High-precision Image Classification Algorithm based on Attention Mechanism and Multi-scale Features

Ling Yang, Liantian Li*
Department of Information Engineering, Yangjiang Polytechnic, Yangjiang 529566, China
E-mail: li2020202207@126.com
*Corresponding author

Keywords: Attention mechanism, multi-scale features, convolutional neural network, high precision image classification, feature pyramid

Received: May 23, 2025

High-precision image classification has steadily emerged as a key area of research interest due to the extensive use of image classification technologies in many different domains. The study enhances the conventional feature pyramid networks (FPN) and suggests a high-precision image classification model in an attempt to further increase the precision and effectiveness of picture classification. The model enhances the ability of convolutional neural network (CNN) to focus on key information by combining the channel attention and spatial attention mechanisms. The outcomes indicated that the improved CNN model achieved 77.50% classification accuracy on the ImageNet dataset and 94.20% on the CIFAR-10 dataset, which was significantly higher than the control model. In addition, in the classification of different types of high-precision images, the improved CNN model performed well in the recall, F1 score, and robustness metrics. Their values were 94.3%, 94.6%, and 93.5%, respectively. The results show that the high-precision image classification model is able to capture the key features and detail information in the image more effectively, which significantly improves the classification accuracy and robustness. This study provides a new technical tool for high-precision image classification tasks.

Povzetek: Raziskave računalniškega vida so avtorji izvedli s pomočjo CNN za klasifikacijo slik z izboljšanim FPN (utežena večsklopna fuzija, prilagoditvena konvolucija) in hibridno kanalno-prostorsko pozornostjo. Validacija je narejena na ImageNet/CIFAR-10/medicinskih naborih.

1 Introduction

One of the fundamental tasks of computer vision, picture classification has extensive use in a variety of domains, including automatic driving, remote sensing image analysis, and medical image diagnostics. In these fields, high-precision image classification is crucial for improving work efficiency, reducing errors and enhancing safety [1-2]. For example, in medical image diagnosis, accurate identification of lesion areas is crucial for early diagnosis and treatment. In autonomous driving, high-precision image classification (IC) can help vehicles better recognize road signs and obstacles, thus improving driving safety. Convolutional neural networks (CNNs) have emerged as one of the primary methods for modern IC due to their impressive performance in IC tasks in recent years. By automatically extracting features of an image, CNN can effectively capture local and global information in an image, thus realizing high-precision classification [3-4]. However, despite its excellent performance in IC, CNN still has some limitations. Traditional CNN models are often difficult to effectively focus on key information in images when dealing with complex image scenes, resulting in limited classification accuracy (CA). In addition, CNNs usually only capture single-scale features in the feature extraction (FE) process, ignoring the multi-scale information in the image [5-6]. FE process is a process of gradually extracting local and global features of the image.

Some scholars have also used CNN model to classify images at this stage. Wu et al. proposed an improved CNN model for multi-label medical IC. The model consisted of three main components of CNN and Transformer branch: multi-label multi-head attention-enhanced feature module, multi-branch residual module, and information interaction module. The results indicated that the framework demonstrated good performance on multiple publicly available datasets with good generalization ability and was applicable to other medical multi-label IC tasks [7]. Alkhatib M et al. proposed a model called Improved CNN model to address the lack of training samples in hyperspectral IC and the failure of traditional CNN to fully utilize the correlation between hyperspectral image bands. The results indicated that it outperformed existing methods in terms of overall accuracy, average accuracy, and Kappa coefficient, and obtained near-optimal classification results even with a small number of training samples [8]. Han et al. proposed a dynamic multi-scale CNN model for the current situation of insufficient feature information extracted by CNN and inaccurate attention weights in medical IC. The results indicated that the model achieved most advanced classification performance and solved the uncertainty quantization problem on publicly available datasets from four different medical domains [9]. The specific retrospective analysis of the above-mentioned literature is shown in Table 1.

Table 1: A specific review and analysis of the literature

References	Method	Advantages	Disadvantages
Reference 7	CTransCNN is a hybrid deep learning model that combines CNN and Transformer, specifically designed for multi-label medical image classification. The model consists of a multi-label multi-head attention enhancement feature module, a multi-branch residual module and an information interaction module.	The implicit correlation between labels is automatically captured, eliminating the need to manually predefine label relationships. A cross-attention mechanism is introduced to allow the model to weight image features according to the importance of each label. Effective fusion of local and global features is achieved through the information interaction module. The feature representation ability of the model is optimized through the embedded and externally embedded residual structures, and the number of parameters is reduced.	The structure of the model is relatively complex, resulting in a long training and reasoning time. The performance of the model depends to a certain extent on the quality and scale of the data set. Due to the complexity of the model, deploying it to mobile devices or other resource-constrained environments may face challenges.
Reference 8	Tri-CNN first uses PCA for dimensionality reduction, and then inputs the data into three branches. Each branch uses 3D-CNN of different scales to extract features. The features extracted from the three branches are flattened and tiled, then classified through the fully connected layer and the softmax layer, and trained using the cross-entropy loss function.	Multi-scale FE can make more comprehensive use of the multi-dimensional information of hyperspectral images. Through feature fusion, the model can better capture features of different scales and types, and improve the classification performance. The model performs well on multiple datasets, demonstrating good adaptability and generalization ability.	Since the model contains multiple branches and multi-scale 3D-CNNs, the computational complexity is relatively high. Although the model performs well on small sample datasets, its performance may depend on a sufficient number of training samples. The complexity of the model may lead to deployment difficulties in practical applications, especially in resource-constrained environments.
Reference 9	The DM-CNN model introduces a dynamic multi-scale feature fusion module, a hierarchical dynamic uncertainty quantization attention mechanism, a multi-scale fusion pooling method, and a multi-objective loss optimization network structure for medical image classification.	The model is capable of extracting feature information at different scales. The attention mechanism can dynamically adjust the attention weights according to different information in each layer, enabling the model to better focus on important feature information. The pooling method can accelerate the computing speed and prevent overfitting while retaining the main and important information. Multi-objective loss can better balance the training process of the model and improve the convergence speed and classification performance of the model	The training and reasoning time of the model is relatively long. Performance may depend on sufficient training samples. Scalability may be limited

The diversity of visual data makes it challenging for a single FE method to match the demand for high-precision classification, even though the aforementioned research has produced superior outcomes. Based on this research, feature pyramid networks (FPN) is improved. Meanwhile, it improves CNN based on attention mechanism (AM) and multi-scale features (MSFs). The research aims to enhance the CNN's ability to focus on key information by introducing an AM and a MSF extraction method, as a way to improve the accuracy and efficiency of IC. The innovation of the study is the introduction of weighted fusion mechanism and adaptive feature adjustment strategy to improve the FPN. Meanwhile, by combining channel attention and spatial attention, the hybrid AM is designed to improve the CNN's capacity to concentrate on important information and increase CA.

2 Methods and materials

2.1 Efficient FE method based on MSF fusion

In IC tasks, FE is one of the key steps to determine the CA. By combining features of several scales, MSFs can enhance the accuracy of IC and collect both global and specific information about an image [10-11]. FE is crucial to the accuracy of image classification. MSFs simultaneously capture image details and global information, improving CA by fusing features of different scales. Multi-scale FE is a key technology for high-precision image classification. Traditional FPN, as a classic architecture, is widely used in object detection and image classification. Its structure is shown in Figure 1.

In the task of image classification, multi-scale FE is one of the key technologies for improving CA. To better understand the multi-scale FE process, the study first introduces the FPN, as shown in Figure 1. The core idea of the traditional feature pyramid lies in combining high-level semantic information with low-level detail information through a top-down path. This generates a multi-scale FPN with rich semantics and details, which effectively captures the multi-scale information of the image [12-13]. The traditional FPN structure can be mainly divided into FE, top-down path, and the generation process of FPN. The key steps of the process are shown in Fig. 2.

