Received: 26 July 2024 | Revised: 15 November 2024 | Accepted: 22 November 2024

IET Image Processing

DOI: 10.1049/ipr2.13295

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

A fine-grained image classification method based on information interaction

Shuo Zhu[1] | Xukang Zhang[2] | Yu Wang[2] | Zongyang Wang[3] | Jiahao Sun[1]

[1]Jiangsu Province Engineering Research Center of Photonic Devices and System Integration for Communication Sensing Convergence, Wuxi University, Wuxi, China

[2]School of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing, China

[3]Wuxi Xiyuan Technology Co., Ltd., Wuxi, Jiangsu, China

Correspondence
Shuo Zhu, Jiangsu Province Engineering Research Center of Photonic Devices and System Integration for Communication Sensing Convergence, Wuxi University, Wuxi 214105, China.
Email: zshuo2011@163.com

Funding information
Jiangsu Province Double Innovation Talents Plan for Doctoral Candidates, Grant/Award Number: JSSCBS20210871; Wuxi Institute of Technology Talent Start-up Fund

Abstract

To enhance the accuracy of fine-grained image classification and address challenges such as excessive interference factors within the dataset, inadequate extraction of local key features, and insufficient channel semantic association, a dual-branch information interaction model that integrates convolutional neural networks (CNN) with Vision Transformers is proposed. This model leverages the Vision Transformer branch to extract global features, which are subsequently combined with the CNN branch to further augment the model's capability for local information extraction. In order to enhance the ability of the CNN branch to extract global information and reduce the loss of feature information, a feature enhancement module is added to the CNN branch. Since the Vision Transformer branch directly convolves with the convolution kernel will result in the inability to learn the underlying features of the image, a shallow feature extraction module is proposed, and the CNN and Vision Transformer branches interact with the information of the dual branches through the down-sampling Down module and the up-sampling UP module. The accuracy of the improved method on CUB-200-2011, Stanford Cars and FGVC-Aircraft fine-grained image classification datasets are 95.2%, 97.1% and 96.9%, respectively. The experimental results show that the method has good generalization on different datasets.

1 | INTRODUCTION

Fine-grained image classification refers to the task of classifying images with similar appearance but belonging to different fine-grained categories in the field of computer vision. Compared to traditional image classification tasks, fine-grained image classification requires models that are able to distinguish objects or scenes with small differences. A common challenge in fine-grained image classification is the presence of similar appearance features in the same category, such as different breeds of dogs or birds. These objects may have similar shapes and structures overall, but subtle visual feature differences may be contained in their local regions, such as texture, colour, or shape details. Consequently, a significant challenge for fine-grained image classification models lies in the comprehensive extraction of discriminative regional features of objects within images to enhance classification accuracy.

Convolutional neural networks (CNNs) have been dominant in computer vision, and CNN architectures have evolved to become more and more powerful, starting with AlexNet [1] and its revolutionary performance on the ImageNet image classification challenge.AlexNet was a major breakthrough for convolutional neural networks. Since then, several deep convolutional neural network models have been developed, including VGG [2], ResNet [3], GoogLeNet [4] and EfficientNet [5]. These models employ deeper network structures, more efficient convolutional kernels, and more efficient feature extraction methods, thus improving the accuracy of image classification [6]. Based on this, a large number of fine-grained networks based on CNN improvements have been proposed. For example, in 2018, the HBP method [7] proposed cross-layer bilinear pooling to capture some of the feature relationships between different layers, and used a hierarchical bilinear pooling framework to integrate multiple cross-layer bilinear features to enhance the feature representation. The HBPASM method [8] improves the classification performance by generating RoI-aware fine-grained feature representations, integrating multilayered masks to reduce the influence of

background and generates more reliable RoI features.Kong et al. [9] proposed to represent covariance features as matrices and apply a low-rank bilinear classifier. The resultant classifier can be evaluated without explicitly computing the bilinear feature mapping, which can significantly reduce the computation time and reduce the number of effective parameters to be learned. Chang [10] embedded the obfuscated channel spatial attention mechanism into the CNN network structure to enhance the feature extraction capability, while proposing a semantic data enhancement method to improve the accuracy of fine-grained model recognition, and utilizing a counterfactual attention network to improve the detection accuracy. Despite the great progress made in deep learning-based approaches for fine-grained image classification, the transformation operators used in CNN architectures impose restrictions on the receptive domain, thus hindering the network's ability to capture pixels at long distances from each other, which is a pressing issue in CNN architectures.

Transformer [11] was initially applied in the field of natural language processing (NLP), which relies on an attentional mechanism to model inputs and outputs. However, it cannot be directly applied to computer vision (CV) because CV involves two-dimensional image data for feature extraction and Transformer requires one-dimensional sequences for training. Vision Transformer (ViT) [12] was the first to apply Transformer to CV by dividing an image into multiple non-overlapping patches and feeding them into Transformer's encoder for training, which achieved satisfactory results in image classification. Based on this, a large number of improved methods based on ViT have been proposed. For example, the HVT method [13] proposes to divide the Transformer into stages and downsample the feature maps as the network goes deeper. However, due to the secondary complexity of the self-concerned modules, high resolution feature maps in the early stages result in high computational and memory costs. The PVT method [14] proposes a network suitable for intensive prediction tasks. Compared to ViT, PVT not only trains on dense partitions of images to achieve high output resolution, which is important for dense prediction, but also uses an asymptotically shrinking pyramid to reduce the computational effort of large feature maps. Meanwhile PVT inherits the advantages of CNN and Vision Transformer. The above methods achieve good results but ignore local feature details and reduce the discriminability between background and foreground.The ConViT method [15] utilizes the Transformer as its foundational framework, incorporating the GPSA attention mechanism and convolutional blocks, while dynamically executing convolution calculations based on gating parameters, thereby enhancing both the trainability and computational efficiency of the model. SADMix [16] introduces a data mixing enhancement strategy that leverages semantic and attention-based data mixing. Initially, images are asymmetrically mixed, with labels for the mixed images generated according to normalized Class Activation Maps (CAM). Key object regions are identified through a semantic and attention positioning module. The sizes and aspect ratios of these selected semantic-focused box regions are randomly altered to integrate more informative content. The TBAL-Net method [17] proposes a dual-branch attention learning network wherein each branch extracts global information and key region details from images; it incorporates a lightweight hybrid attention module to accurately locate key regions and learn fine-grained feature representations. A fine-grained contrastive instance learning (CIL) training framework is employed during testing to enhance the model's adaptability to varying environments.

