IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 January 2024, accepted 29 January 2024, date of publication 22 February 2024, date of current version 5 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3369032

RESEARCH ARTICLE

Multi-Model Fusion Fine-Grained Image Classification Method Based on Migration Learning

WENYING ZHANG AND YAPING WANG
School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou 450001, China
Corresponding author: Wenying Zhang (wyzhangzzu@163.com)

ABSTRACT Current single-model methods for fine-grained image classification suffer from insufficient generalisation ability, while multi-model fusion methods suffer from weight curing. The study suggests and experimentally tests a dynamic weight multi-model fusion strategy for transfer learning-based fine-grained picture classification. The results of the experiment showed that the suggested fusion model enhanced recognition accuracy by 1.33%, 1.19%, and 0.83% compared to the single model on the medical dataset and 3.25%, 1.34%, and 7.28% on the agronomy dataset, respectively. Furthermore, when compared to the comparison method, the models under the proposed method of the study improved recognition accuracy by 0.18%, 0.61%, 0.43%, and 0.43% on the medical dataset, and the experimental time consumed was 3.25 minutes less than that of the sum-of-maximum-probabilities method; however, the fusion models of the proposed method of the study had higher recognition accuracy than that of the comparison met Overall, the proposed dynamic weight multi-model fusion method for fine-grained image classification using migration learning has better performance and generalisation ability, which can improve performance while reducing time cost, and has higher application value for the actual fine-grained image classification task.

INDEX TERMS Migration learning, multi-model, fine-grained image, generalization capability, CNN.

I. INTRODUCTION

The advancement of computer and modern communication technology has made it increasingly easier for people to acquire image data and classify images. The actual task of image classification is to use the network model to propose corresponding features to the objects to be classified in the image in order to achieve the goal of judging the classification of the category to which they belong [1]. The current image classification system includes both coarse- and fine-grained image (FGI) classification; the latter is more closely related to human life and has greater practical value, thus it has attracted a lot of attention [2]. Early FGI classification is generally accomplished through the use of machine learning algorithms, which typically analyse the lower level information in the image and do not involve the higher level abstraction information, resulting in a limitation that cannot satisfy the needs of the actual scene [3]. Deep learning (DL) is a new notion for FGI classification since it can automatically extract the actual feature information of an image layer by layer as artificial intelligence has progressed [4]. When using fine-grained image classification, there is often a problem of insufficient training samples. Compared to previous methods, transfer learning allows researchers to conduct experiments using pre trained models, which are typically trained on large-scale datasets and can provide rich feature representations. Meanwhile, transferring learned features can achieve more accurate classification on relatively small datasets. In addition, transfer learning can effectively integrate the pre training knowledge of various models and improve classification performance.

However, in fine-grained image classification, although single model classification methods based on deep learning can achieve good results, they still have some shortcomings

The associate editor coordinating the review of this manuscript and approving it for publication was Hongjun Su.

compared to multi model fusion methods, and significantly increase time costs, making recognition accuracy difficult to meet practical requirements. However, the fixed weight multi model fusion method adopts an empirical assignment method for the contribution weights of each sub model, resulting in fixed weight values and randomness in the results, making it difficult to ensure that the model converges to the optimal solution. In view of this, the experiment proposes a multi-model fusion fine-grained image classification method based on transfer learning, aiming to explore efficient methods for fine-grained image classification.

During the experiment, a dynamic weight multi model fusion method (DWMF) for fine-grained image classification was proposed by using transfer learning. It is expected to solve the problem of insufficient generalization ability of single model methods and the problem of traditional multi-model fusion methods. The problem of value solidification. Compared with existing single model classification methods using deep learning (such as attention paired interaction networks, weakly supervised fine-grained image classification networks based on attention guided image enhancement, etc.) and fine-grained image classification methods based on multi model fusion (such as image recognition methods based on self coding and convolutional neural network fusion, etc.), the experimental method provides unique insights by combining transfer learning and multi model fusion. This model not only improves the accuracy of classification, but also enhances the recognition ability of the model for fine-grained features. In addition, this method also demonstrates how to utilize pre trained models in resource constrained situations, which is not common in current fine-grained image classification research.

The whole study is organised into four parts, the first of which summarises and examines existing research on FGI classification methods for DL under DL. The second section provides a summary of the theoretical underpinnings of the FGI classification approach for MMF as well as the proposed research method. The final section is an experimental verification of the research method's validity. The fourth section provides a synopsis of the entire article.

II. RELATED WORK

According to the quantity of manually labelled data required for the algorithmic model to be trained, the current FGI classification methods using DL can be divided into two categories: those that use strongly supervised information and those that use weakly supervised information [5]. Among them, the latter combines the target detection algorithm, attention mechanism, reinforcement learning and other methods, which not only saves the cost of manual annotation, but also has more research value for practical application and promotion on the basis of meeting or even exceeding the FGI classification method based on strongly supervised information [6]. Wang et al. addressed the problems related to the classification of FGIs due to intra- and inter-class differences by proposing a neural model incorporating a new attentional mechanism on the basis of the aggregated attention module, which effectively improves the accuracy of recognizing and classifying FGIs [7]. Wei et al. addressed the problem of FGI analysis in computer vision and pattern recognition by reviewing the research of DL in FGI classification, which redefined and broadened the field of FGI analysis, and thus effectively solved the key problems in FGI classification [8]. Adem et al. addressed the problem of FGI in agricultural vegetable leaves by proposing a model for image recognition classification based on image processing and DL, thus effectively reducing the diagnosis time of actual disease categories while improving the recognition accuracy [9]. By putting up a model for picture recognition classification based on image processing and DL, Ngugi L.C. et al. attempted to address the issue of plant disease identification. On the basis of image processing methods and DL, a model for picture identification and classification was put out, effectively cutting the diagnosis time for actual disease categories while increasing the recognition accuracy [9]. Ngugi et al. addressed the related problems in plant disease detection by proposing a model for automatic recognition and classification of pictures on the basis of image processing techniques and DL, thus effectively improving the diagnosis and recognition accuracy of plant diseases [10].

In addition, Chen et al. proposed an improved model fusing segmented linear representation and weighted support vector machine for the problems related to FGI classification by applying it in real stock image analysis, thus effectively improving the recognition accuracy of FGI [11]. Liu et al. addressed the problem of recognizing and classifying hyperspectral FGIs by organically integrating convolutional and multimodal neural network models on the basis of dynamic stochastic resonance, thus effectively improving the recognition accuracy of FGIs based on the use of MMFs [12]. Jenisha and Dickson proposed a fusion model for automated FGI segmentation by using deep ML and attention mechanism for problems related to FGI recognition segmentation of liver tumors, thus in providing help in enhancing the segmentation of liver tumor images [13]. Vo et al. addressed the problems related to DL in medical FGI recognition and classification, and constructed an X-network model by fusing a residual network model with a squeezing and excitation model, which effectively improved the predictability of medical image classification [14].