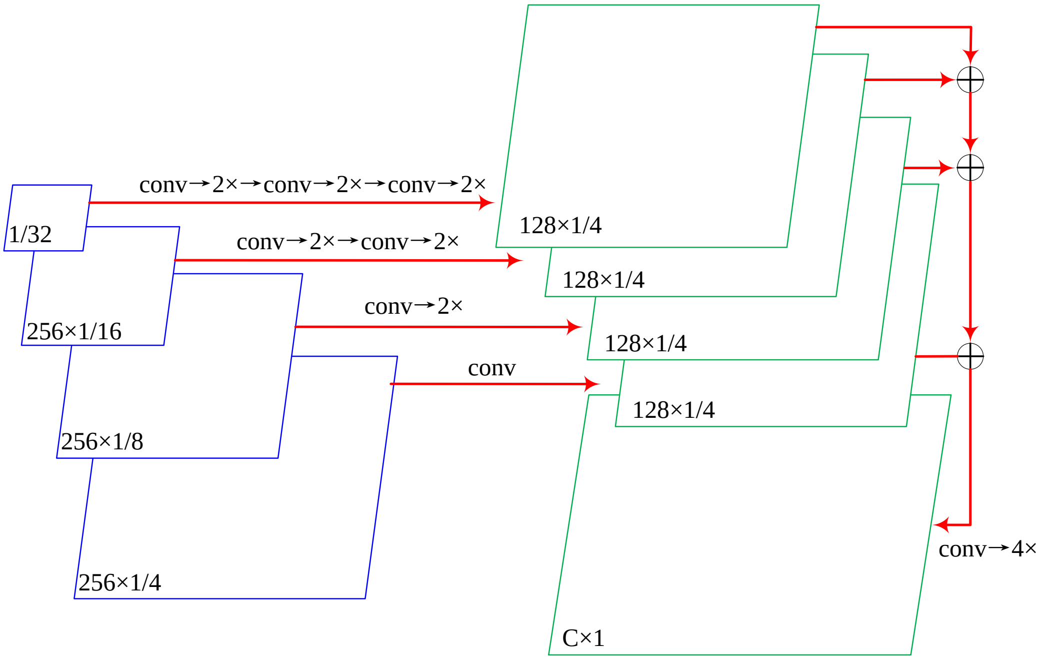


Figure 1: Schematic diagram of the MSF pyramid structure

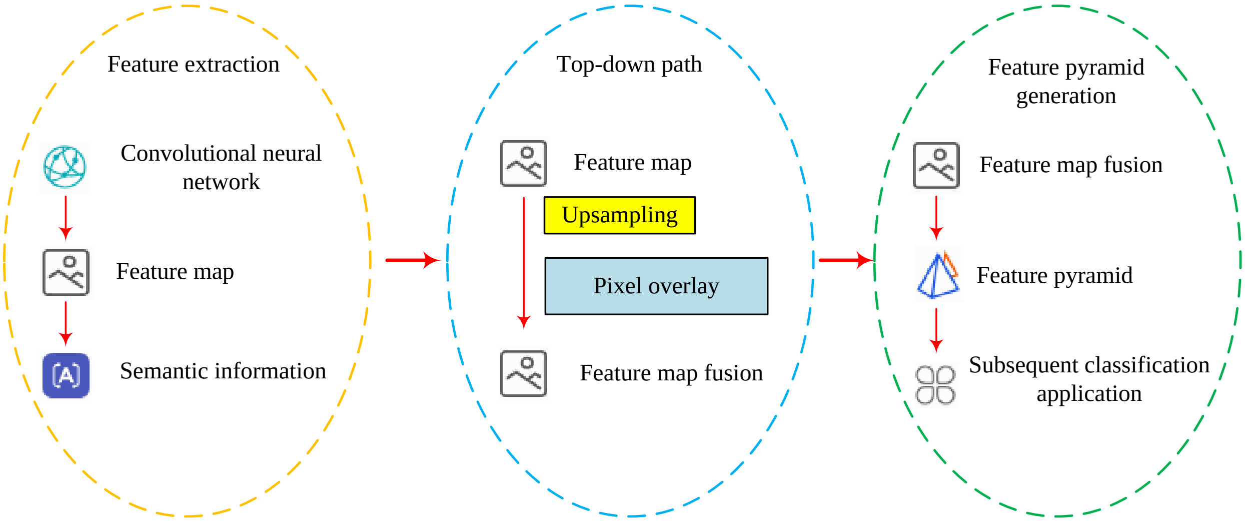


Figure 2: The generation process of the feature pyramid

In Fig. 2, the first is the FE phase, which is based on the first few layers of the CNN to extract different levels of the feature map (FM) $F_1, F_2, \cdots F_n$. Among them, F_i denotes the FM of layer i. Moreover, the resolution of F_i gradually decreases and the semantics gradually increases. FM is the result of the convolution operation. It is a two-dimensional or three-dimensional that represents how the input image (InI) responds to a specific convolution kernel. This is followed by a top-down path stage. That is, starting from the high-level FM, the FM is gradually up-sampled. Meanwhile, pixel-by-pixel summation is performed with the bottom layer FMs, so as to generate a series of fused feature maps (FFMs). For the FM F_i of the i th layer, the mathematical expression of the FFM is shown in Equation (1).

$$F_i^{\sim} = F_i + U(F_{i+1}^{\sim}) \quad (1)$$

In Equation (1), $F_i^{\sim}$ denotes the FFM. U denotes the up-sampling operation. The FFMs are generated through upsampling operations. First, each FM is averaged and pooled globally to generate a one-dimensional vector (1DV). Then, the vector is input into the fully connected layer (FCL) to calculate the weights. Finally, it is the generation stage of FPN. The FFMs are composed into a feature pyramid, which is used for subsequent classification tasks. Although the traditional FPN has achieved remarkable results in MSF extraction, it has some shortcomings in the feature fusion process. For example, the simple pixel-by-pixel summing operation does not take into account the difference in importance of features at different scales. To overcome these limitations, the study proposes an improved FPN network. First, a weighted fusion mechanism is introduced. That is, a weight is assigned to each scale in the process of feature fusion as a way to highlight the feature information with higher value [14-15]. Defining the FM weight of layer i as ω_i , then the mathematical expression of the FFM is shown in Equation (2).

$$F_i^{\sim} = \omega_i \cdot F_i + U(F_{i+1}^{\sim}) \qquad (2)$$

In Equation (2), the weights ω_i are dynamically computed by a weight learning module, which is implemented through global average pooling (GAP) as well as a FCL. A GAP operation is first performed on each FM F_i to obtain a 1DV. The related mathematical expression is shown in Equation (3).

$$v_i = P_a(F_i) \qquad (3)$$

In Equation (3), the corresponding weight information is obtained after inputting a IDV v_i into the fully connected form, as shown in Equation (4).

$$\omega_i = \mathrm{Re}LU(FC(v_i)) \qquad (4)$$

In Equation (4), FC means the FCL. $\mathrm{Re}LU$ denotes the activation function (AF). After introducing the weighted fusion mechanism to optimize the FPN, in order to further the effect of feature fusion, the study introduces an adaptive adjustment module on the FFM. This module optimally adjusts the fused features through a convolutional layer (CL) as a way to enhance the expression of the relevant features [16]. Specifically based on the FFM $F_i^{\sim}$, the mathematical expression of the FM through optimization and adjustment is shown in Equation (5).

$$F_i^{adjust} = Conv(F_i^{\sim}) \qquad (5)$$

In Equation (5), F_i^{adjust} denotes the FM after adjustment and optimization. By introducing the weighted fusion mechanism as well as the adaptive feature adjustment strategy, the FPN network structure proposed in the study is able to fuse MSFs more effectively. Meanwhile, more valuable feature information is highlighted, which in turn improves the performance of IC.