In response to the problem that CNN only focuses on local features and ViT ignores local features, researchers have taken the combination of CNN and ViT as the main research trend. For example, the CeiT method [18] combines the advantages of convolutional neural networks (CNNs) in extracting low-level features and enhancing localization, and the advantages of ViT in establishing long-distance dependencies, to solve the problem of labelling directly from the original input image, and achieves good performance while reducing the training cost. The CvT method [19] introduces a convolutional labelling embedding that contains a new converter hierarchy, and a convolutional converter block that utilizes convolutional projection. These changes introduce beneficial properties of convolutional neural networks while maintaining the benefits of the Transformer. The LocalViT method [20] enhances the Vision Transformer (ViT) architecture by incorporating locality through the addition of multiple Transformer blocks, thereby enabling exploration of deeper architectures. ViT serves as a neural network model for image processing, while LocalViT facilitates the exchange of local features by integrating 2D deep convolution and nonlinear activation functions into the feedforward network of ViT, significantly augmenting its performance via the introduction of a localized mechanism.

To further enhance the accuracy of fine-grained image classification, preserve local features and global representations of images to the greatest extent, and improve the model's classification capability, this study proposes a dual-branch information interaction model that integrates convolutional neural networks (CNN) with Vision Transformers.

1. A multi-scale fusion attention mechanism is incorporated within the CNN branch, enabling the model to effectively capture critical information in images while enriching feature details. Additionally, a feature enhancement module is employed to augment the model's capacity for feature representation, thereby facilitating the capture of more salient information and improving overall classification accuracy.
2. A low-level feature extraction module is introduced at the front end of the Vision Transformer branch to ensure that the model can thoroughly identify and capture image details during its initial processing stage. Subsequently, a feature selection self-attention module enhances the model's focus on key features within images while mitigating background interference.
3. The integration of CNNs and Vision Transformers culminates in a dual-branch information interaction network that fuses local and global information across varying resolutions, significantly enhancing the model's ability to recognize fine-grained features.

2 | METHOD

CNN and ViT bilinear interactive network, CB-ViT, is a bilinear information interaction that combines CNN and ViT, and its purpose is to make full use of the advantages of CNN and ViT to improve the detection accuracy of fine-grained image classification. First the CB-ViT model uses the SFE module to obtain the low-level features of the input image, and then these features are fed into the ViT model for global feature extraction. Secondly the input image is entered into ResNet50 for feature extraction and the extracted information is input into the CNN module. In order to enhance the ability of the CNN module to extract global information and to reduce the loss of feature information, a feature enhancement module is added to the model. In the CB-ViT module, we continuously feed the global context of the ViT branch into the CNN to enhance the global sensing ability of the CNN. At the same time, the local features of the CNN branch are gradually fed into patch embedding to enrich the local details of the ViT branch. Such a process forms the information interaction. Since the feature dimensions of CNN and ViT are inconsistent in the process of interaction, the dimension of CNN feature mapping is C × H × W (C, H, and W are the channel, height, and width, respectively), and the shape of patch embedding in ViT is $(K+1) \times E$, where K, 1, and E denote the number of image patches, the class markers, and the embedding dimension, respectively. When fed to the ViT branch, the feature map first needs to be aligned to the number of channels of the patch embedding by 1 × 1 convolution. Then the spatial dimension alignment is done using the downsampling Down module. Finally, the feature maps are added to the patch embedding. When the global information obtained from the ViT branch is fed back to the CNN branch, UP upsampling of the patch embedding is required to align the spatial scale. The channel dimensions are aligned to the dimensions of the CNN feature map by 1 × 1 convolution.

2.1 | Vision transformer encoders

ViT works by first dividing the input image into fixed-size image patches, that is, patches. each patch is then converted into a one-dimensional vector by a linear mapping, which is usually achieved by spreading the pixel values. To introduce position information, the position of each patch is embedded into the vector representation using position encoding position embedding. The patch embedding is concatenated with the position encoding to form an input embedding matrix, which is passed as an input to the Transformer encoder. The Transformer encoder mainly consists of multi-head self attention (MSA) as well as multi-layer perceptron MLP to capture the global image information and the relationship between patches. The global image representation is obtained using a position-aware pooling operation, which is usually performed on the output of the last encoder layer. Finally, the image classification task is performed by concatenating the location-aware pooling outputs to the fully connected layer and adding a softmax layer.

2.2 | Feature extraction module

In the ViT model, the input image is convolved with a convolutional kernel of size $p \times p$ for patch tokenization. Where the step size is p, usually set to 16, which is the size of each patch. But directly with the convolution kernel for convolution there will be unable to learn the underlying features of the image, for this problem, this paper proposes a shallow feature extraction module (shallow feature enhancement, SFE), the detailed network structure is shown in Figure 1 after the SFE processing results. the SFE contains five convolutional layers, convolutional layers of the first four layers of the application of the convolution kernel is 3 The first four layers of the convolutional layers apply a network structure with a step size of 2 and padding of 2, and ReLU and batch normalization operations are added. The fifth layer applies a network structure with a convolution kernel of 1, a step size of 1, and a padding of 2. This operation improves the model performance and convergence speed.