FGI classification methods based on weakly supervised information mainly include two kinds of FGI classification using SM and MMF. However, the current SM method only makes use of parameter tuning or fine-tuning the model structure based on the current network model to improve recognition accuracy in image classification, which not only adds to the time cost but also makes it challenging to obtain satisfactory results. The MMF method for image classification has fixed weights for each sub-model and the researchers assign values to each sub-model based on their own experience, so it cannot reflect the contribution of each sub-model in the actual classification task. Based on this, the study proposes

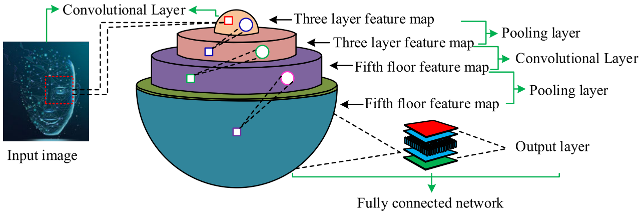


FIGURE 1. Schematic diagram of the specific architecture of CNN in the field of image recognition and classification.

a dynamic weighting MMF method for FGI classification by utilizing ML, which not only can solve the problem of insufficient GC, but also can realize the dynamic adjustment of the weights, so it is innovative. Meanwhile, in theory, research methods have a significant improvement in classification performance compared to single model methods; Compared to traditional fixed weight multi model fusion methods, the research method not only improves classification performance, but also achieves the optimal state faster, saving time and computational costs. Therefore, it has superiority and innovation.

III. ANALYSIS OF CLASSIFICATION METHODS FOR FGI OF MMFS

Numerous studies have demonstrated that the MMF approach has better performance than a single model, and its practical application mainly depends on the network model used and the fusion strategy of different models. Therefore, this section focuses on the main models of MMF and analyzes the proposed MMF approach.

A. CNN MODELING AND MMF ANALYSIS

Aiming at the GC deficiency of the current SM method and the solidification of the weights of the traditional MMF method, the study proposes a dynamic weighted MMF method for FGI classification by utilizing ML. Convolutional Neural Networks (CNN) and Deep Neural Networks are the basic technologies for the approach, so the study first examines both of them.CNN is a feed-forward neural network that was developed using DL, and its operation was influenced by the visual cortex of the human eye, which is naturally better at recognising images [15], [16]. CNN uses a "end-to-end" approach to image recognition, but it can be intuitively divided into two parts based on the functions played by each component of its structure: feature extraction and classification. The feature extraction component uses alternating convolutional and pooling layers to perform, and the classification component is typically implemented by a fully connected layer [17], [18]. Thus, the specific CNN structure in the area of image identification and classification is depicted in Figure 1.

As shown in Figure 1, the study sets the feature maps in the CNN's convolutional layer to 3 layers, the pooling layer's feature maps to 3 layers, the convolutional layer's feature maps to 5 layers, the secondary pooling layer's feature maps to 5 layers, and then it moves on to the fully-connected layer and the output layer. The central part of the CNN is the convolutional layer, whose primary job is to automatically extract the input data and local features using convolutional operations and numerous convolutional kernels to produce the feature map. This process can be viewed as filtering the entire image with a filter of a specific size and specific rules, and then filtering out the content that matches the filtering rules of that filter to get a feature map. The major convolution kernel parameters are its size, the motion step size, and the filling of the feature map. This filter is also known as a convolution kernel. In this case, the convolution is computed as the sum of the products of the convolution kernel weights and the corresponding pixel values, which is expressed as shown in equation (1).

$$conv_{p,q} = \sum_{i}^{x*y} \gamma_i \lambda_j \tag{1}$$

In equation (1), $conv_{p,q}$ denotes the convolution result value; p and q denote the horizontal and vertical coordinates of the actual feature map; $x \times y$ denotes the maximum size of the convolution kernel, and i and j are the internal size sizes of the two, respectively; γ denotes the visual weight of the convolution kernel; and λ denotes the relevant pixel value of the corresponding point of the actual image. The pooling layer mainly plays the role of pooling operation on the feature map, which improves the robustness of the CNN. The pooling operation adopts a method similar to downsampling, which can both reduce the computational amount of the model and compress the features, thus effectively eliminating the redundant information in the data and preventing the overfitting of the data. In the pooling layer, the pooling operation can be divided into average pooling and maximum pooling, which reduces all the values on each block to a single number after average or maximum pooling of the input image. The corresponding expression is shown in equation (2).

$$B_k^l\left(i,j\right) = \left[\sum_{p=1}^{f}\sum_{q=1}^{f} B_k^l\left(z_o i + p, z_o j + q\right)^w\right]^{\frac{1}{w}} \tag{2}$$

In equation (2), B_k^l denotes the simplified value; f denotes the feature map size; z_o denotes the step size; and w denotes the pre-specified value, which when it is 1 denotes the average pooling, and the corresponding region takes the mean value, and when it tends to infinity denotes the maximum pooling, and the corresponding region takes the maximum value. In addition, CNN internal important also contains activation function, commonly used activation function contains S-type activation function and Linear Rectification Function (ReLu), the expression of the two as shown in equation (3) and equation (4).

$$f\left(h\right) = \frac{1}{1+e^{-h}} \tag{3}$$

In equation (3), the first line denotes the S-type activation function expression and h denotes the function

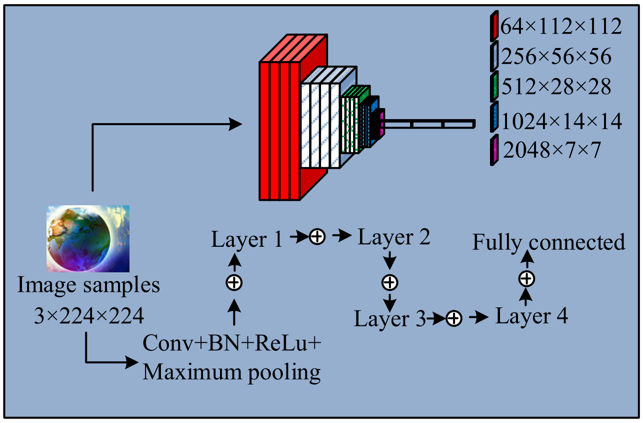


FIGURE 2. Typical ResNet network structure diagram.

independent variable.

$$f'(h)=\begin{cases} h, & h\geq 0\\ 0, & h<0\end{cases} \tag{4}$$

Equation (4) represents the ReLu function expression.The most important network model in CNN modeling is the Residual Network (ResNet), mainly because of its solution to the long unsolved problem of gradient vanishing. Figure 2 displays a typical ResNet network structure.