2.2 Construction of IC model with improved CNN based on AM

The introduction of weighted fusion mechanism and adaptive feature tuning strategy through improved FPN has effectively enhanced the performance of FE. However, in complex image scenes, certain regions or channels may be more critical to the classification task. To further improve the accuracy of IC, this study introduces an AM to enhance the CNN's ability to focus on critical information on the basis of FE. The primary function of CNN's central convolutional operation is to extract an image's local features. The related schematic is shown in Fig. 3.

The convolution operation is the core of convolutional neural networks and is primarily used to extract the local features of images. Figure 3 illustrates the specific operation flow to better understand the process of convolution operation. In Fig. 3, the InI is defined as $I \in R^{H \times W \times C}$. Among them, H , W , and C denote the height, width, and quantity of channels of the image. Meanwhile, the convolution kernel $K \in R^{k \times k \times C}$ is slid over the InI and after multiplying and summing by element-by-element, this generates the FM F , as expressed in Equation (6).

$$F_{ij} = \sum_{m=0}^{k-1} \sum_{n=0}^{k-1} \sum_{c=0}^{C-1} K_{m,n,c} \cdot I_{i+m, j+n, c} + b \qquad (6)$$

In Equation (6), b means the bias term. F_{ij} means the value of the output FM at position (i, j) . k means the size of the convolution kernel. The primary purpose of pooling procedures is to decrease the FM's spatial dimension (SD) and computational burden. Both maximum pooling (MP) and average pooling (AP) are common pooling operations (POs). The related schematic is shown in Fig. 4.

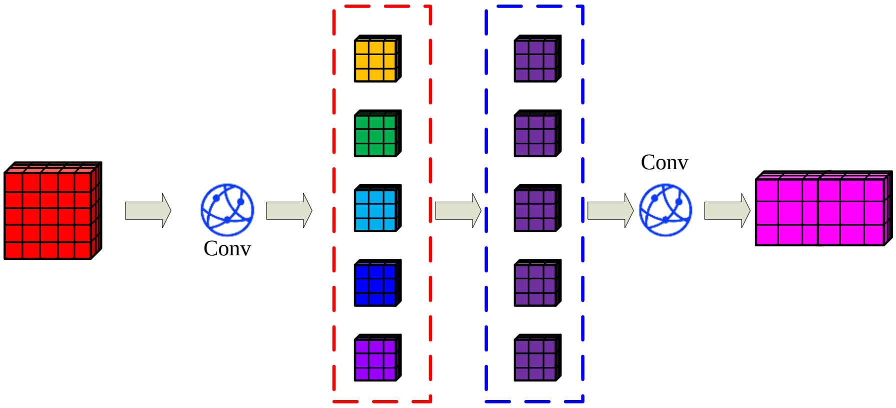
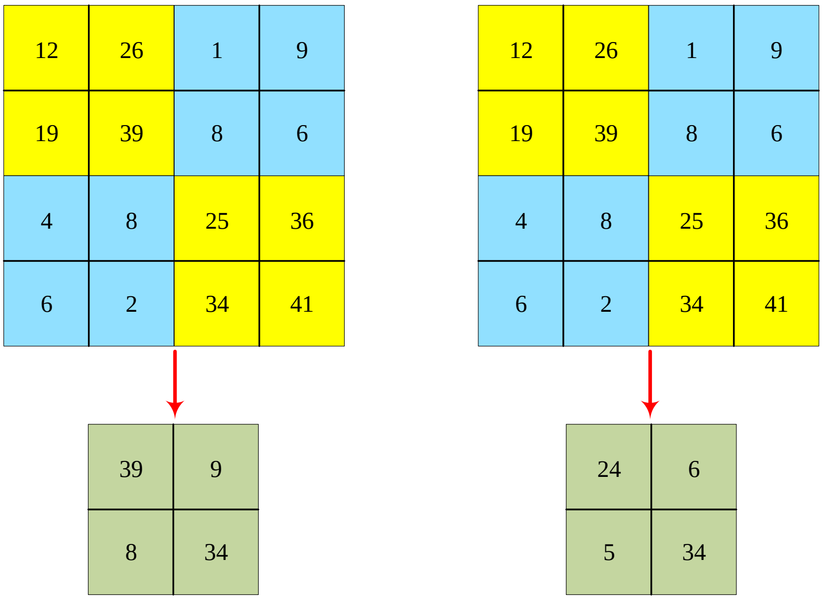


Figure 3: Schematic diagram of CO



(a) Schematic diagram of maximum pooling

(b) Average pooling schematic diagram

Figure 4: Schematic diagram of PO

The PO is an important step in CNNs. It is mainly used to reduce the SD of FMs and decrease the computational load. To present the process of POs more intuitively, the study illustrates the specific pooling process through Figure 4. Fig. 4(a) and Fig. 4(b) illustrate the process of MP as well as AP, respectively. For the input FM, the study uses a combination of MP and AP, i.e., for the input F. The mathematical expression related to MP and AP is shown in Equation (7).

$$\begin{cases} F_{\max,i,j} = \max\limits_{m=0}^{p-1} \max\limits_{n=0}^{p-1} F_{i,p+m,j\cdot p+n} \\ F_{avg,i,j} = \dfrac{1}{p^2} \sum\limits_{m=0}^{p-1} \sum\limits_{n=0}^{p-1} F_{i,p+m,j\cdot p+n} \end{cases} \tag{7}$$

In Equation (7), $p \times p$ denotes the size of the convolutional kernel. In the CNN model, the study introduces a hybrid AM. It mainly consists of two types of channel attention as well as spatial attention. First, channel attention mainly learns the weights between channels as a way to highlight the channel features with better values. For the input FM F, GAP, and global MP are applied to individual channels to obtain two IDVs $F_{avg,c}$ and $F_{\max,c}$, as shown in Equation (8).

$$\begin{cases} F_{avg,c} = \dfrac{1}{H \cdot W} \sum\limits_{i=0}^{H-1} \sum\limits_{j=0}^{W-1} F_{i,j,e} \\ F_{avg,i,j} = \dfrac{1}{p^2} \sum\limits_{m=0}^{p-1} \sum\limits_{n=0}^{p-1} F_{i,p+m,j\cdot p+n} \end{cases} \tag{8}$$

After obtaining two IDVs based on Equation (8), the channel attention weights α are obtained by splicing them and inputting them into the two FCLs as shown in Equation (9).

$$\alpha = Sigmoid(FC_2(\mathrm{Re}\,LU(FC_1([F_{avg}, F_{\max}])))) \tag{9}$$

In Equation (9), $Sigmoid$ denotes the AF. Then the mathematical expression of the final channel attention FM is displayed in Equation (10).

$$F_{ca} = \alpha \odot F \tag{10}$$

In Equation (10), F_{ca} is the channel attention FM. ⊙ denotes element-by-element multiplication. The spatial AM focuses on the weights between spatial locations as a way to highlight more valuable spatial regions [17-18]. First, two 2D FMs F_{avg}^{t} and $F_{\max}^{t}$ are obtained after channel AP and channel MP of the FMs, as shown in Equation (11).

$$\begin{cases} F_{avg,i,j}^{t} = \dfrac{1}{C} \sum\limits_{c=0}^{C-1} F_{i,j,e} \\ F_{\max,i,j}^{t} = \max\limits_{c} F_{i,j,c} \end{cases} \tag{11}$$

After obtaining a 2D FM based on Equation (11), it is spliced and input into a CL as a way to obtain the spatial attention weight β, as shown in Equation (12).

$$\beta = Sigmoid(Conv([F_{avg}, F_{\max}])) \tag{12}$$

In Equation (12), $Conv$ denotes the convolution operation (CO). The final mathematical expression based on the spatial attention FM F_{sa} is shown in Equation (13).

$$F_{sa} = \beta \odot F \tag{13}$$

The building of the enhanced CNN model is finished once the channel AM and the spatial AM have been incorporated into each CL. Fig. 5 displays the schematic of the CNN model based on the AM.