2.3 | CNN branch feature extraction

When the target to be classified only occupies a part of the input image, the features of this target may be affected by the complex background information or noise of the image. It is difficult for the traditional convolutional module to accurately capture the key features of the object, thus affecting the accuracy of the classification results. To address the above problems, the CNN branch of this model is based on Resnet50 for feature extraction, and the feature enhancement module (FEM) is used to enhance the information interaction, and at the same time, enhance the information characterization ability to obtain more detailed features. The network structure of the CNN branch is shown in Figure 2.The network model of the FEM module is shown in Figure 3. Based on the definition in Resnet50, this paper divides the main branch into five main stages, each of which consists of multiple convolutional modules, including 1 × 1 down-projection convolutional module, 3 × 3 spatial convolutional module, FEM module, 1 × 1 up-projection convolutional module, and the residual connection between the inputs and the outputs, in which the convolutional module consists of the ordinary convolutional layer Conv2d, BatchNorm, and the ReLU activation function. In the main branch of the CNN, the 3 × 3 convolutional kernel slides over feature mappings with overlap, which provides the ability to preserve fine localized features.

For the FEM module, firstly the module performs convolution operation on the input feature map R using convolution kernels of sizes 3 × 3 and 5 × 5 to obtain two outputs R_1 and R_2 respectively, and each convolution layer is followed by batch normalization and SiLU activation operation in turn. Subsequently, element-by-element summing operation is performed on the above two outputs R_1 and R_2 to obtain a new output feature map R_m:

$$R_1 = \sigma(BN(Conv_{3\times3}(R))) \quad (1)$$

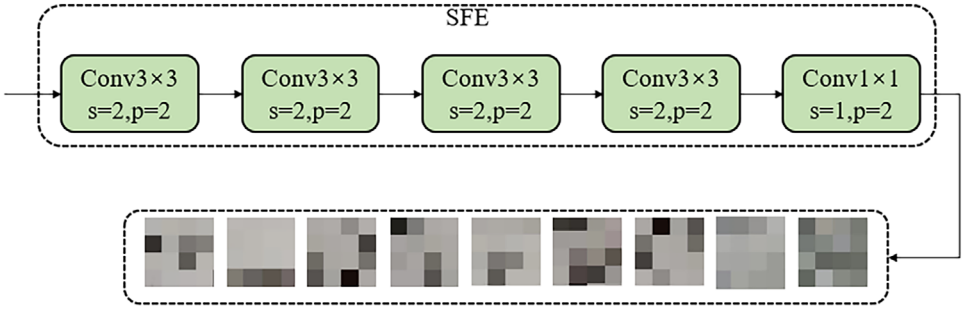


FIGURE 1 Network structure diagram of SFE.

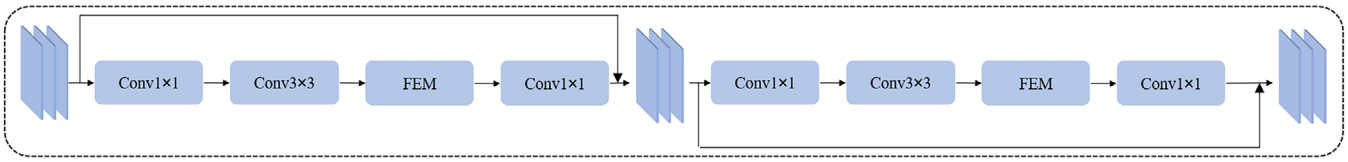


FIGURE 2 Network structure of convolutional neural network branching.

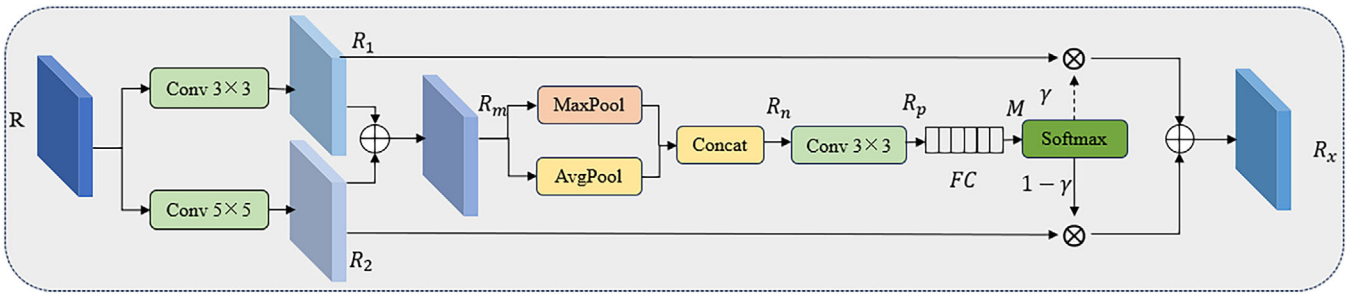


FIGURE 3 Network structure of feature enhancement module.

$$R_2 = \sigma(BN(Conv_{5\times5}(R))) \tag{2}$$

$$R_m = R_1 \oplus R_2 \tag{3}$$

In the above equation $BN(\cdot)$ denotes the batch normalization and $\sigma(\cdot)$ denotes the SiLU activation function.

Then the feature map R_m is downsampled by maximum pooling and average pooling operations and then spliced to obtain R_n, where the maximum pooling operation is used to extract the stronger feature information in the feature map to enhance the invariance and robustness of the model, and the average pooling operation reduces the amount of computation and the number of parameters while obtaining the features with global information and helps to prevent the model from overfitting. R_p is then obtained by integrating R_n with a convolutional kernel of size 3×3, followed by a fully-connected layer to obtain a new feature vector M:

$$R_n = Cat\left(\mathcal{F}_{Map}(R_m); \mathcal{F}_{Avp}(R_m)\right) \tag{4}$$

$$R_p = Conv_{3\times3}(R_n) \tag{5}$$

$$M = FC(R_P) \tag{6}$$

In the above equation $\mathcal{F}_{Map}(\cdot)$ denotes the maximum pooling operation, $\mathcal{F}_{Avp}(\cdot)$ denotes the average pooling operation, and $FC(\cdot)$ denotes the fully connected operation.