As can be seen in Figure 2, a typical ResNet network structure starts with image samples of $3 \times 224 \times 224$ size, and on the one hand, it is output through the form of continuous reduction, which can make the original image construction level significantly improved; on the other hand, it is output through the fully connected layer after multi-layer network through convolution operation, batch normalization, ReLu function and maximum pooling operation. For the ResNet network model, the most important component is the residual learning unit, the basic idea is to draw on the residual learning, using a technique called "jump connection" to skip multiple levels of input training, and will be directly connected to the output. The expression of the residual equation is shown in equation (5).

$$G\left(p'\right)=E\left(p'\right)+p' \Rightarrow E\left(p'\right)=G\left(p'\right)-p' \tag{5}$$

In equation (5), p' denotes the residual value; $E\left(p'\right)$ the observed value; $G\left(p'\right)$ the actual value. In addition to this, deep neural network models have developed many models on the basis of CNN, such as Attention Mechanism, Transformer Structure and so on. Among them, the typical model in the Transformer structure is the Vision Transformer (Vit) model, which outperforms most of the CNN models on image networks with a recognition accuracy of 88.55%. The traditional CNN is built by associating the pixel points in the whole image over a long period of time, while for a very small convolutional kernel granularity can not completely cover the whole image, and it is necessary to expand the perceptual field by increasing the depth of the model, but such a method not only can not achieve the expected results, but also has some undesirable effects.Transformer, on the other hand, adopts the use of self-attention layer to build the long-distance dependency, reducing the degree of dependency on external information so as to capture the internal relevance of the data or features and maximize the use of the information inherent in the features themselves to interact with the attention.

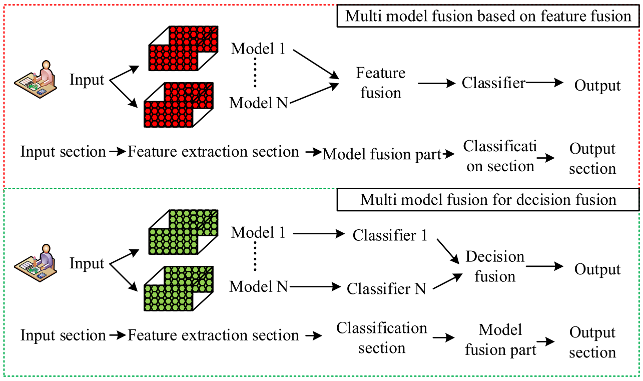


FIGURE 3. Schematic diagram of multi model fusion under feature fusion and decision fusion.

The models constructed for the study are then fused with multiple models.The principle of MMF is to train multiple base models simultaneously, after which these base models are fused using a model fusion strategy, and finally the fused integrated network model is obtained.The MMF usually contains feature and decision fusion, where the schematic diagram of the MMF under feature fusion and decision fusion is shown in Figure 3.

As can be seen from Figure 3, feature fusion refers to the fusion of features extracted from multiple sub-models to form a feature map that is more informative and more conducive to classification, after which the relevant classification is carried out through the fully connected layer. Decision fusion is by fusing and re-calculating the classification probability results output from each sub-model, and finally outputting the actual classification results.

B. RESEARCH ON DYNAMIC WEIGHTED MMF METHOD

On the basis of the theory of CNN in deep neural network part of the model and the theory of MMF, the research began to propose the MMF method for FGI. In practice, MMF mainly depends on the network model used and the fusion strategy of different models, by applying the network model constructed on the basis of ML to the actual MMF not only can achieve better results than a single model, but also can effectively improve the efficiency. However, the current MMF method realizes fusion by fixing the weights, but the method ignores the role played by each sub-model in the whole process of model operation and the differences in the actual classification, and also fails to take into account the problem that the experimenter assigns the values based on his/her own experience. Therefore, the study utilizes the idea of dynamic adjustment to propose the DWMF method, the framework of which is shown in Figure 4.

The primary components of the DWMF structure include input and data pre-processing, model library, weight

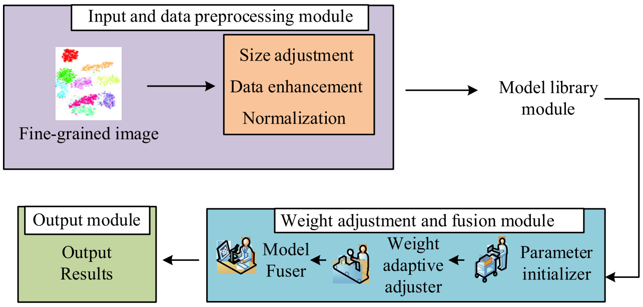


FIGURE 4. Schematic diagram of the structural framework of dynamic weight multi model fusion method for fine-grained image classification.

adjustment and fusion, and output modules, as shown in Figure 4. The input and data pre-processing module contains the actual data input and data pre-processing parts. Among them, the actual data input part is mainly responsible for receiving the input image sampling data, and this method can accept Red Green Blue (RGB) image sampling. The data pre-processing part is responsible for pre-processing the received image samples to eliminate the inconsistency of size and data distribution in the image samples, which provides the basis for the subsequent use of image samples in the model. The role of the model library module is to store the sub-models involved in the MMF process and to provide users with the functions of managing and configuring the sub-models. There are three network models placed inside the model library, namely, ResNet50, Efficient Network_b0 (EfficientNet_b0) model, and the model about Vit chunking (vit_base_patch16_224), which are all experimental models without structural modification and have been subjected to the corresponding changes in the ImageNet dataset. structural modifications and all three network models are experimental models, which have not been structurally modified and have been pre-trained accordingly in the ImageNet dataset.

The primary function of the weight adjustment and fusion module is to correct and combine the weights of each sub-model. During the model fusion process in this module, the weight value of each sub-model will be changed in real time. Each sub-model's weight value will be dynamically modified in accordance with the accuracy of the previous round; the higher the accuracy, the better the sub-model's performance, the higher the weight value of the model will be, and vice versa. The module is divided into a parameter initializer, a weight adjuster, and a model fuser. The parameter initializer initializes the parameters before the implementation of the actual method, and in the actual application, it is mainly complicated to initialize the parameters in the method automatically according to the model user's own choice. Among them, the model weight parameter is particularly important for comparing the expression of sub-models on the dataset before model fusion, and its calculation expression is shown in equation (6).

$$M_l = \frac{1}{F}, \quad l \in F \tag{6}$$

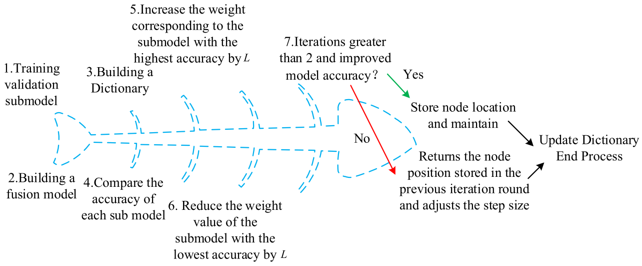

5.Increase the weight corresponding to the submodel with the highest accuracy by L
1.Training validation submodel
3.Building a Dictionary
7.Iterations greater than 2 and improved model accuracy?
Yes
Store node location and maintain
No
Update Dictionary End Process
2.Building a fusion model
4.Compare the accuracy of each sub model
6. Reduce the weight value of the submodel with the lowest accuracy by L
Returns the node position stored in the previous iteration round and adjusts the step size

FIGURE 5. Schematic diagram of the weight dynamic adaptive adjustment algorithm flow.

