems for Video Technology, 2018, 29(5): 1394–1407.
[28] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multiBox detector. Proc. of the European Conference on Computer Vision, 2016: 21–37.
[29] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779–788.
[30] REDMON J, FARHADI A. YOLO9000: better, faster, stronger. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6517–6525.
[31] REDMON J, FARHADI A. YOLOv3: an incremental improvement. https://arxiv.org/abs/1804.02767.
[32] BOCHKOVSKIY A, WANG C Y, LIAO H Y. YOLOv4: optimal speed and accuracy of object detection. https://arxiv.org/abs/2004.10934.
[33] YAN Z X, HOU Z Q, XIONG L, et al. Fine-grained image classification based on the fusion of YOLOv3 and bilinear feature. Journal of Image and Graphics, 2021, 26(4): 847–856. (in Chinese)
[34] BRANSON S, VAN HORN G, BELONGIE S, et al. Bird species categorization using pose normalized deep convolutional nets. https://arxiv.org/abs/1406.5952.
[35] LIN D, SHEN X Y, LU C W, et al. Deep LAC: deep localization, alignment and classification for fine-grained recognition. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1666–1674.
[36] YU C J, ZHAO X Y, ZHENG Q, et al. Hierarchical bilinear pooling for fine-grained visual recognition. Proc. of the European Conference on Computer Vision, 2018: 595–610.
[37] ZHENG H L, FU J L, ZHA Z J, et al. Learning deep bilinear transformation for fine-grained image representation. Proc. of the 33rd International Conference on Neural Information Processing Systems, 2019, 385: 4277–4286.
[38] KONG S, FOWLKES C. Low-rank bilinear pooling for fine-grained classification. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 7025–7034.
[39] ZHANG Y B, JIA K, WANG Z X, et al. Part-aware fine-grained object categorization using weakly supervised part detection network. IEEE Trans. on Multimedia, 2020, 22(5): 1345–1357.
[40] ZHANG C J, LIANG C, LI L, et al. Fine-grained image classification via low-rank sparse coding with general and class-specific codebooks. IEEE Trans. on Neural Networks and Learning Systems, 2017, 28(7): 1550–1559.
[41] LI X L, MONGA V. Group based deep shared feature learning for fine-grained image classification. https://arxiv.org/abs/2004.01817.
[42] HU J, CHEN J N, LIU S, et al. TransFG: a transformer architecture for fine-grained recognition. https://arxiv.org/abs/2103.07976.
[43] WANG J, YU X H, GAO Y S. Feature fusion vision transformer for fine-grained visual categorization. https://arxiv.org/abs/2107.02341.
[44] LIANG J Y, GUO J L, LIU X, et al. Fine-grained image classification with Gaussian mixture layer. IEEE Access, 2018, 6: 53356–53367.
[45] LUO W, YANG X T, MO X J, et al. Cross-X learning for fine-grained visual categorization. Proc. of the IEEE/CVF International Conference on Computer Vision, 2019: 8241–8250.
[46] MING S, YUAN Y C, ZHOU F, et al. Multi-attention multi-class constraint for fine-grained image recognition. Proc. of the European Conference on Computer Vision, 2018: 834–850.
[47] ZHENG X T, QI L, REN Y T, et al. Fine-grained visual categorization by localizing object parts with single image. IEEE Trans. on Multimedia, 2020, 23: 1187–1199.
[48] BEHERA A, WHARTON Z, HEWAGE P, et al. Context-aware attentional pooling (CAP) for fine-grained visual classification. https://arxiv.org/abs/2101.06635.
[49] ZHUANG P Q, WANG Y L, QIAO Y. Learning attentive pairwise interaction for fine-grained classification. Proc. of the AAAI Conference on Artificial Intelligence, 2020: 13130–13137.
[50] GAO Y, HAN X T, WANG X, et al. Channel interaction networks for fine-grained image categorization. Proc. of the AAAI Conference on Artificial Intelligence, 2020: 10818–10825.
[51] WANG Z H, WANG S J, LI H J, et al. Graph-propagation based correlation learning for weakly supervised fine-grained image classification. Proc. of the AAAI Conference on Artificial Intelligence, 2020: 12289–12296.
[52] SUN G L, CHOLAKKAL H, KHAN S, et al. Fine-grained recognition: accounting for subtle differences between similar classes. Proc. of the AAAI Conference on Artificial Intelligence, 2020: 12047–12054.
[53] RAO Y M, CHEN G Y, LU J W, et al. Counterfactual attention learning for fine-grained visual categorization and re-identification. Proc. of the IEEE/CVF International Conference on Computer Vision, 2021: 1025–1034.
[54] CHANG D L, DING Y F, XIE J Y, et al. The devil is in the channels: mutual-channel loss for fine-grained image classification. IEEE Trans. on Image Processing, 2020, 29: 4683–4695.
[55] GE Z Y, MCCOOL C, SANDERSON C, et al. Subset feature learning for fine-grained category classification. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015: 46–52.
[56] GE Z Y, BEWLEY A, MCCOOL C, et al. Fine-grained classification via mixture of deep convolutional neural networks. Proc. of the IEEE Winter Conference on Applications of Computer Vision, 2016. DOI: 10.1109/WACV.2016.7477700.
[57] WANG Z H, WANG X X, WANG G. Learning fine-grained features via a CNN tree for large-scale classification. https://arxiv.org/abs/1511.04534.
[58] ZHANG X F, ZHOU F, LIN Y Q, et al. Embedding label structures for fine-grained feature representation. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1114–1123.
[59] JI R Y, WEN L Y, ZHANG L B, et al. Attention convolutional binary neural tree for fine-grained visual categorization. Proc. of the IEEE/CVF Conference on Computer Vision