Then different convolutional kernel weights are calculated by softmax function and weighted and fused with each branch respectively to get two weighted feature maps. Finally, the branches are summed element by element to obtain the final feature map R_x.

$$R_x = (R_1 \otimes \gamma) \oplus (R_2 \otimes (1-\gamma)) \tag{7}$$

The above FEM module can selectively focus on the target region in the image by adaptively adjusting the convolutional kernel receptive field size, effectively reducing the feature information loss and optimizing the feature information interaction in the context.

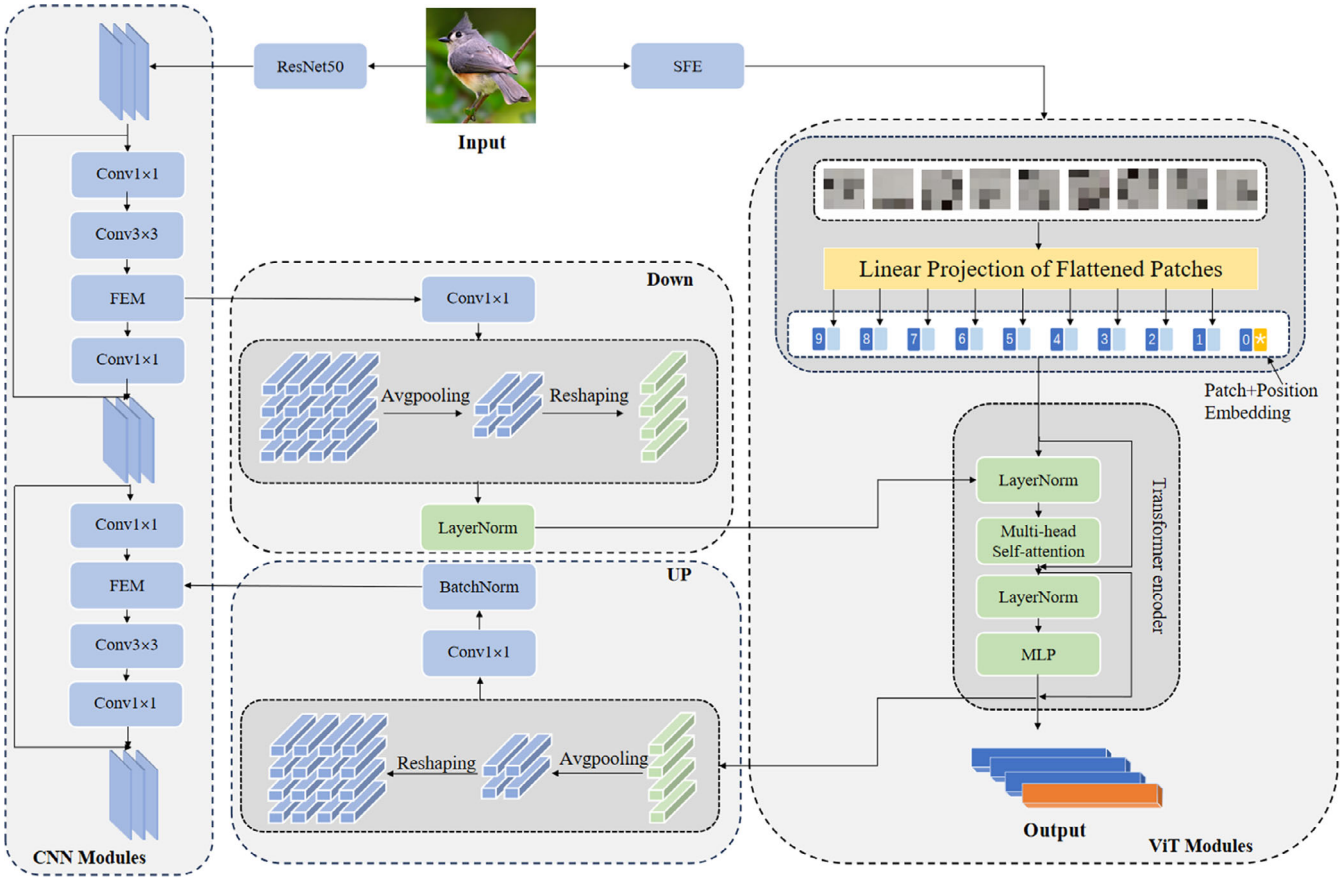


FIGURE 4 CB-ViT model network structure.

2.4 | CNN and ViT bilinear interactive network modelling

To utilize both local features and global representations, we design a CNN with ViT bilinear interaction network model (CB-ViT). Considering the complementary nature of the two styles of features, in CB-ViT, we successively feed the global context of the ViT branch into the feature mapping to enhance the global perception of the CNN branch. Similarly, the local features of the CNN branch are gradually fed into patch embedding to enrich the local details of the ViT branch. Such a process constitutes the interaction, where the CNN branch and the ViT branch consist of N (e.g., 12) repeated convolutional and ViT blocks, respectively. In this case, the network structure of CB-ViT is shown in Figure 4 through the down-sampling Down block and up-sampling UP block as connecting blocks.