In equation (6), M_l represents the actual weight value of the model; F represents the actual number of models. The weight adjustment step size is also important for the model itself, and the actual fusion model will take too much time if it wants to reach the ideal convergence state, while setting it too large will make its change magnitude too large, which in the extreme case will lead to the emergence of the sub-model with weights greater than 1 and the other sub-model with a negative value. The weight adjustment step size calculation expression is shown in equation (7).

$$L = \frac{1}{F * V} \tag{7}$$

In equation (7), L denotes the weight adjustment step size; V denotes the actual number of model training. The weight adjuster is the core component of the module, which is responsible for the dynamic adaptive adjustment of the visual weight values of each submodel involved in model fusion, and mainly contains the weight adjustment strategy, the step size adjustment strategy and the weight adaptive adjustment algorithm. Among them, the weight adjustment equation in the weight adjustment strategy is expressed as shown in equation (8) and equation (9).

$$\begin{cases} M_l = M_l + L \\ M_l \end{cases} \tag{8}$$

The first line of equation (8) represents the computed expression of the weight values when the accuracy is maximized.

$$M_l = M_l - L \tag{9}$$

Equation (9) represents the expression of weight value calculation when the accuracy is the minimum value. The weight adaptive adjustment algorithm is the core algorithm in the weight adjustment and fusion module, based on the proposed algorithm can be realized in the actual self-training process of the model to adaptive adjustment of the weights, and its flow is shown in Figure 5.

As can be seen from Figure 5, for iteration round $o \in (1, 2, \cdots, V)$, the algorithm first builds a fusion model by training and validating each sub-model accordingly, and then composes a dictionary of the actual validation accuracy of each sub-model and the corresponding weight values. Secondly, it compares the accuracy of each sub-model,

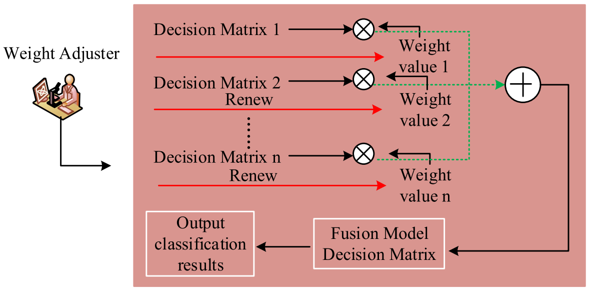


FIGURE 6. Schematic diagram of the actual working process of the weight fusion device.

increases the weight of the sub-model with the highest accuracy by L, and decreases the weight of the sub-model with the lowest accuracy by L, and then starts to determine that if the accuracy of the fusion model increases compared to the previous iteration after more than 2 iteration rounds of the experimental iteration, then it stores the actual position of the current node and continues to change in the original step size. If there is no rise then the node position stored in the previous iteration round is returned and the step size is adjusted and then the detection is performed with the changed small step size. Then the dictionary is updated to end the iteration and end the process. In this, the dictionary expression is constructed as shown in equation (10).

$$dict = (a_1 : M_1, a_2 : M_2, \cdots, a_F : M_F) \tag{10}$$

In equation (10), *dict* denotes the dictionary composed of the combination of accuracy and weight values; a denotes the accuracy rate. And in the step of algorithm step adjustment, the adjustment equation is expressed as shown in equation (11).

$$L = L * \alpha \tag{11}$$

In equation (11), α denotes the step decay rate, and its value is determined to be 0.5 in the actual experiments.In addition, the model fuser is the key of MMF, which is mainly used to generate the decision matrix of the final fusion model by fusing the decision matrices of the actual outputs of various sub-models in a corresponding weighted manner, and the actual flow of its work is shown in Figure 6.

As shown in Figure 6, the decision matrix for the final model is produced by first multiplying the weights by the actual output decision matrix of each sub-model to produce the weighted decision matrix of each sub-model. Where the final model decision matrix is expressed as shown in equation (12).

$$Q(u) = \sum_{k=1}^{m} M_k X_k = M_1 X_1 + M_2 X_2 + \cdots + M_m X_m \tag{12}$$

In equation (12), $Q(u)$ denotes the decision matrix of the fusion model; k denotes the serial number of the sub-model, m denotes the actual total number of sub-models; and X_k denotes the decision matrix of the actual output of the kth sub-model, which is actually expressed as shown in equation (13).

$$X_m = \begin{bmatrix} \gamma_{11} & \gamma_{21} & \cdots & \gamma_{b1} \\ \gamma_{12} & \gamma_{22} & \cdots & \gamma_{b2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{1c} & \gamma_{2c} & \cdots & \gamma_{bc} \end{bmatrix} \tag{13}$$

In equation (13), γ_{bc} represents the actual probability that sample c belongs to category b. Finally, the output module refers to the decision matrix from the weight adjustment and fusion module after the fusion of the maximum probability value by rows, this maximum probability value corresponding to the category of the column that is the fusion model of the sample's predicted classification, and after comparing with the real label, the final classification probability is output. In addition, the algorithm is evaluated from the classification effect and algorithm efficiency respectively in the experiment, in which the recall, accuracy, precision and F1 value are selected as the evaluation indexes for the classification effect, and the corresponding computational expressions are shown in equation (14) and equation (15).

$$\begin{cases} A = \dfrac{\varphi + \kappa}{\varphi + \psi + \kappa + \rho} \\ R = \dfrac{\varphi}{\varphi + \psi} \end{cases} \tag{14}$$

In equation (14), A denotes the accuracy rate; φ denotes the number of correctly classified positive samples; κ denotes the number of correctly classified negative samples; ψ denotes the number of negative samples misidentified as positive samples; ρ denotes the number of underreported positive samples; and R denotes the recall rate.

$$\begin{cases} \mu = \dfrac{\varphi}{\varphi + \rho} \\ F1 = \dfrac{2(\mu * R)}{\mu + R} \end{cases} \tag{15}$$

In equation (15), μ denotes the precision rate; $F1$ denotes the F1 value.

IV. PERFORMANCE ANALYSIS OF FGI CLASSIFICATION METHODS BASED ON MMF FOR ML

The study chose two additional traditional FGI classification datasets in the medical and agricultural domains, respectively, for trials to confirm the validity and generalizability of the research approach in FGI classification. Among them, the medical dataset is selected as Human Against Machine with 10000 training images (HAM10000) in the skin cancer related dataset. The HAM10000 data set is a fine-grained image data set with a simple background. It comes from the International Skin Imaging Society ISIC2018 Challenge. The data set contains 10,015 dermoscopic images of different people, divided into 7 categories. The cases basically include those in the field of pigmented lesions. All important diagnostic categories. And the cassava leaf disease (CLD) dataset is selected as the dataset in the agricultural field. The CLD dataset is a fine-grained image dataset with complex

TABLE 1. Original dataset, enhanced dataset and experimental environment content.