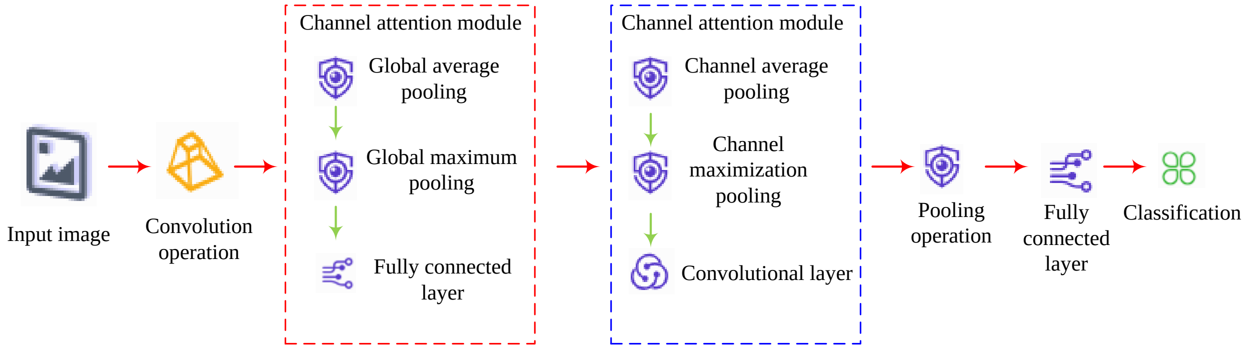


Figure 5: The structure of CNN model based on the AM

In Fig. 5, first, for the InI I , the FM F is obtained based on Equation (14) after the CO through the CL.

$$F = Conv(I) \tag{14}$$

Based on the FM F , it is applied to the channel and spatial attention module and the corresponding FM is obtained. The PO is performed on the final FM as a way to reduce the SD of the FM, as shown in Equation (15).

$$\begin{cases} F_{pool} = Pool(F_{final}) \\ F_{final} = \beta \otimes F_{ca} \end{cases} \tag{15}$$

The pooled FM is expanded into a IDV, which is fed into the FCL, as shown in Equation (16).

$$Y = \mathrm{Re}\,LU(WF_{pool} + b) \tag{16}$$

Finally, the result of the FCL is output to the classifier, which can complete the process of classifying high-precision images. Overall, the study proposes a FE model based on improved FPN. By introducing a weighted fusion mechanism and an adaptive feature adjustment strategy, it is able to fuse MSFs more effectively and highlight valuable information. Second, a CNN classification model based on the AM is constructed. By combining the channel and spatial AMs, the model's ability to focus on key information is enhanced. To facilitate the description of subsequent experiments, the overall model combining these two modules is defined as AM CNN with improved FPN (AM-CNN-FPN) model.

3 Results

3.1 Performance evaluation of CNN models with the introduction of an AM

The above study introduces a weighted fusion mechanism, an adaptive feature adjustment strategy, and an AM to optimize the CNN model. The operation of deep learning (DL) models often possesses high requirements on the computer environment. The experimental environment Settings are shown in Table 2 as follows.

After the experimental environment is set up, the specific parameters and architecture of the model are elaborated in detail in the study. The relevant parameter settings and values are shown in Table 3.

Based on the network structure parameter information shown in Table 3, the study first validates the performance of the improved FPN. It introduces the traditional CNN model, support vector machine (SVM) model, and random forest (RF) for controlled experiments. Meanwhile, the dataset is selected from ImageNet dataset and CIFAR-10 dataset.

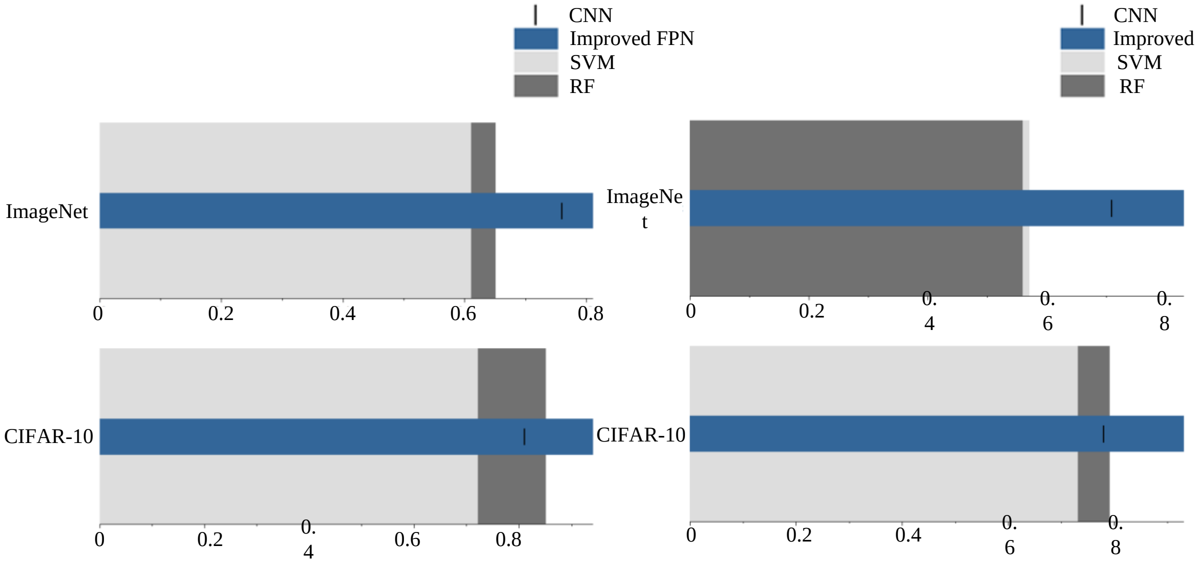
Table 2: Experimental environment setting

Name	Model and configuration	Name	Model and configuration
Operating system	Ubuntu 20.04 LTS	Deep learning framework	PyTorch 1.9.0
CPU	Intel Core i7-9700K	Programming version	Python 3.8
GPU	NVIDIA GeForce RTX 2080 Ti	CUDA version	CUDA 11.1
Memory	32GB DDR4 (3200 MHz)	cuDNN version	cuDNN 8.0
Hard disk	1TB NVMe SSD + 2TB HDD	/	/

Table 3: Parameter settings and values

Layer type	Layer name	Filter size	Number of filters	Stride	Padding	Activation function	Dropout rate
Input layer	Input	/	/	/	/	/	/
Convolutional layer	Conv1	3x3	32	1	Same	ReLU	0.2
Pooling layer	Pool1	2x2	-	2	/	/	/
Convolutional layer	Conv2	3x3	64	1	Same	ReLU	0.3
Pooling layer	Pool2	2x2	-	2	/	/	/
Convolutional layer	Conv3	3x3	128	1	Same	ReLU	0.4
Pooling layer	Pool3	2x2	-	2	/	/	/
Fully connected layer	FC1	/	256	/	/	ReLU	0.5
Fully connected layer	FC2	/	128	/	/	ReLU	0.5
Output layer	Output	/	Num_classes	/	/	Softmax	/

Large-scale IC and target identification tasks are the primary applications for the ImageNet dataset, which has over 14 million annotated images. The CIFAR-10 dataset is frequently used to assess how well IC models perform, particularly in terms of their capacity to classify small-size pictures. In the setting of hyperparameters, the attenuation factor of the learning rate is set to 0.01, and the minimum learning rate is set to 1×10-6. The initial learning rate of the optimizer Adam is 0.001, where β1=0.9, β2=0.999, and the weight attenuation coefficient is 1×10-5. According to the early stop standard, training will be stopped if the loss of the validation set does not improve within 20 consecutive epochs. Meanwhile, if the improvement of the validation set loss is less than 1×10-4, it is considered that there is no improvement. To improve the model's generalization ability and robustness, a variety of data augmentation techniques are adopted during the study's training process. First, there is a random horizontal flip, meaning the image is flipped horizontally at random with a probability of 0.5. Second, there is a random vertical flip, meaning the image is flipped vertically at random with a probability of 0.5. Next comes the random rotation, that is, the random rotation of the image, with the rotation Angle ranging from -10 to +10 degrees. Finally, there is random cropping, which involves cropping a random part of the image. The size of the cropped image is 224×224. The study simultaneously divides the dataset into the training set, the validation set and the test set in a ratio of 7:1:2. The study starts by comparing the four models' CA comparability. The results are shown in Fig. 6.