For the CNN branch, the input image is firstly subjected to preliminary feature extraction through ResNet50, convolution operation with 1 × 1 down projection convolution module and 3 × 3 spatial convolution module, followed by enhancing the information interaction through the FEM module, preserving the fine-grained local features and inputting them into the down sampling Down module, and completing the spatial dimensionality alignment through the number of channels of the 1 × 1 convolutional alignment, and the average pooling reconstruction. Finally, the feature mapping is added to the patch embedding in the ViT branch through the LayerNorm module. For the ViT branch, the input image first undergoes primary feature extraction by the SFE module to obtain the patches, and then the patches are fed into the linear projection of flattened patches layer to flatten the input sequence, that is, the two-dimensional features are converted into a one-dimensional patch embedding by a linear transformation, and the positional information is added due to the sequential order of the patches. Since the patches are sequential, position embedding is added, and finally the sequence is fed into the Transformer Encoder, which combines with the local information fed by the CNN branch through the down-sampling Down module, and finally outputs the classification result through MLP classification. At the same time, the classification result is converted into two-dimensional sequence through the upsampling UP module, entered into the 1 × 1 convolution for dimensional alignment, the BatchNorm regularizes the features and inputs them into the CNN branch, the global information obtained from the ViT branch is fed back to the CNN branch, and combined with the FEM module to achieve the information interaction between the local and global information, and ultimately realize the fine-grained image classification of the image.

TABLE 1 Dataset specific information.

Data set	Class	Training set	Test set
CUB-200-2011	200	5994	5794
Stanford Cars	196	8144	8041
FGVC-Aircraft	100	6667	3333

3 | ENVIRONMENT CONFIGURATION

3.1 | Dataset

Fine-grained image datasets are datasets used for fine-grained classification tasks where each category contains objects or scenes with subtle differences. These datasets are commonly used to train and evaluate computer vision models to recognize and distinguish objects that are similar but belong to different fine-grained categories. This paper uses three public datasets, CUB-200-2011, Stanford Cars and FGVC-Aircraft, to validate the algorithm, and the specific information of the datasets is shown in Table 1.

3.2 | Experimental environment

The open source PyTorch deep learning framework is used to implement a fine-grained image classification network based on Vision Transformer. The operating systems are Windows 10, the GPU model is NVIDIA GeForce RTX 3090, the CPU is Inter(R)Core(TM)i9-9900CPU@3.10 GHz, the computing architecture CUDA version 11.3, Python as the programming language with version 3.9.7, Pytorch version 1.12.1, and GPU acceleration using CUDA and CUDNN to improve the computing power of the computer. All the images were scaled to 600 × 600 size as input data. In the training phase, random horizontal flipping, random cropping and normalized enhancement were used to process the images to 448 × 448 size and fed into the network. The initial learning rate of the network is set to 0.001, the initial momentum is 0.9, the weight decay is 0.00005, the BatchSize is set to 8, and the Epoch is 100.

4 | EXPERIMENTAL RESULTS AND ANALYSIS

To facilitate the observation and evaluation of the proposed method, mean average precision (mAP) and F1 score are employed to assess the classification capability of the model, while the number of model parameters (Params) is selected to evaluate its complexity, and frame rate (FPS) is utilized to measure computational efficiency.

Average precision (AP) quantifies the classification accuracy of the model for a specific target type. It is represented as a precision-recall curve, with precision and recall plotted on the horizontal and vertical axes, respectively. The AP values across all target types in the dataset are then averaged to obtain mAP, which serves as an overall performance metric for the network model. The F1 score incorporates both precision and recall of the classification model, thereby reflecting its comprehensive classification capability. Its formula is presented below, where n denotes the number of categories.

$$AP = \int_0^1 P(R)\, dR \tag{8}$$

$$mAP = \frac{1}{n} \sum_{i=1}^{n} AP \tag{9}$$

$$F1 = 2\frac{P \times R}{P + R} \tag{10}$$

In this context, P denotes the precision value and R signifies the recall value.

The number of parameters serves as an indicator of the network's structural complexity; a more intricate architecture poses greater challenges for practical deployment. The frame rate (FPS) reflects the model's real-time responsiveness, quantifying the number of image frames processed by the network per second. A higher frame rate indicates enhanced computational efficiency of the network.

4.1 | Ablation test

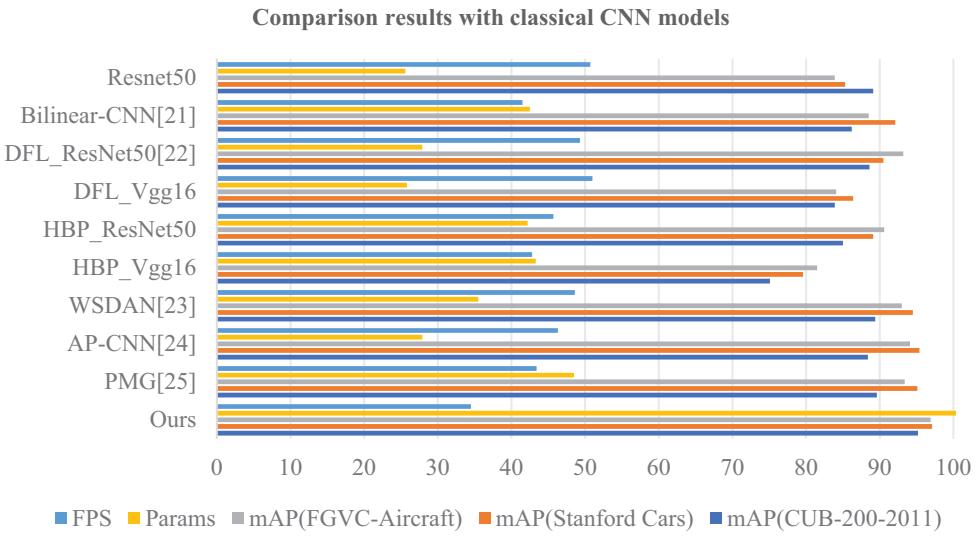
In order to verify the effectiveness of different modules, this paper conducts inter-module ablation experiments on CUB-200-2011 public data set, and the experimental results are shown in Table 2 below.