-	Original data set		Enhanced experimental dataset	
-	HAM10000	CLD	HAM10000	CLD
Types of	Fine grained images of human skin diseases	Fine grain image of cassava disease	Fine grained images of human skin diseases	Fine grain image of cassava disease
Number of images	10015	21367	51454	96322
Dataset partitioning	7	5	7	5
Category	-		8:1:1	8:1:1
Memory	32G		Graphics card	GeForce RTX 2080 Ti
Processor	Intel(R)Core(TM)i7-7800X CPU @ 3.40 GHz		Operating system	Ubuntu 18.0.4
Development environment	PyTorch		Software environment	CUDA11.0

background. The dataset has a total of 21,367 cassava disease images, divided into four disease categories and a fifth category representing healthy leaves. Most of the images were obtained from farmers taking photos of their own vegetable gardens and were annotated by professional institutions. This is a form that most truly represents the diagnosis that farmers need in real life, and has fine-grained features such as complex backgrounds, illumination changes, different angles, etc. It is more suitable for experimental laboratories than other widely used data sets in the field of plant diseases. Build the model. Due to the problem of uneven data distribution in the HAM10000 dataset and severe long-tailed distribution in the CLD dataset, the study utilizes data enhancement to process the two datasets. Table 1 displays a selection of them, including the original dataset, the augmented dataset, and the components of the experimental setting.

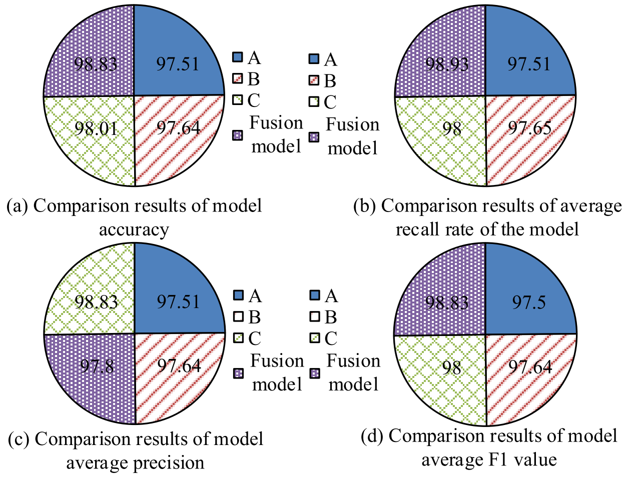
Table 1 shows that following data augmentation, the number of HAM10000 photos increases by 51,454 from 10015; also, the number of images in the CLD dataset doubles, going from 21367 to 96322 images. In addition, the categories before and after data enhancement of the two datasets remain unchanged, and the two are 7 and 5. Based on this foundation, the study was carried out to validate the comparisons from three perspectives, respectively, i.e. original dataset and enhanced dataset comparison, SM vs. MMF comparison, and different MMF methods comparison experiments. Among them, the comparison of the recognition results of the proposed fusion model of the study in the two datasets is shown in Table 2.

In Table 2, the accuracy rate of the HAM10000 dataset before data enhancement was 88.42%, the average recall rate was 76.63%, the average precision rate was 86.41%, and the average F1 value was 80.58%; whereas the values of the four metrics after data enhancement grew to 98.83%, and the recognition effect was significantly enhanced. In addition, the accuracy rate of CLD dataset before data enhancement was 86.60%, the average recall rate was 72.98%, the average precision rate was 77.97, and the average F1 value was 75.07%; while the values of the four metrics after data enhancement were 98.19%, 98.24%, 98.14%, and 98.18%, which were also significantly enhanced. Taken together, the operation of offline data augmentation positively affects the final recognition results of the fusion model, with the recognition accuracy metrics improving by 10.42% and other metrics improving by around 12.43% to 22.21% on the HAM10000 dataset in a simple context. While on the CLD dataset in complex context the recognition accuracy was improved by 11.60% and other metrics were improved by around 20.18% to 25.27%. Thus the operation of data enhancement not only makes the recognition accuracy of the fusion model effectively improved, but also makes the actual difference between the evaluation indicators become smaller. On this basis, the study compares the fine-grained image recognition performance of the three single models in the model library in Figure 4 with the fusion model proposed in the study on two datasets, and sets the three single models as A∼C. Set ResNet50 (the main feature of ResNet50 is the introduction of ''Residual Block'', which allows the network to better learn the differences between input and output, rather than directly learning output, which helps improve the performance of the model) to Model A and Efficient Network_ B0 (which uses the trained parameters, freezes the high-level, only fine tunes the classifier's parameters using the training set, and uses the entire model to recognize the test set) is set to B model, vit_ Base_ Patch16_ 224 is set as the C model. These three models are all experimental models and are officially released network models without any structural changes. They have all been pre trained on relevant datasets. All three models are already mature pre trained models. The recognition results on the HAM10000 dataset are shown in Figure 7.

TABLE 2. The recognition results of the fusion model proposed in the study on two datasets.

-	Original data set		After data enhancement	
-	HAM10000	CLD	HAM10000	CLD
Accuracy(%)	88.42	86.60	98.83	98.19
Average recall rate(%)	76.63	72.98	98.83	98.24
Average precise(%)	86.41	77.97	98.83	98.14
Average F1 value(%)	80.58	75.07	98.83	98.18

Detailed Figure 7 demonstrates that on the HAM10000 dataset, Model A has accuracy values of 97.51%, average recall values of 97.51%, average precision values of 97.51%, and average F1 values of 97.50%, whereas the values of the four metrics of Model B are 97.65%, 97.65%, 97.64%, and 97.64%, respectively. The values of the four metrics of Model C are 98.01%, 97.80%, 98.00%, and 98.00%, while the values of the four metrics of the fusion model proposed in the study are 98.83%, which are higher than the comparison models. When compared to models A, B, and C collectively, the fusion model's recognition accuracy is increased by 1.33%,



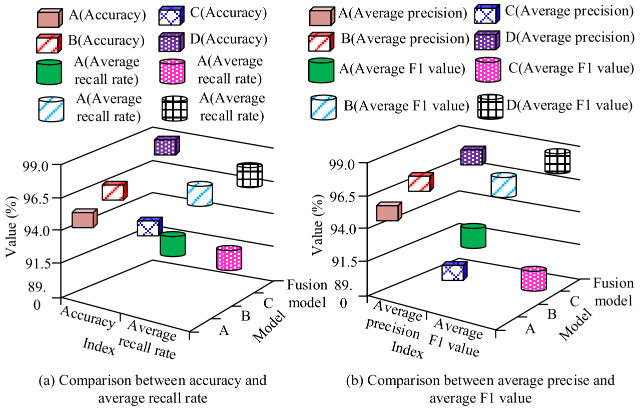
(a) Comparison results of model accuracy

(b) Comparison results of average recall rate of the model

(c) Comparison results of model average precision

(d) Comparison results of model average F1 value

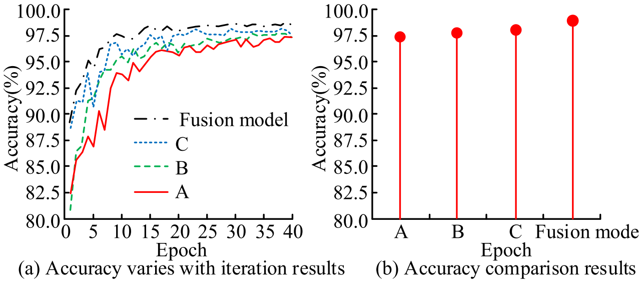
FIGURE 7. Recognition results of four models on the HAM10000 dataset.