and Pattern Recognition, 2020: 10465–10474.
[60] SERMANET P, FROME A, REAL E. Attention for fine-grained categorization. https://arxiv.org/abs/1412.7054.
[61] ZHAO B, WU X, FENG J S, et al. Diversified visual attention networks for fine-grained object classification. IEEE Trans. on Multimedia, 2017, 19(6): 1245–1256.
[62] LEI C X, JIANG L F, JI J S, et al. Weakly supervised learning of object-part attention model for fine-grained image classification. Proc. of the IEEE 18th International Conference on Communication Technology, 2018: 1222–1226.
[63] ZHANG L B, HUANG S L, LIU W, et al. Learning a mixture of granularity-specific experts for fine-grained categorization. Proc. of the IEEE/CVF International Conference on Computer Vision, 2019: 8331–8340.
[64] ZHANG F, LI M, ZHAI G S, et al. Multi-branch and multi-scale attention learning for fine-grained visual categorization. Proc. of the International Conference on Multimedia Modeling, 2021: 136–147.
[65] WEI X S, XIE C W, WU J X. Mask-CNN: localizing parts and selecting descriptors for fine-grained image recognition. https://arxiv.org/abs/1605.06878.
[66] XIAO T J, XU Y C, YANG K Y, et al. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 842–850.
[67] PENG Y X, HE X T, ZHAO J J. Object-part attention model for fine-grained image classification. IEEE Trans. on Image Processing, 2018, 27(3): 1487–1500.
[68] HE X T, PENG Y X. Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification. Proc. of the 31st AAAI Conference on Artificial Intelligence, 2017: 4075–4081.
[69] WANG Y M, MORARIU V, DAVIS L. Learning a discriminative filter bank within a CNN for fine-grained recognition. Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 4148–4157.
[70] WANG Z H, WANG S J, YANG S H, et al. Weakly supervised fine-grained image classification via Guassian mixture model oriented discriminative learning. Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 9746–9755.
[71] ZOU C, WANG R, CAO X C, et al. Weighted focus-attention deep network for fine-grained image classification. Proc. of the IEEE International Conference on Big Data, 2019: 5116–5125.
[72] DING Y, ZHOU Y Z, ZHU Y, et al. Selective sparse sampling for fine-grained image recognition. Proc. of the IEEE/CVF International Conference on Computer Vision, 2019: 6598–6607.
[73] DU R Y, CHANG D L, BHUNIA A K, et al. Fine-grained visual classification via progressive multi-granularity training of Jigsaw patches. Proc. of the European Conference on Computer Vision, 2020: 153–168.
[74] LIU C B, XIE H T, ZHA Z J, et al. Filtration and distillation: enhancing region attention for fine-grained visual categorization. Proc. of the AAAI Conference on Artificial Intelligence, 2020: 11555–11562.
[75] LIU X, XIA T, WANG J, et al. Fully convolutional attention networks for fine-grained recognition. https://arxiv.org/abs/1603.06765.
[76] ZHU Y X, LI R C, YANG Y, et al. Learning cascade attention for fine-grained image classification. Neural Networks, 2020, 122: 174–182.
[77] GE W F, LIN X R, YU Y Z. Weakly supervised complementary parts models for fine-grained image classification from the bottom up. https://arxiv.org/abs/1903.02827.
[78] WAN C Q, WU Y, TIAN X M, et al. Concentrated local part discovery with fine-grained part representation for person re-identification. IEEE Trans. on Multimedia, 2020, 22(6): 1605–1618.
[79] LIN Z Q, JIA J D, GAO W L, et al. Fine-grained visual categorization of butterfly specimens at sub-species level via a convolutional neural network with skip-connections. Neurocomputing, 2020, 384: 295–313.
[80] YANG Y D, WANG X F, ZHAO Q, et al. Two-level attentions and grouping attention convolutional network for fine-grained image classification. Applied Sciences, 2019, 9(9): 1939–1954.
[81] YANG S K, LIU S, YANG C, et al. Re-rank coarse classification with local region enhanced features for fine-grained image recognition. https://arxiv.org/abs/2102.09875.
[82] LIU X D, WANG L L, HAN X G. Transformer with peak suppression and knowledge guidance for fine-grained image recognition. Neurocomputing, 2022, 492: 137–149.
[83] ZHANG Y, CAO J, ZHANG L, et al. A free lunch from ViT: adaptive attention multi-scale fusion transformer for fine-grained visual recognition. Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2022: 3234–3238.
[84] HU Y Q, JIN X, ZHANG Y, et al. RAMS-Trans: recurrent attention multi-scale transformer for fine-grained image recognition. Proc. of the 29th ACM International Conference on Multimedia, 2021: 4239–4248.
[85] WANG D Q, SHEN Z Q, SHAO J, et al. Multiple granularity descriptors for fine-grained categorization. Proc. of the IEEE International Conference on Computer Vision, 2015: 2399–2406.
[86] ZHAO J J, PENG Y X, HE X T. Attribute hierarchy based multi-task learning for fine-grained image classification. Neurocomputing, 2020, 395: 150–159.
[87] HE X T, PENG Y X. Fine-grained image classification via combining vision and language. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 5994–6002.
[88] SONG K T, WEI X S, SHU X B, et al. Bi-modal progressive mask attention for fine-grained recognition. IEEE Trans. on Image Processing, 2020, 29: 7006–7018.
[89] MAFLA A, DEY S, BITEN A F, et al. Multi-modal reasoning graph for scene-text based fine-grained image classification and retrieval. Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021: 4023–4033.
[90] ZHANG C Y, KAESER C, VESOM G, et al. The iMet collection 2019 challenge dataset. https://arxiv.org/abs/1906.00901.
[91] BEERY S, VAN HORN G, MAC AODHA O, et al. The iWildCam 2018 challenge dataset. https://arxiv.org/abs/1904.05986.