(a) Parallel experiment 1 classification accuracy comparison (b) Parallel experiment 2 classification accuracy comparison

Figure 6: Comparison of CA of the four models

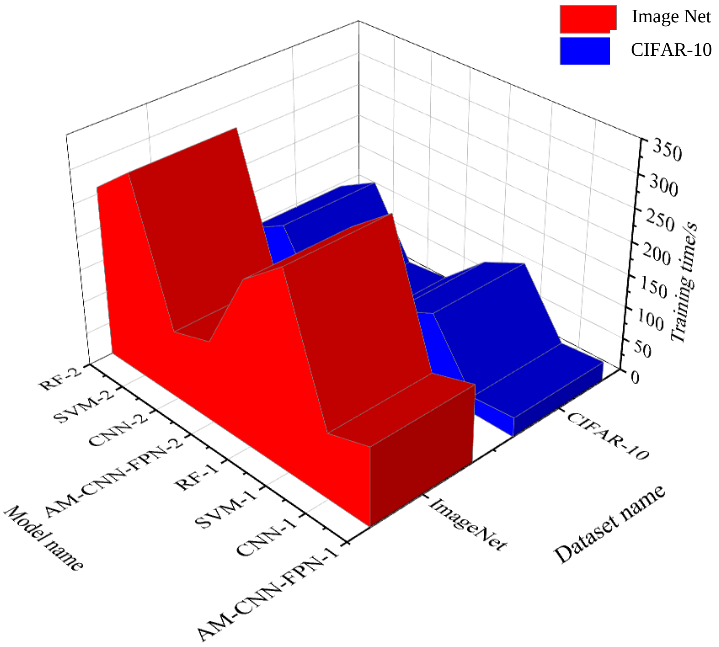


Figure 7: Comparison of computational efficiency among the four models

Fig. 6(a) shows the comparison of CA of the four models in parallel experiment 1. Fig. 6(b) shows the comparison of CA of the four models in parallel experiment 2. The enhanced FNN model's CA in the two simultaneous runs on the ImageNet dataset is 0.81 and 0.83, which is noticeably higher than that of the other three models. This indicates that the improved FNN model has obvious advantages in dealing with large-scale and complex IC tasks. The CA of the traditional CNN model is 0.76 and 0.71 in the two experiments, respectively. Although it also shows better performance, it still falls short of the improved FNN model. On the CIFAR-10 dataset, the CA of the improved FNN model is 0.94 and 0.93 in two parallel experiments, which is also significantly higher than the other three models. The findings reveals that the improved FNN model not only has an obvious advantage in CA, but also excels in performance stability and consistency. The study further compares the computational efficiency of the AM-CNN-FPN model with the control model. The results are shown in Fig. 7.

Fig. 7 shows the training time (TT) comparison of the four models. TT refers to the total time required for a model to complete its training process. This time depends on various factors, including the model's complexity, the dataset's size, and the hardware configuration. The results show that the TT of the AM-CNN-FPN model is 120s and 125s in parallel experiment 1 and parallel experiment 2, respectively. Its TT is slightly longer compared with that of the traditional CNN model, which may be due to the introduction of the AM and MSF fusion, and the computational complexity is increased. However, the overall TT is still within the acceptable range. The computational efficiency of the improved model is not significantly reduced while the performance is improved. The TT of SVM model is significantly higher than that of CNN model, probably due to the fact that SVM needs to extract features manually. Moreover, its computational complexity is higher when dealing with large-scale datasets. The TT of RF model is also longer. Although it shows better performance in some tasks, it is less computationally efficient on large-scale datasets. In summary, the AM-CNN-FPN model does not significantly decrease the computational efficiency while improving its performance. It shows that it has good application prospects in high-precision IC tasks. The study further introduces more advanced models, ResNet, DenseNet, and EfficientNet, for comparative experiments. Therefore, the performance comparison of the four models is shown in Table 4.

Table 4 shows that the performance parameters of the AM-CNN-FPN proposed in the study are all superior to those of the control model, with respective accuracy rates of 77.5% and 94.2%. The recall rates are 94.3% and 94.0%, respectively. The F1 is 94.6% and 94.3%, respectively. The robustness indicators are 93.5% and 94.5%. The TTs are 120 s and 30 s. The results show that the AM-CNN-FPN model performs well in both datasets. Meanwhile, its TT is the shortest, indicating that it also has advantages in computational efficiency. In conclusion, the AM-CNN-FPN model is highly effective and efficient at complex image classification tasks, making it a high-precision image classification model. The study verifies the performance of the AM-CNN-FPN model with ablation experiments. The results are shown in Table 5.

In Table 5, the CA of the AM-CNN-FPN model on the ImageNet dataset is 77.50%, and the TT is 120s. The CA of the AM-CNN-FPN model on CIFAR-10 is 94.20%, and the TT is 30s. The results show that the full AM-CNN-FPN model on both datasets exhibits the highest CA. It demonstrates how these two mechanisms working together can significantly enhance the model's performance. Meanwhile, the basic CNN model has an advantage in TT, but still has room for improvement in CA. It shows that the base CNN has achieved a better balance between performance and efficiency. However, its performance can be further improved by introducing the AM and improved FNN.

Table 4: Model performance comparison

Model	Dataset	Accuracy (%)	Recall (%)	F1 score (%)	Robustness (%)	Training time (s)
AM-CNN-FPN	ImageNet	77.5	94.3	94.6	93.5	120
ResNet	ImageNet	76.8	93.5	93.8	92	110
DenseNet	ImageNet	76.2	92.8	93.2	91.5	130
EfficientNet	ImageNet	76.5	93.2	93.5	92.5	115
AM-CNN-FPN	CIFAR-10	94.2	94	94.3	94.5	30
ResNet-50	CIFAR-10	93.5	92.5	92.8	93	28
DenseNet	CIFAR-10	93	91.5	91.8	92	32
EfficientNet	CIFAR-10	93.3	92.8	93	93.2	31