(a) Comparison between accuracy and average recall rate

(b) Comparison between average precise and average F1 value

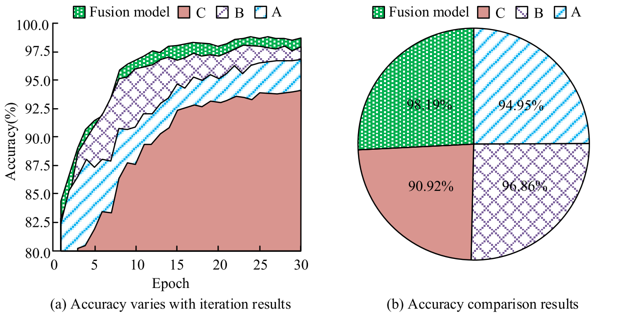
FIGURE 8. Recognition results of four models on the CLD dataset.



(a) Accuracy varies with iteration results

(b) Accuracy comparison results

FIGURE 9. The results of accuracy changes with iteration for four models on the HAM10000 dataset.



(a) Accuracy varies with iteration results

(b) Accuracy comparison results

FIGURE 10. The accuracy of four models on the dataset CLD varies with iteration results.

1.19%, and 0.83%, while the other three metrics are improved by 0.84% to 1.33%, 0.84% to 1.33%, and 0.84% to 1.34%, respectively, when compared to the SM. In addition, the recognition results on the CLD dataset are shown in Figure 8.

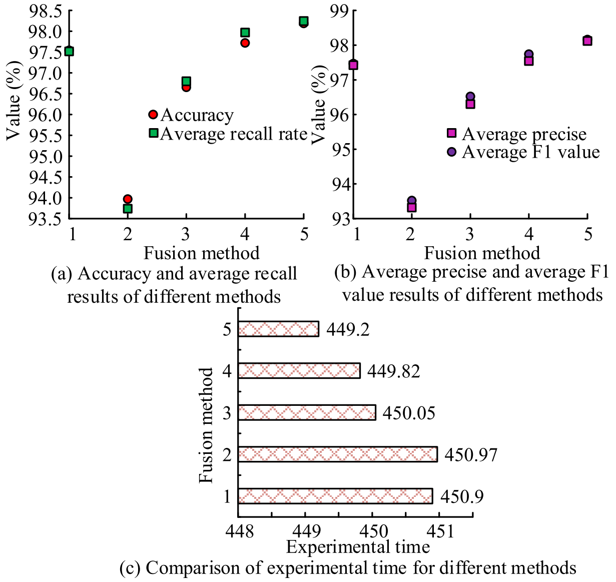
Comprehensive Figure 8 demonstrates that on the CLD dataset, Model A has accuracy values of 94.95%, average recall levels of 94.46%, average precision values of 95.53%, and average F1 values of 94.49%. In contrast, Model B has accuracy values of 96.86%, 97.11%, 96.74%, and 96.90% for each of the four metrics. The values of the four indicators of Model C were 90.92%, 89.77%, 89.85%, and 89.80%, while the values of the four indicators of the fusion model proposed in the study were 98.19%, 98.24%, 98.14%, and 98.18%, which were higher than those of the comparison model. In conclusion, while the other three metrics are improved by 1.14%-8.48%, 1.41%-8.30%, and 1.38%-8.39%, respectively, compared to SM, the recognition accuracy of the fusion model is enhanced by 3.25%, 1.34%, and 7.28% compared to models A, B, and C, respectively. The fusion model outperforms the comparison model and has the best picture recognition result, as can be demonstrated by combining Figure 7 and Figure 8. In order to visualize the difference between the SM and fusion models, the study continued by comparing the results of the four models on two datasets in terms of the change in accuracy with iterations. One of the results on the dataset HAM10000 is shown in Figure 9.

As can be seen from the combined Figure 9, due to the use of ML techniques for the four models, the accuracy improvement is realized quickly and the models reach convergence with only a small number of iterations. Additionally, the fusion model's accuracy curve is always above the comparison model and exhibits minimal fluctuation. For instance, when the number of iterations is between 5 and 10, models A and C exhibit significant fluctuation while model B is comparatively smooth, the fusion model exhibits minimal fluctuation. This indicates that the fusion model successfully combines the traits of the three SMs to improve recognition. Taken together, on the HAM10000 dataset, the fusion model proposed by the study outperforms the SMs in all four indicators, indicating the effectiveness of the research method. And Figure 10 displays the outcomes for the dataset CLD.

Comprehensive Figure 10 shows that model C has the best actual recognition effect among the three SMs when on the dataset HAM10000, but the worst performance on the CLD dataset, indicating that model C has an advantage in recognizing data samples with simple backgrounds, but not in recognizing image samples with complex background noise. The fusion model converges faster in the case of growing number of iterations, and basically starts to stabilize when the number of iterations is around 8 and is used higher than the comparison model, which proves that the fusion model has better robustness. Combining Figure 9 and Figure 10,

TABLE 3. Comparison of recognition effect and algorithmic efficiency on data set HAM10000.

-	SP	PP	SMV	SMP	Research method
Accuracy(%)	98.65	98.22	98.40	98.40	98.83
Average recall rate(%)	98.65	98.22	98.40	98.40	98.83
Average precise(%)	98.65	98.23	98.40	98.39	98.83
Average F1 value(%)	98.65	98.22	98.39	98.39	98.83
Experimental time	282.05min	282.80min	286.72min	287.45min	284.2min



(a) Accuracy and average recall results of different methods

(b) Average precise and average F1 value results of different methods

(c) Comparison of experimental time for different methods

FIGURE 11. Comparison results of recognition effect and Algorithmic efficiency on data set CLD.

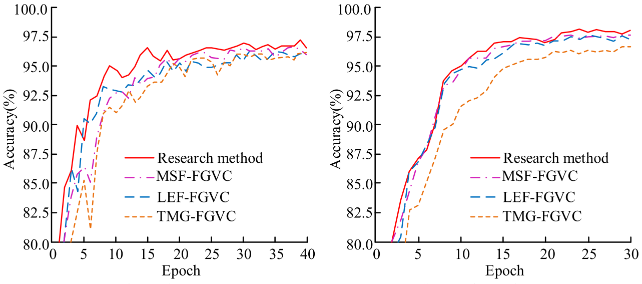
it can be seen that the fusion model exhibits high recognition performance despite the significant variation in the actual recognition performance of individual sub-models. This suggests that it is less impacted by the actual performance of individual sub-models that exhibit poor classification performance, while at the same time, it consistently outperforms other models in the recognition and classification of both simple and complex background images. The findings also demonstrate the superiority of the proposed DWFM technique over SM in the application of fusing simple network models using machine learning, as well as its superior robustness, demonstrating the efficacy of the study methodology on such network models.

Finally, to increase the reliability of the test, the experiment was repeated multiple times for each method on two data sets, and the algorithm efficiency, actual recognition effect and accuracy were analyzed. The maximum value of the three performance indicators is 100%. The higher the values obtained by different methods, the better the relative performance of the method.