From the ablation experiments in Table 2, the accuracy rate of Resnet50 and ViT network model is 88.7% and 89.5%, respectively, and the accuracy rate of Resnet50+ ViT network model is 90.2% at the time of the ablation experiments, which is accurately higher than that of the Resnet50 and ViT network model trained alone. However, compared to a single-network configuration, this two-branch structure increases computational demand to 134.5 M, leading to a reduction in computational efficiency; consequently, the number of frames is reduced to 40.8 fps, but the performance of the model is improved. The input image is processed using the ResNet50 + ViT network model. Following the incorporation of the FEM module, the model achieves a mAP of 93% and an F1 score of 90.3%, representing an increase of 1.1% in mAP and 2.5% in F1 score compared to the baseline ResNet50 + ViT model. This improvement indicates that the FEM module effectively mitigates background interference, allowing the network to learn more relevant feature information. Additionally, shallow feature extraction conducted on top of the ResNet50 + ViT architecture yields a mAP of 93.3% and an F1 score of 90.6%, suggesting that the SFE module enhances low-level feature learning capabilities within the network. When both FEM and SFE modules are integrated into the ResNet50 + ViT framework simultaneously, we observe a further increase in mAP to 96.4% and an F1 score rise to 94.1%. However, this dual-branch configuration results in a total parameter count of 162.6 M and

TABLE 2 Ablation experiments.

Methods	Precision (%)	mAP (%)	F1 (%)	Params (M)	FPS
Resnet50	88.7	89.1	86.7	25.6	50.7
ViT	89.5	91	87.1	86	46.7
Resnet50+ ViT	90.2	91.9	87.8	134.5	40.8
Resnet50+ ViT+FEM	91.9	93	90.3	155.8	38.4
Resnet50+ ViT+SFE	91.6	93.3	90.6	149.6	36.8
Ours	**95.2**	**96.4**	**94.1**	**162.6**	**34.5**

TABLE 3 Accuracy comparison with classical CNN model (%).



a frame rate reduction to 34.5 fps, indicating that while it successfully integrates low-level features from SFE with global features from FEM, it also leads to increased computational demands with a slight decrease in efficiency overall. In summary, these enhancements demonstrate effective improvements in model performance.

4.2 | Comparison test

In order to evaluate the performance of the algorithms in this paper in fine-grained image classification tasks, we conducted comparative experiments with other weakly-supervised mainstream CNN-based network models. To ensure the reliability of the experimental results, we train and test all network models on the same dataset. Table 3 below shows the classification accuracy comparison results of this paper's algorithm with other algorithms:

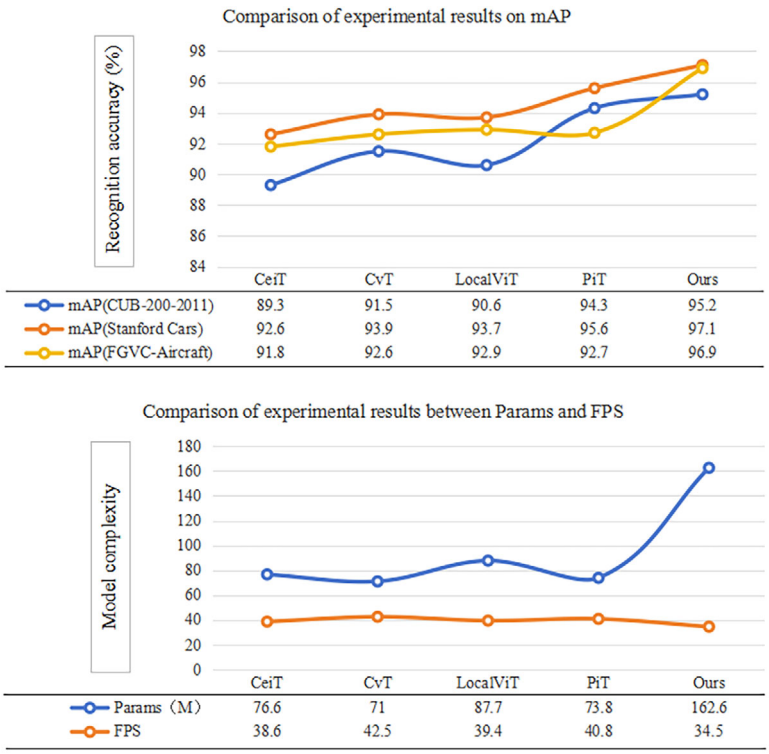
In order to verify the superiority of this scheme and analyse the classification performance of this paper's algorithm, Table 3 shows the results of the accuracy comparison between this paper's algorithm and the improved CNN algorithm model on three public fine-grained datasets, from the data in the table, this paper's algorithm obtains the highest detection results on three public datasets, and the average classification accuracy on the CUB-200-2011 bird dataset is 94.2%, which is 4.6% higher than the PMG algorithm with relatively better detection results by 4.6%. The average classification accuracy on the Stanford Cars automobile dataset is 96.8%, which is 1.7% better than the PMG algorithm. The average classification accuracy on the FGVC-Aircraft dataset is 95.8%, which is 2.4% better than the PMG algorithm. The above shows that the algorithm in this paper is able to extract more effective features and can achieve better results compared with the traditional weakly supervised CNN model.