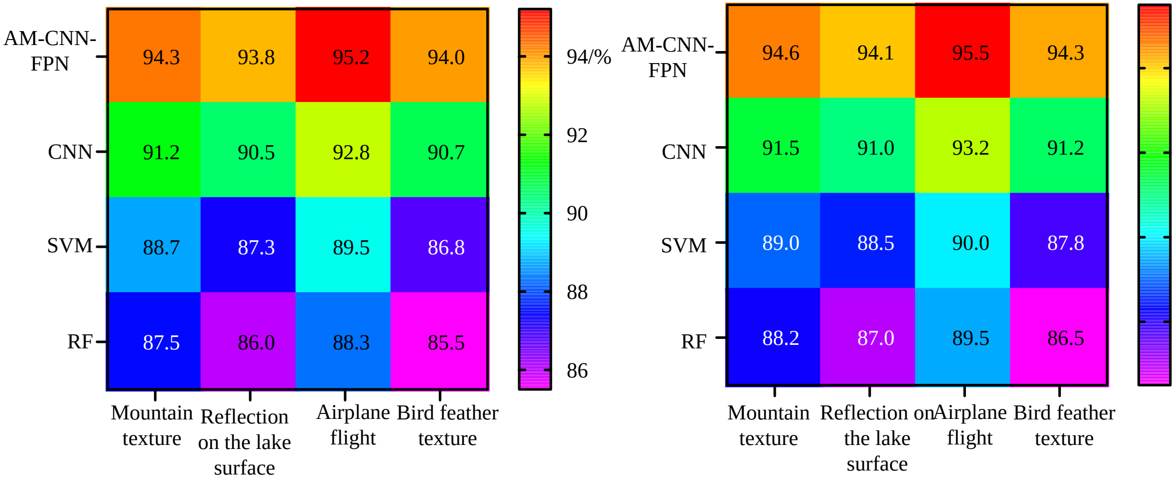
Table 5: Ablation experiment results

Data set	Model configuration	Classification accuracy	Training time
ImageNet	Complete improvement of CNN (AM+improved FPN)	77.50%	120
	Attention-free mechanism (only improving FPN)	76.64%	115
	No improved FPN (AM only)	76.25%	118
	Attention-free mechanism and improved FPN (basic CNN)	76.05%	100
CIFAR-10	Complete improvement of CNN (AM+improved FPN)	94.20%	30
	Attention-free mechanism (only improving FPN)	93.50%	28
	No improved FPN (AM only)	93.80%	29
	Attention-free mechanism and improved FPN (basic CNN)	93.06%	25

3.2 Performance comparison of different types of high-precision IC

After validating the performance of the AM-CNN-FPN model, the study further compares the classification performance of the model for different types of high-precision images. Among them, the high-precision images include mountain texture, reflection on the lake surface, airplane flight, and bird feather texture. First, the recall and F1 score (F1) comparisons based on the four high-precision images are shown in Fig. 8.

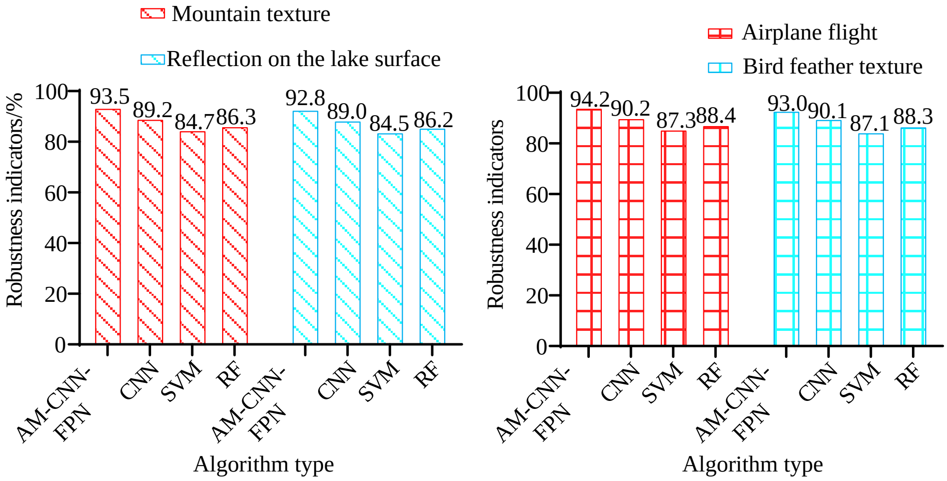
Fig. 8(a) shows the recall comparison of the model against four high-precision images. Fig. 8(b) shows the comparison of F1 of the model for four high-precision images. The results show that the recall of the AM-CNN-FPN model on the mountain texture image reaches 94.3% and the F1 is 94.6%. It is 3.1% and 3.1% higher than the traditional CNN, respectively. This illustrates that AM-CNN-FPN is able to recognize the target object more accurately while maintaining a high precision rate when recognizing complex natural scenes like mountain texture. In contrast, the SVM and RF models have lower recall and F1 of 88.7%, 89.0%, 87.5%, and 88.2%, respectively. The AM-CNN-FPN model has a recall of 93.8% and an F1 of 94.1% in the classification of the reflection on the lake surface image. It is also higher than 90.5% and 91.0% for the traditional CNN model. For airplane flight IC, the AM-CNN-FPN model achieves 95.2% recall and 95.5% F1, respectively. This compares favorably with 92.8% and 93.2% for traditional CNN. Finally, in the classification of bird feather texture images, the AM-CNN-FPN model has a recall of 94.0% and an F1 of 94.3%. It is 3.3% and 3.1% higher than the traditional CNN, respectively. In summary, the AM-CNN-FPN model enables the model to capture the key features and detail information in the image more effectively by introducing the AM and the improved FPN network. Meanwhile, it achieves a better balance between precision and recall. The study further compares the robustness metrics of the model for high-precision images. The results are shown in Fig. 9.



(a) Recall rate comparison

(b) Comparison of F1 scores

Figure 8: Recall rate of high-precision images and comparison of F1



(a) Comparison of robustness indicators between mountain texture and the reflection on the lake surface

(b) Comparison of robustness indicators between airplane flight and bird feather texture

Figure 9: Comparison of model robustness indicators

Fig. 9(a) shows the comparison of model robustness metrics for mountain texture and reflection on the lake surface high-precision images. Fig. 9(b) shows the comparison of model robustness indexes for airplane flight and bird feather texture high-precision images. The results show that the robustness index of AM-CNN-FPN model is 93.5%, which is significantly higher than that of traditional CNN (89.2%), SVM (84.7%), and RF (86.3%). It shows that AM-CNN-FPN is better able to resist the influence of noise and illumination changes and maintain high classification performance when dealing with complex natural scenes. In classification of reflection on the lake surface images, the robustness index of AM-CNN-FPN model is 92.8%, which is significantly higher than the rest of the models. It shows that it can recognize key features more effectively and reduce the possibility of misclassification when dealing with images with complex light and shadow effects. The robustness index of AM-CNN-FPN in airplane flight IC is 94.2%. It shows that it can locate and classify more accurately when recognizing images with clear target objects. Finally, the robustness index of AM-CNN-FPN in bird feather texture is 93.0%. It shows that it can extract key features more effectively and improve the CA when dealing with images with rich details and complex textures. In summary, the AM-CNN-FPN model, through the introduced AM and improved FPN network, enables the model to capture key features and detail information in images more effectively. Meanwhile, it shows stronger stability when facing complex conditions such as noise, light changes, and occlusion. A comparison of the model's capacity to extract fine-grained features for four high-precision images rounds out the study. Fig. 10 displays the findings.

The model's capacity to collect fine-grained features for four high-precision images is compared in Fig. 10. The results show that the AM-CNN-FPN model's detailed feature capturing ability for mountain texture images is 95.2%, which is significantly higher than that of traditional CNN (92.1%), SVM (87.6%), and RF (88.9%). It shows that it is able to capture features such as mountain texture and contours more effectively when dealing with complex natural scenes. For the reflection on the lake surface image, the detailed feature capturing ability of AM-CNN-FPN is 94.5%, which is higher than that of traditional CNN (91.3%), SVM (86.7%), and RF (87.8%). It shows that it can recognize and capture key details more effectively when processing images with complex lighting effects. For airplane flight, the detailed feature capturing ability of AM-CNN-FPN is 96.0%, which is significantly higher than that of traditional CNN (93.4%), SVM (88.2%), and RF (90.1%). It shows that it is able to capture detailed features more accurately when recognizing images with clear target objects. Finally, in terms of bird feather texture image, the detailed feature capturing ability of AM-CNN-FPN is 94.8%, which is higher than that of traditional CNN (91.7%), SVM (87.3%), and RF (89.5%).