The study compared the actual effects of different multi-model fusion methods in two fine-grained images. The comparative fusion methods are the Sum of the Probabilities (SP), the Product of the Probabilities (PP), and Sum of the Maximal Probabilities (SMP) and Simple Majority Voting (SMV). The recognition effect and algorithm efficiency results on the data set HAM10000 are shown in Table 3.

As can be seen in Table 3, on the dataset HAM10000 the accuracy, average recall, average precision, and average F1 value of the SP method is 98.65%, and the values of the four metrics of the PP method are 98.22%, 98.22%, 98.23%, and 98.22%. the values of the four metrics of the SMV method are 98.40%, 98.40%, 98.39%, and 98.39%, and the values of the four metrics of the SMP method are 98.40%, 98.40%, 98.39%, and 98.39%, respectively, 98.40%, and 98.39%, and the values of the four indicators for the SMP method were 98.40%, 98.40%, 98.39%, and 98.39%, respectively. Whereas, the values of all the research methods were 98.83%, which was significantly higher than the comparison methods. The fusion model recognition accuracy is improved by 0.18%, 0.61%, 0.43%, and 0.43% for the research method compared to the comparison methods, while it is improved by 0.18% to 0.61%, 0.18% to 0.60, and 0.18% to 0.61% in the other three metrics, respectively. In addition, the experimental times of the four comparison methods are 282.05 min, 282.80 min, 286.72 min, and 287.45 min, respectively, while the experimental time of the research method is in the middle at 284.2 min. Taken as a whole, the fusion model constructed by the research method has a better practical performance, while the experimental time consumed is 3.25 min less than that of SMP, and the four fusion algorithms' overall time is not much different, and the efficiency of each model is basically the same. And the recognition effect and algorithm efficiency on the dataset CLD are shown in Figure 11.

In Figure 11, 1 to 5 denote SP, PP, SMV, SMP, and research methods, respectively. Combining Figure 11, it can be seen that the values of the four indicators of the SP method on the dataset CLD are 97.54%, 97.50%, 97.43%, and 97.46%, respectively; the values of the PP method are 93.97%, 93.74%, 93.32%, and 93.52%, respectively. the values of the SMV method are 96.67%, 96.77%, respectively, 96.30%, and 96.52%, and the values for the SMP method were 97.72%, 97.97%, 97.54%, and 97.74%, respectively. Whereas, the values of 98.19%, 98.24%, 98.14%, and 98.18% for the research methods were higher than those of the comparison methods. In addition, the experimental time of the four comparison methods was 450.9 min, 450.97 min, 450.05 min, and 449.82 min, while the experimental time of the research method was 449.2 min, which was significantly lower than the comparison methods. Taken together, the recognition accuracy of the fusion model utilizing the research method is higher than the comparison method by 0.68%, 4.22%, 0.47%, and 1.52%, while the other three metrics are improved by 0.27% to 4.50%, 0.48% to 4.82%,



(a) Results on dataset HAM10000

(b) Results on dataset CLD

FIGURE 12. The accuracy of four methods varies with iteration on two datasets.

and 0.44% to 4.66%, respectively. The experimental time is also lower than the comparison method, indicating higher performance. Similarly, in order to see the actual effect of the research method more intuitively, the study compared the research method with the three most advanced methods currently. These three methods are respectively a fine-grained image classification method based on multi-scale feature fusion (MSF-FGVC), a fine-grained image classification method based on learning enhanced features and double inference (LEF-FGVC), and a trusted multi-granular information fusion method. Fine-grained image classification method (TMG-FGVC) [19], [20], [21]. The accuracy changes with iterations of the research method and three advanced methods on the two data sets were analyzed, and the results are shown in Figure 12.

Comprehensive Figure 12 shows that the trend on the dataset HAM10000 shows that the fusion model under the research method can reach the fitting state faster compared to other model fusion methods, which is more obvious when the number of iterations is in the range of 8 to 15. When the number of iterations is 15, the recognition accuracy of the fusion model under the research method has reached 98.51%, compared with the highest accuracy of 0.32% of the golden cross, while the model under the other methods in the same period maintains between 96% and 97%, significantly lower than the fusion model under the research method. The trend on the CLD of the dataset can be seen that there is a phase intersection between the MSF-FGVC and LEF-FGVC methods and the research method, which performs better among the four compared methods. Compared with HAM10000, the dataset CLD has a higher complexity and is closer to FGI recognition in the real environment, so different models show significant differences in actual sample classification due to the influence of the environment and external conditions, which in turn leads to the fusion of different fusion methods due to the differences in fusion mechanisms and also shows a large difference in the fusion of decision matrices. Taken together, the fusion model under the proposed method has a slight advantage on the dataset HAM10000 and is able to utilize fewer iterations to achieve the highest level of recognition, which effectively saves time costs. Meanwhile, it shows better results on the dataset CLD, which proves that the research method has a better GC.

V. CONCLUSION

To improve the problem of insufficient generalization ability of the current single model method and the solidification of weights of the traditional multi-model fusion method, the experiment proposes a multi-model fusion fine-grained image classification method based on transfer learning. On the basis of the existing multi-model fusion method, the fine-grained image classification method of multi-model fusion is improved. At the same time, in order to adapt to the characteristics of fine-grained images, transfer learning and adding attention mechanisms are adopted for the network models used. Improve. The results show that in the comparison of fusion methods, the four index values of the fusion model under the research method on the data set HAM10000 are 98.83%; the values of the fusion model under the research method on the data set CLD are 98.19%, 98.24%, 98.14% and 98.18 respectively. %, are higher than the comparative method, and the actual time consumption is lower than the comparative method. In addition, the accuracy curve of the fusion model is always above the comparison model with smaller fluctuations, and converges faster as the number of iterations continues to increase. In the accuracy comparison, on the data set HAM10000, when the number of iterations is 15, the recognition accuracy of the fusion model under the research method has reached 98.51%, while the models under other methods during the same period remained between 96% and 97%, which is obvious. The fusion model is lower than the research method. On the dataset CLD, the accuracy of the research method has been significantly higher. In summary, it can be seen that the fusion method proposed in the study can reduce the time cost of the model while effectively improving the recognition and classification performance, and has good generalization ability. The research method has very significant competitiveness, and has more obvious advantages when facing fine-grained image data sets with complex backgrounds, which further reflects the huge application value of the research method. However, the experimental method relies on a large amount of rich and diverse pre-training data, which limits its application scope to a certain extent. For those fields where large-scale annotated data is difficult to obtain, experimental methods may not achieve optimal performance. In addition, although the multi-model fusion strategy improves the accuracy of classification, it also increases the complexity of the model and the demand for computing resources. This may lead to difficulties in practical application in resource-constrained environments. If the source data set is too different from the target task, the effect of transfer learning may be reduced. In order to deal with these limitations and potential error sources, future research will start from the following three aspects: first, explore more effective data enhancement and transfer learning strategies to reduce reliance on large-scale pre-training data. Second, study more lightweight model fusion methods to reduce computational costs and improve model usability. Third, test the generalization ability of the method in different fields and tasks to further optimize the migration strategy.