[92] THAPA R, SNAVELY N, BELONGIE S, et al. The plant pathology 2020 Challenge Dataset to classify foliar disease of apples. https://arxiv.org/abs/2004.11958.
[93] SU J C, MAJI S. The semi-supervised iNaturalist-Aves challenge at FGVC7 workshop. https://arxiv.org/abs/2103.06937.
[94] BEERY S, AGARWAL A, COLE E, et al. The iWildCam 2021 competition dataset. https://arxiv.org/abs/2105.03494.
[95] SU J C, MAJI S. The semi-supervised iNaturalist challenge at the FGVC8 workshop. https://arxiv.org/abs/2106.01364.

Biographies

XIE Yuxiang was born in 1976. She received her B.S., M.S., and Ph.D. degrees from National University of Defense Technology in 1998, 2001, and 2004 respectively. She is a professor in the School of Information System and Management, National University of Defense Technology. Her research interests include computer vision and image and video analysis, classification, and retrieval.
E-mail: yxxie@nudt.edu.cn



GONG Quanzhi was born in 1998. He received his B.S. degree from 2020. He is pursuing his M.S. degree in National University of Defense Technology. His research interests include fine-grained image classification and action recognition.
E-mail: Charles_g27@qq.com

LUAN Xidao was born in 1976. He received his B.S. degree in applied mathematics in 1998, M.S. and Ph.D. degrees in systems engineering in 2005, 2009 respectively, from National University of Defense Technology. Now he is a professor in the School of Computer Engineering and Applied Mathematics, Changsha University. His research interests include computer vision and image and video analysis, classification, and retrieval.
E-mail: xidaoluan@ccsu.cn

YAN Jie was born in 1999. She received her B.S. degree from 2020. She is pursuing her M.S. degree in National University of Defense Technology. Her research interests include computer vision and image caption.
E-mail: yjierrr@163.com

ZHANG Jiahui was born in 1996. He received his B.S. degree from 2019. He is pursuing his M.S. degree in the National University of Defense Technology. His research interests include computer vision and deep learning.
E-mail: 100634004@qq.com