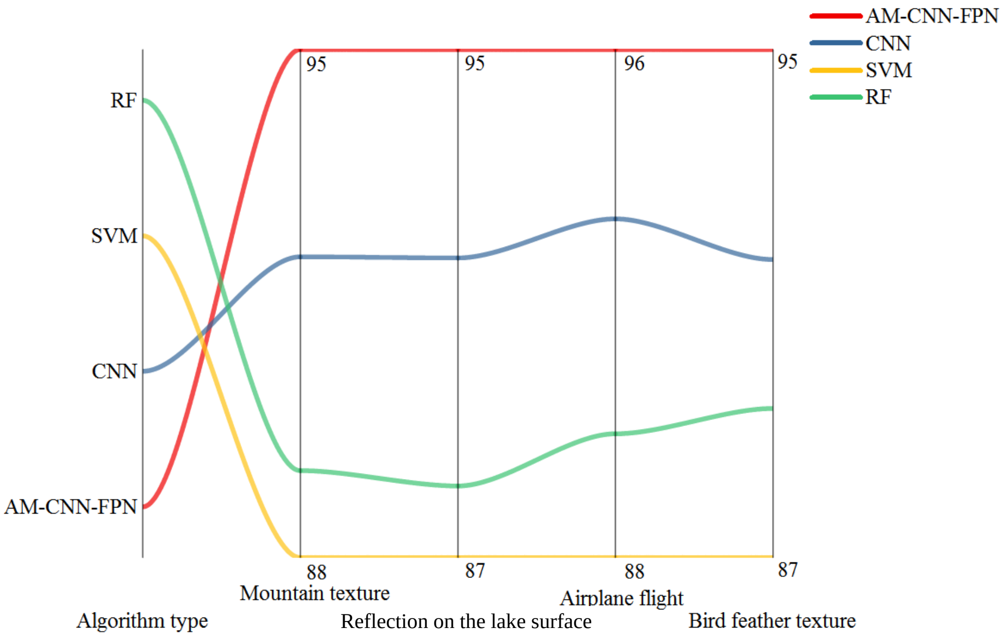


Figure 10: Comparison of detail feature capture capabilities

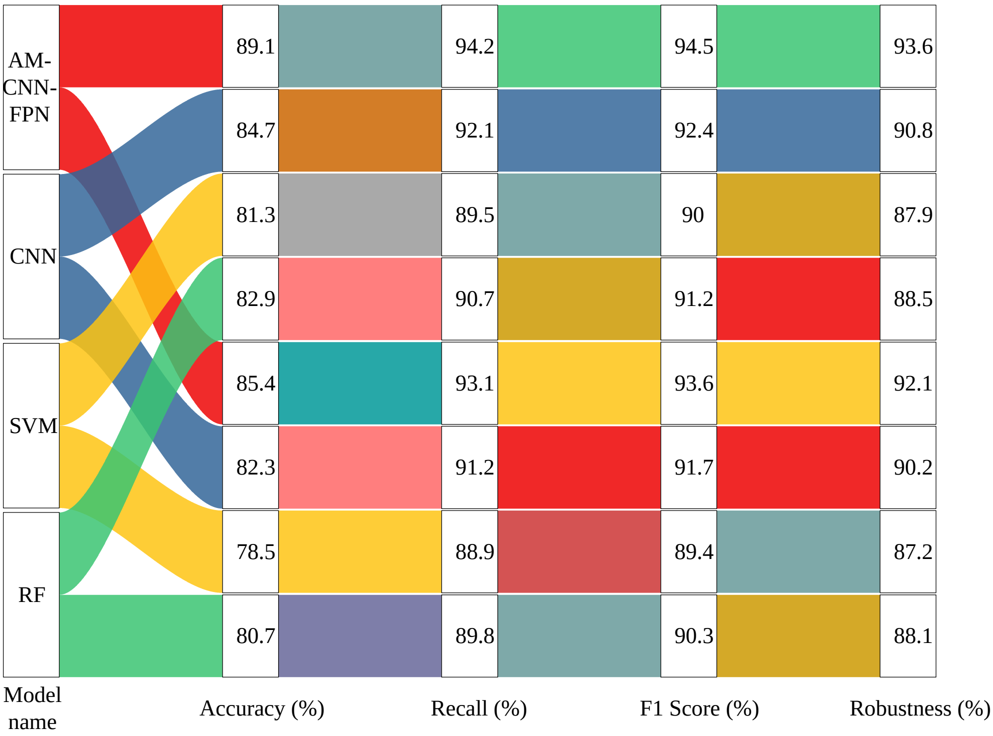


Figure 11: Performance comparison of four models for the medical image dataset

3.3 Comparison of model performance indicators in the medical field

Although the model performed well in general applications in the above-mentioned research, its use is limited to specific fields. To explore the broader application of the model, its performance in the field of medical images is studied and investigated. The MIMIC-CXR and CheXpert datasets are selected for the study on medical images. MIMIC-CXR is a large-scale dataset of chest X-ray images, containing over 370,000 chest X-ray images, and is used for the diagnosis of various diseases. The CheXpert dataset contains over 220,000 chest X-ray images for classifying various chest diseases. First, the performance comparison of the four models for the medical image dataset is shown in Figure 11.

As shown in Figure 11, the performance indicators of the AM-CNN-FPN model proposed in this study are superior to those of the control model. Its accuracy rates are 89.1% and 85.4%, respectively. The recall rates are 94.2% and 93.1% respectively. The F1 is 94.5% and 93.6%, respectively. The robustness indicators are 93.6% and 92.1% respectively. The SVM model performs the worst in the medical dataset, achieving an accuracy rate of less than 80%. All of the other performance indicators are below 90%. In summary, the AM-CNN-FFN model outperforms traditional CNN, SVM and RF models on medical image datasets. It indicates that the AM-CNN-FFN model not only performs well on general datasets, but also has good applicability and generalization ability on domain-specific datasets.

4 Discussion and conclusion

To enhance the performance of high-precision IC, the study improved the traditional FPN model and constructed an improved CNN model. Meanwhile, experimental validation was carried out on ImageNet and CIFAR-10 datasets. The numerical results indicated that on the ImageNet dataset, the CA of the improved FPNN model reached 77.50%, which had a significant advantage over the traditional CNN model (76.05%), SVM model (65.0%), and RF model (61.0%). On the CIFAR-10 dataset, the CA of the improved FNN model was 94.20%, which was also significantly higher than that of the traditional CNN model (93.06%), SVM model (85.0%), and RF model (72.0%). This outcome was in line with Yu F et al.'s findings, which suggested an enhanced FPN model for the advanced in entire crop growth cycle IC and recognition applications. The results indicated that the classification ability of the images of this improved FPN model was significantly better than that of the control model. Similar to this study, the CNN model incorporating the improved FPNN also introduced the AM as well as multi-feature fusion [19]. AM-CNN-FPN could capture the key features and detail information in the image more efficiently, which significantly improved the classification precision and robustness. In addition, AM-CNN-FPN performed well in different types of high-precision images (e.g., mountain texture, reflection on the lake surface, airplane flight, bird feather texture) in terms of recall, F1, and robustness metrics. The performance was excellent in terms of recall, F1, and robustness. For example, in mountain texture IC, the recall of AM-CNN-FPN reached 94.3%, the F1 was 94.6%, and the robustness index was 93.5%, which were higher than other models. The results differed from those of Zhou X et al. which proposed a tool wear classification method based on CNN and time series images. The method classified cutting force signals by converting them into time series images and then inputting them into a CNN model. Unlike this study, the classification ability of this model was weaker than the improved RF model. The possible reason for this could be that the improved RF was highly robust to noise and outliers and did not require complex preprocessing steps. Meanwhile, the model complexity of the improved RF was relatively low and the TT was shorter. Even on ordinary computing resources, its training could be completed quickly [20].

The findings demonstrate that the enhanced CNN model offers notable benefits for handling intricate natural sceneries and fine-grained characteristics. In summary, the numerical results reflect the effectiveness and superiority of the improved CNN model in high-precision IC tasks, which provides a new research direction and technical means for the field of IC. The AM-CNN-FPN model proposed in the research improves the classification performance by introducing the AM and multi-scale FE. However, the complexity of deep neural networks typically makes the models difficult to interpret. The model's AM provides clues about which parts of the InI are important for classification. For example, the channel attention module emphasizes the most informative FM, and the spatial attention module highlights important spatial regions within the FM. These attention maps can be visualized to provide intuitive insight into the model's decision-making process.