Table 4 shows the accuracy comparison between this paper's algorithm and the classical improved ViT algorithm, from the data in the table, this paper's algorithm achieved the highest classification results on the three public datasets, specifically when compared to Swim-T, ViViT, HVT-S, PVT-S, BeiT, and CPVT-S on the CUB-200-2011 bird dataset. Notably, in comparison with DeiT-S and T2T-ViT algorithms, the mAP improves by 7.6%, 11.8%, 11.3%, 9.3%, 8.9%, 6.3%, 7.6% and 3.9% respectively. On the Stanford Cars automobile dataset, mAP increases by 4.3%, 11.5%, 9.9%, 6.8%, 4.6%, 6.5%, 3.8% and 5.3%. Similarly, for the FGVC-Aircraft aircraft dataset, mAP shows improvements of 5%, 11.5%, 8.3%, 7.5%, 4.8%, 7.4%, 6.3% and 6.5%. The comparative analysis of the data demonstrates that the proposed algorithm constructs a dual-branch network integrating CNN and ViT, enabling the extraction of more local

TABLE 4 Comparison of accuracy with improved ViT performed.

Methods	mAP (%) CUB-200-2011	mAP (%) Stanford Cars	mAP (%) FGVC-Aircraft	Params (M)	FPS
Swim-T [26]	87.6	92.8	91.9	88	38.8
ViViT [27]	83.4	85.6	85.4	19	52.3
HVT-S [28]	83.9	87.2	88.6	21.7	48
PVT-S	85.9	90.3	89.4	25	45.3
BeiT [29]	86.3	92.5	92.1	24.6	43.3
CPVT-S [30]	88.9	90.6	89.3	22	46.5
DeiT-S [31]	87.6	93.3	90.6	22	42.8
T2T-ViT [32]	91.3	91.8	90.4	21.6	46.9
Ours	**95.2**	**97.1**	**96.9**	**162.6**	**34.5**

TABLE 5 Comparison of accuracy with the ViT algorithm model incorporating CNNs.



features and yielding superior classification results compared to traditional improved ViT models.

Table 5 presents the results of a comparative analysis between the proposed algorithm and the ViT model integrated with CNN. The findings indicate that the mAP of the proposed algorithm on the CUB-200-2011 dataset surpasses that of CeiT, CvT LocalViT, and PiT [33] algorithms by 5.9%, 3.7%, 4.6%, and 0.9%, respectively. On the Stanford Cars dataset, the mAP is higher than those three algorithms by 4.5%, 3.2%, and 1.5%, respectively. Similarly, on the FGVC-Aircraft dataset, it exceeds these algorithms by margins of 5.1%, 4.3%, and 4%. A thorough examination of these statistical data leads to the conclusion that while integrating CNN with ViT in a dual-branch network increases parameter count to 162.6 M, it also reduces model inference speed to 34.5 fps. Nevertheless, this computational overhead facilitates improved classification performance, underscoring the effectiveness of our model in fine-grained image classification tasks.

4.3 | Deployment experiment of hardware equipment

To demonstrate the classification performance and computational efficiency of the improved model in practical applications, the dual-branch information interaction model was trained

TABLE 6 Hardware device deployment experiments.

Mobile devices	Method	FPS	Params
Windows10	Ours	34.5	162.6
EC-R3588SPC		20.3	62.6
NVIDIA Jetson Orin NX		25.6	62.6

on a PC to obtain optimal weights. These weights were then deployed on two hardware devices, EC-R3588SPC and NVIDIA Jetson Orin NX, to perform fine-grained image classification tasks. The frame rate and number of model parameters served as evaluation metrics to assess the feasibility of the improved model in real-world application scenarios. Table 6 presents the results of the actual deployment experiment.

Based on the experimental results, the frame rate of the proposed algorithm in the Windows system is 34.5 fps, with a model size of 162.6 M. Initially, the model was converted to Linux format and imported into the EC-R3588SPC terminal, where it achieved a frame rate of 20.3 fps during execution. To accommodate edge devices, the model parameters were compressed to 62.6 M, indicating that although the computing power of the EC-R3588SPC host is lower than that of the PC terminal—resulting in a decreased FPS value—the algorithm still demonstrates good computational efficiency in practical scenarios. Conversely, NVIDIA Jetson Orin NX possesses greater computing power than EC-R3588SPC; therefore, under identical model size conditions, the dual-branch information interaction model deployed on NVIDIA Jetson Orin NX achieves a higher frame rate of 25.6 fps and enhanced computational efficiency. In summary, the proposed algorithm exhibits superior performance when deployed in low-resource environments.

5 | CONCLUSION

In order to improve the detection accuracy of fine-grained image classification and make full use of the potential of CNN focusing on extracting local features and ViT for global features, this paper proposes a bilinear interactive network model of CNN and ViT, which maximizes the retention of local features and global representations by parallelizing the CNN branch and ViT branch. And the two are interactively connected by down-sampling Down module and up-sampling UP module, which fuses the local features in the CNN branch with the global representation in the ViT branch. In order to avoid the influence of background noise and the phenomenon that shallow features cannot learn the underlying features of the image, a feature enhancement module and a shallow feature extraction module are proposed. In order to verify the effectiveness and generalization of this paper's algorithm, a large number of tests are carried out on three public fine-grained datasets and a large number of comparative experiments are conducted, which show that this paper's algorithm achieves a more considerable result through the experimental data. Nevertheless, the algorithm exhibits certain limitations, including a substantial number of model parameters and relatively low computational speed. In future work, we will undertake comprehensive research on lightweight networks to enhance both the effectiveness and efficiency of the network.

AUTHOR CONTRIBUTIONS

Shuo Zhu: Funding acquisition; writing—review and editing. **Xukang Zhang**: Formal analysis; software. **Yu Wang**: Writing—original draft. **Zongyang Wang**: Resources. **Jiahao Sun**: Investigation.