REFERENCES

[1] Y. Yang and X. Song, "Research on face intelligent perception technology integrating deep learning under different illumination intensities," *J. Comput. Cognit. Eng.*, vol. 1, no. 1, pp. 32–36, Jan. 2022, doi: 10.47852/bonviewjcce19919.

[2] H. Huang, J. Zhang, L. Yu, J. Zhang, Q. Wu, and C. Xu, "TOAN: Target-oriented alignment network for fine-grained image categorization with few labeled samples," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 2, pp. 853–866, Feb. 2022, doi: 10.1109/TCSVT.2021.3065693.

[3] P. Koniusz and H. Zhang, "Power normalizations in fine-grained image, few-shot image and graph classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 591–609, Feb. 2022, doi: 10.1109/TPAMI.2021.3107164.

[4] S. A. Bargal, A. Zunino, V. Petsiuk, J. Zhang, K. Saenko, V. Murino, and S. Sclaroff, "Guided zoom: Zooming into network evidence to refine fine-grained model decisions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4196–4202, Nov. 2021, doi: 10.1109/TPAMI.2021.3054303.

[5] A. Abusnaina, M. Abuhamad, H. Alasmary, A. Anwar, R. Jang, S. Salem, D. Nyang, and D. Mohaisen, "DL-FHMC: Deep learning-based fine-grained hierarchical learning approach for robust malware classification," *IEEE Trans. Dependable Secure Comput.*, vol. 19, no. 5, pp. 3432–3447, Sep. 2022, doi: 10.1109/TDSC.2021.3097296.

[6] Q. Zheng, P. Zhao, H. Wang, A. Elhanashi, and S. Saponara, "Fine-grained modulation classification using multi-scale radio transformer with dual-channel representation," *IEEE Commun. Lett.*, vol. 26, no. 6, pp. 1298–1302, Jun. 2022, doi: 10.1109/LCOMM.2022.3145647.

[7] X. Wang, J. Shi, H. Fujita, and Y. Zhao, "Aggregate attention module for fine-grained image classification," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 7, pp. 8335–8345, Jul. 2023, doi: 10.1007/s12652-021-03599-7.

[8] X.-S. Wei, Y.-Z. Song, O. M. Aodha, J. Wu, Y. Peng, J. Tang, J. Yang, and S. Belongie, "Fine-grained image analysis with deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8927–8948, Dec. 2022, doi: 10.1109/TPAMI.2021.3126648.

[9] K. Adem, M. M. Ozguven, and Z. Altas, "A sugar beet leaf disease classification method based on image processing and deep learning," *Multimedia Tools Appl.*, vol. 82, no. 8, pp. 12577–12594, Mar. 2023, doi: 10.1007/s11042-022-13925-6.

[10] L. C. Ngugi, M. Abelwahab, and M. Abo-Zahhad, "Recent advances in image processing techniques for automated leaf pest and disease recognition—A review," *Inf. Process. Agricult.*, vol. 8, no. 1, pp. 27–51, Mar. 2021, doi: 10.1016/j.inpa.2020.04.004.

[11] X. Chen, K. Hirota, Y. Dai, and Z. Jia, "A model fusion method based on multi-source heterogeneous data for stock trading signal prediction," *Soft Comput.*, vol. 27, no. 10, pp. 6587–6611, May 2023, doi: 10.1007/s00500-022-07714-4.

[12] X. Liu, H. Wang, J. Liu, S. Sun, and M. Fu, "HSI classification based on multimodal CNN and shadow enhance by DSR spatial–spectral fusion," *Can. J. Remote Sens.*, vol. 47, no. 6, pp. 773–789, Aug. 2021, doi: 10.1080/07038992.2021.1960810.

[13] J. Jenisha and A. J. Dickson, "Automated liver tumor segmentation using deep transfer learning and attention mechanisms," *EPRA Int. J. Res. Develop.*, vol. 8, no. 7, pp. 144–150, Jul. 2021, doi: 10.36713/epra2016.

[14] M. T. Vo, A. H. Vo, and T. Le, "A robust framework for shoulder implant X-ray image classification," *Data Technol. Appl.*, vol. 56, no. 3, pp. 447–460, Nov. 2021, doi: 10.1108/dta-08-2021-0210.

[15] H.-Y. Chien, Y.-C. Wang, and G.-C. Chen, "Application of image recognition in workpiece classification," *Adv. Mech. Eng.*, vol. 13, no. 6, Jun. 2021, Art. no. 168781402110260, doi: 10.1177/16878140211026082.

[16] R. Guo, Y. Zhou, J. Zhao, Y. Man, M. Liu, R. Yao, and B. Liu, "Point cloud classification by dynamic graph CNN with adaptive feature fusion," *IET Comput. Vis.*, vol. 15, no. 3, pp. 235–244, Apr. 2021, doi: 10.1049/cvi2.12039.

[17] C.-C. Lin, C.-H. Kuo, and H.-T. Chiang, "CNN-based classification for point cloud object with bearing angle image," *IEEE Sensors J.*, vol. 22, no. 1, pp. 1003–1011, Jan. 2022, doi: 10.1109/JSEN.2021.3130268.

[18] X. Zhao, P. Huang, and X. Shu, "Wavelet-attention CNN for image classification," *Multimedia Syst.*, vol. 28, no. 3, pp. 915–924, Jan. 2022, doi: 10.1007/s00530-022-00889-8.

[19] Y. Shang and H. Huo, "MSFF: Multi-scale feature fusion for fine-grained image classification," *Academic J. Comput. Inf. Sci.*, vol. 6, no. 2, pp. 109–119, Feb. 2023, doi: 10.25236/AJCIS.2023.060215.

[20] X. Nie, B. Chai, L. Wang, Q. Liao, and M. Xu, "Learning enhanced features and inferring twice for fine-grained image classification," *Multimedia Tools Appl.*, vol. 82, no. 10, pp. 14799–14813, Apr. 2023, doi: 10.1007/s11042-022-13619-z.

[21] Y. Yu, H. Tang, J. Qian, Z. Zhu, Z. Cai, and J. Lv, "Fine-grained image recognition via trusted multi-granularity information fusion," *Int. J. Mach. Learn. Cybern.*, vol. 14, no. 4, pp. 1105–1117, Apr. 2023, doi: 10.1007/s13042-022-01685-6.

WENYING ZHANG received the B.E. degree from Southwest Jiaotong University, in 1994, and the M.E. degree in information and communication systems from Zhengzhou University, in 2002. She is currently a Lecturer with the School of Electrical and Information Engineering, Zhengzhou University. Her research interests include signal and information processing, image processing, and pattern recognition.

YAPING WANG received the B.E. degree in automation specialty and the M.E. and D.E. degrees in control science and engineering from Northwestern Polytechnical University, in 2004, 2007, and 2013, respectively. She is currently an Associate Professor with the School of Electrical and Information Engineering, Zhengzhou University. Her research interest includes image processing and analysis.