Although the AM-CNN-FPN model has shown promising results, it still has several limitations. First, the model's performance depends heavily on the quality and quantity of the training data. Insufficient or biased training data may lead to poor model performance and limited generalization ability. Second, introducing the AM and multi-scale FE increases the model's complexity and computational requirements. Compared with simpler models, this may lead to longer TT and higher resource consumption. Third, the interpretability of deep neural networks remains a challenge. Although AMs offer some interpretability, they cannot fully reveal the model's decision-making process. To enhance the robustness and generalization ability of the model in future research, more advanced data augmentation techniques and regularization methods should be explored. Meanwhile, techniques such as pruning, quantization, and knowledge distillation are used to reduce the model's complexity and computational requirements. In addition, layer-by-layer correlation propagation, SHAP or other advanced methods are adopted to explain the prediction results of the model in more detail. Ultimately, integrating the proposed model with other modalities could result in a more comprehensive and robust classification system.

References

[1] Rajdeep Kaur, Rakesh Kumar, and Meenu Gupta. Deep neural network for food image classification and nutrient identification: A systematic review. Reviews in Endocrine and Metabolic Disorders, 24(4):633-653, 2023.https://doi.org/10.1007/s11154-023-09795-4

[2] Rushit Dave, and Joy Purohit. Leveraging deep learning techniques to obtain efficacious segmentation results. Archives of Advanced Engineering Science, 1(1):11-26, 2023.https://doi.org/10.47852/bonviewAAES32021220

[3] Muthukrishnan Ramprasath, M.Vijay Anand, and Shanmugasundaram Hariharan. Image classification using convolutional neural networks. International Journal of Mechanical Engineering Research and Technology, 16(2):173-181, 2024.

[4] Shiwei Liu, Liejun Wang, and Wenwen Yue. An efficient medical image classification network based on multi-branch CNN, token grouping Transformer and mixer MLP. Applied Soft Computing, 153(4):111323-111342, 2024.https://doi.org/10.1016/j.asoc.2024.111323

[5] Mengxuan Zhang, Long Liu, Yaochu Jin, Zhikun Lei, Zhigang Wang, and Licheng Jiao. Tree-shaped multiobjective evolutionary CNN for hyperspectral image classification. Applied Soft Computing, 152(3):111176-111189, 2024.https://doi.org/10.1016/j.asoc.2023.111176

[6] Oleh Berezsky, Petro Liashchynskyi, Oleh Pitsun, and Ivan Izonin. Synthesis of convolutional neural network architectures for biomedical image classification. Biomedical Signal Processing and Control, 95(3):106325-106339, 2024.https://doi.org/10.1016/j.bspc.2024.106325

[7] Xin Wu, Yue Feng, Hong Xu, Zhuosheng Lin, Tao Chen, Shengke Li, Shihan Qiu, Qichao Liu, Yuangang Ma, and Shuangsheng Zhang, CTransCNN: Combining transformer and CNN in multilabel medical image classification. Knowledge-Based Systems, 281(2):111030-111048, 2023.https://doi.org/10.1016/j.knosys.2023.111030

[8] Mohammed Q. Alkhatib, Mina Al-Saad, Nour Aburaed, Saeed Almansoori, Jaime Zabalza, Stephen Marshall, and Hussain Al Ahmad. Tri-CNN: A three branch model for hyperspectral image classification. Remote Sensing, 15(2):316-335, 2023.https://doi.org/10.3390/rs15020316

[9] Qi Han, Xin Qian, Hongxiang Xu, Kepeng Wu, Lun Meng, Zicheng Qiu, Tengfei Weng, Baoping Zhou, and Xianqiang Gao. DM-CNN: Dynamic multi-scale convolutional neural network with uncertainty quantification for medical image classification. Computers in Biology and Medicine, 168(6):107758-107775, 2024.https://doi.org/10.1016/j.compbiomed.2023.107758

[10] Xiangzuo Huo, Gang Sun, Shengwei Tian, Yan Wang, Long Yu, Jun Long, Wendong Zhang, and Aolun Li. HiFuse: Hierarchical multi-scale feature fusion network for medical image classification. Biomedical Signal Processing and Control, 87(5):105534-105549, 2024.https://doi.org/10.1016/j.bspc.2023.105534

[11] Yang Yu, Yi Zhang, Zeyu Cheng, Zhe Song, and Chengkai Tang. Multi-scale spatial pyramid attention mechanism for image recognition: An effective approach. Engineering Applications of Artificial Intelligence, 133:108261,

2024.https://doi.org/10.1016/j.engappai.2024.108261

[12] Leyuan Fang, Yifan Jiang, Yinglong Yan, Jun Yue,and Yue Deng. Hyperspectral image instance segmentation using spectral–spatial feature pyramid network. IEEE Transactions on Geoscience and Remote Sensing, 61(4):1-13, 2023.https://doi.org/10.1109/TGRS.2023.3240481

[13] Yu Liu, Parma Nand, Md Akbar Hossain, Minh Nguyen, and Weiqi Yan. Sign language recognition from digital videos using feature pyramid network with detection transformer. Multimedia Tools and Applications, 82(14):21673-21685, 2023.https://doi.org/10.1007/s11042-023-14646-0

[14] Yi Zhang, Yushuang Zhu, Xiongwei Liu, Yingjian Lu, Chan Liu, Xixin Zhou, and Wei Fan. In-field tobacco leaf maturity detection with an enhanced MobileNetV1: Incorporating a feature pyramid network and attention mechanism. Sensors, 23(13):5964-5979, 2023.https://doi.org/10.3390/s23135964

[15] Yonglin Yu, Haifeng Li, Hanrong Shi, Lin Li, and Jun Xiao. Question-guided feature pyramid network for medical visual question answering. Expert Systems with Applications, 214(5):119148-119169, 2023.https://doi.org/10.1016/j.eswa.2022.119148

[16] Xiaoxia Meng, Xiaowei Wang, Shoulin Yin, and Hang Li. Few-shot image classification algorithm based on attention mechanism and weight fusion. Journal of Engineering and Applied Science, 70(1):14-31, 2023.https://doi.org/10.1186/s44147-023-00186-9

[17] Zia UrRehman, Yan Qiang, Long Wang, Yiwei Shi, Qianqian Yang, Saeed Ullah Khattak, Rukhma Aftab, and Juanjuan Zhao. Effective lung nodule detection using deep CNN with dual attention mechanisms. Scientific Reports, 14(1):3934-3949, 2024.https://doi.org/10.1038/s41598-024-51833-x

[18] Jia Liang, Xingyu Gu, Dong Jiang, and Qipeng Zhang. CNN-based network with multi-scale context feature and attention mechanism for automatic pavement crack segmentation. Automation in Construction, 164(5):105482-105499, 2024.https://doi.org/10.1016/j.autcon.2024.105482

[19] Feng Yu, Qian Zhang, Jun Xiao, Yuntao Ma, Ming Wang, Rupeng Luan, Xin Liu, Yang Ping, Ying Nie, Zhenyu Tao, and Hui Zhang. Progress in the application of cnn-based image classification and recognition in whole crop growth cycles. Remote Sensing, 15(12):2988-3010, 2023.https://doi.org/10.3390/rs15122988

[20] Xingying Zhou, Tianyu Yu, Guangzhou Wang, Ruiyang Guo, Yanxu Fu, Yazhou Sun, and Mingjun Chen. Tool wear classification based on convolutional neural network and time series images during high precision turning of copper. Wear, 522(2):204692-204716, 2023.https://doi.org/10.1016/j.wear.2023.204692