CONFLICT OF INTEREST STATEMENT

The authors declared no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Shuo Zhu https://orcid.org/0009-0005-0781-233X

REFERENCES

1. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM 60(6), 84–90 (2012)
2. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
3. Targ, S., Almeida, D., Lyman, K.: Resnet in resnet: Generalizing residual architectures. arXiv:1603.08029 (2016)
4. Szegedy, C., Liu, W., Jia, Y.Q., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9. IEEE, Piscataway, NJ (2015)
5. Tan, M.X., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the International Conference on Machine Learning, pp. 6105–6114. PMLR, Microtome Publishing, Brookline, MA (2019)
6. Zhao, X., Wang, L.M., Zhang, Y.F., et al.: A review of convolutional neural networks in computer vision. Artif. Intell. Rev. 57(4), 99 (2024)
7. Yu, C.J., Zhao, X.Y., Zheng, Q., et al.: Hierarchical bilinear pooling for fine-grained visual recognition. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 574–589. Springer, Cham (2018)
8. Tan, M., Wang, G.J., Zhou, J., et al.: Fine-grained classification via hierarchical bilinear pooling with aggregated slack mask. IEEE Access 7, 117944–117953 (2019)
9. Kong, S., Fowlkes, C.: Low-rank bilinear pooling for fine-grained classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 365–374. IEEE, Piscataway, NJ (2017)
10. Chang, P.S.: Fine-grained model identification based on deep learning. Dissertation, Nanjing University of Information Engineering (2022). https://doi.org/10.27248/d.cnki.gnjqc.2022.001004
11. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 6000–6010. Curran Associates Inc., Red Hook, NY (2017)
12. Dosovitskiy, A.: An image is worth 16×16 words: Transformers for image recognition at scale. arXiv:2010.11929 (2020)
13. Pan, Z.Z.H., Zhuang, B.H., Liu, J., et al.: Scalable vision transformers with hierarchical pooling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 377–386. IEEE, Piscataway, NJ (2021)
14. Wang, W.H., Xie, E.Z., Li, X., et al.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. arXiv.2102.12122 (2021). https://doi.org/10.48550/arXiv.2102.12122

15. d'Ascoli, S., Touvron, H., Leavitt, M.L., et al.: Convit: Improving vision transformers with soft convolutional inductive biases. In: Proceedings of the International Conference on Machine Learning, pp. 2286–2296. PMLR, Microtome Publishing, Brookline, MA (2021)
16. He, M.G., Cheng, Q.L., Qi, G.Q.: Weakly supervised semantic and attentive data mixing augmentation for fine-grained visual categorization. IEEE Access 10, 35814–35823 (2022)
17. Guo, J.Q., Qi, G.Q., Li, X.Y., et al.: Two-branch attention learning for fine-grained class incremental learning. Electronics 10(23), 2987 (2021)
18. Yuan, K., Guo, S.P., Liu, Z.W., et al.: Incorporating convolution designs into visual transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 579–588. IEEE, Piscataway, NJ (2021)
19. Wu, H.P., Xiao, B., Codella, N., et al.: CVT: Introducing convolutions to vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 22–31. IEEE, Piscataway, NJ (2021)
20. Li, Y.W., Zhang, K., Cao, J.Z., et al.: LocalViT: Bringing locality to vision transformers. arXiv.2104.05707 (2021). https://doi.org/10.48550/arXiv.2104.05707
21. Lin, T.Y., Roy Chowdhury, A., Maji, S.: Bilinear CNNs for fine-grained visual recognition. arXiv preprint arXiv:1504.07889 (2015)
22. Wang, Y.M., Morariu, V.I., Davis, L.S.: Learning a discriminative filter bank within a CNN for fine-grained recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4148–4157. IEEE, Piscataway, NJ (2018)
23. Hu, T., Qi, H.: See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. arXiv.1901.09891 (2019). https://doi.org/10.48550/arXiv.1901.09891
24. Ding, Y.F., Ma, Z.Y., Wen, S.G., et al.: AP-CNN: Weakly supervised attention pyramid convolutional neural network for fine-grained visual classification. IEEE Trans. Image Process. 30, 2826–2836 (2021). https://doi.org/10.1109/TIP.2021.3055617
25. Du, R.Y., Chang, D.L., Bhunia, A.K., et al.: Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In: Proceedings of the European Conference on Computer Vision, pp. 153–168. Springer International Publishing, Cham (2020)
26. Liu, Z., Lin, Y.T., Cao, Y., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022. IEEE, Piscataway, NJ (2021)
27. Arnab, A., Dehghani, M., Heigold, G., et al.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6836–6846. IEEE, Piscataway, NJ (2021)
28. Pan, Z.Z., Zhuang, B.H., Liu, J., et al.: Scalable vision transformers with hierarchical pooling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 377–386.4. IEEE, Piscataway, NJ (2021)
29. Bao, H.B., Dong, L., Wei, F.R., BEiT: BERT pre-training of image transformers. arXiv.2106.08254 (2021). https://doi.org/10.48550/arXiv.2106.08254
30. Chu, X.X., Zh, T., Zhang, B., et al.: Conditional positional encodings for vision transformers. arXiv:2102.10882 (2023)
31. Alotaibi, A., Alafif, T., Alkhilaiwi, F., et al.: VIT-DEIT: An ensemble model for breast cancer histopathological images classification. In: Proceedings of the 2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC), pp. 1–6. IEEE, Piscataway, NJ (2023)
32. Qi, Y.H., Wang, L., Li, K.Y., et al.: Latent diffusion model-based T2T-ViT for SAR ship classification. In: CCF Conference on Computer Supported Cooperative Work and Social Computing, pp. 296–305. Springer Nature, Singapore (2023)
33. Chen, J.F., Wu, K.L.: Positional knowledge is all you need: Position-induced transformer (PiT) for operator learning. arXiv:2405.09285 (2024)

How to cite this article: Zhu, S., Zhang, X., Wang, Y., Wang, Z., Sun, J.: A fine-grained image classification method based on information interaction. IET Image Process. 18, 4852–4861 (2024). https://doi.org/10.1049/ipr2.13295