Volume 4, Issue 1, January 2025

Multi-Scale Feature Fusion Network for Object Recognition in Mountain Farmland Environments

Zhu Zhenglong[*], Zhang Qiang, Zhang Nanqing

College of Engineering, Zunyi Normal University, Zunyi 563006, Guizhou, China

*Corresponding to: Zhu Zhenglong

Abstract: Object recognition in mountainous field environment is a crucial part of intelligent operation. Since the diversity of crops, weeds and other targets in the mountainous field environment, as well as the complexity of scale changes, object recognition faces great challenges. In order to solve these problems, it is particularly critical to effectively fuse global and local multi-scale features. To this end, this paper proposes a three-branch parallel multi-scale feature fusion network (MFFNet) for object recognition in mountain and field environments. The MFFNet contains specially designed global and local feature extraction modules, which can capture the global and local feature information in the image in parallel. In addition, the MFFNet introduces a feature fusion strategy based on channel attention and spatial attention to effectively integrate global and local features with different semantic depths. The experimental results show that our proposed MFFNet is superior to the comparison method in multiple index results on the self-built mountain field environment image dataset.

Keywords: Deep learning, Object recognition, Multi-scale feature fusion, Mountain farmland environments

1 INTRODUCTION

Object recognition is one of the key areas in the field of computer vision, and its application scope has been rapidly expanded with the breakthrough of deep learning technology [1-3]. Among them, the mountainous field environment has become an emerging and key field for the application of object recognition technology due to its unique geographical and ecological characteristics [4]. The development of this technology will not only improve the efficiency and accuracy of agricultural monitoring, but also contribute to more intelligent and sustainable agricultural management, and by extending advances and innovations in agricultural science and technology [5].

However, target identification in mountain field environments also faces a number of challenges. Due to the presence of a variety of crops and natural vegetation in mountain fields, their size and morphology vary with different species and growth stages, resulting in the diversity and scale of targets in mountain fields, which undoubtedly increases the difficulty of identification tasks [6-8]. Therefore, it is particularly urgent to develop an object recognition technology that can adapt to the special needs of the mountain field environment.

Recent object recognition methods are mainly divided into two categories, one is based on convolutional neural networks (CNNs) and the other is based on Transformer architectures [9]. For CNN-based methods [10], due to the inherent limitations of convolutional kernels, they can usually only capture local features of the image. In contrast, Transformer-based methods can capture global semantic information through complex spatial transformations and modeling long-distance feature dependencies [11]. This capability gives Transformer an advantage in processing the global context of the image, helping to improve the accuracy and robustness of target recognition.

In the field of object recognition and image classification, the research of multi-scale feature fusion is becoming more and more important. For example, the latest ViTAE [12], Transfuse [13], and CCTNet [14].and other methods, their core goal is to achieve the effective integration of information at different scales. These advanced technologies have overcome the challenges of target diversity and scale change in natural images to a certain extent by combining the local perception of convolutional neural networks and the global contextual ability of self-attention mechanisms. However, these methods still face some limitations when directly applied to target recognition in complex mountain field environments.

In order to solve the above problems, this paper innovatively proposes a three-branch parallel multi-scale feature fusion network (MFFNet) for target recognition in mountainous field environments. At the same time, this paper designs a novel global and local feature module to capture the global and local feature information in the image in parallel respectively. In addition, this paper introduces a feature fusion strategy based on channel attention and spatial attention, which can effectively integrate global and local features with different semantic depths. This fusion strategy significantly improves the performance of the proposed method, making it more efficient and accurate than those recognition methods that rely heavily on traditional convolutional neural networks or Transformer architectures when dealing with complex environments. Finally, this paper conducts extensive testing on a carefully collected and annotated dataset of mountain field images. The results show that the MFFNet proposed in this paper achieves significant performance improvement, which further verifies the effectiveness and practicability of the proposed method.

The main contributions of this paper are as follows:

- In this paper, we propose a novel end-to-end three-branch parallel multi-scale feature information fusion network (MFFNet) for target recognition in mountainous field environments.
- In this paper, we innovatively design the global and local feature extraction modules to effectively capture the local spatial features and global semantic information at different scales, respectively.
- An Adaptive Feature Fusion Module (AFF) was proposed to fuse the semantic information between the different scale features of each branch to solve the problem of target diversity and scale change.
- The method proposed in this paper achieves good performance on the collected and labeled mountain field image dataset, and is better than the most advanced method.

2 Related Work

This chapter mainly introduces the application of object recognition technology and attention mechanism based on CNN and Transformer in the field of computer vision.

2.1 Convolutional Neural Networks

As a deep learning model, convolutional neural networks have achieved remarkable success in the fields of natural image classification and object recognition [15,16]. The core of this method is to automatically extract hierarchical features from images through multi-layer convolution and pooling operations, and capture representations from edges, textures, and higher-level semantic information layer by layer [17]. In recent years, classical convolutional neural network architectures such as VGGNet [18], DenseNet [19], and ResNet [20] have been developed one after another, which has promoted the improvement of the accuracy of image classification tasks. These models not only achieve excellent classification performance on large-scale datasets, such as ImageNet [21], but also serve as the basis for many vision tasks, significantly improving the ability of computers to automatically recognize and classify natural images [22]. However, the limitation of convolutional neural networks is that they are difficult to capture global features, and due to the local receptive field of convolutional operations, the model pays more attention to local information and has a less understanding of the global context, which can lead to performance degradation when dealing with complex scenarios or tasks that require overall understanding [23,24].

2.2 Transformer

Transformer models initially made breakthroughs in the field of natural language processing [25], but in recent years they have also been successfully applied to natural image classification and object recognition tasks [26]. Different from traditional convolutional neural networks, Transformer captures global features in images through a self-attention mechanism and does not rely on local receptive fields, which enables a better understanding of long-distance dependencies in images [27]. Models such as Vision Transformer [28] and Swin Transformer [29] use the embedded representation of image blocks as input, and use the global attention mechanism to model the global information of images, which significantly improves the classification performance on large-scale datasets. In addition, Transformer has also been used for tasks such as object detection [30] and segmentation [31], which can achieve more accurate identification and localization of targets in images by fusing local and global information [32]. However, the limitation of Transformer is that it is weak in capturing local information, especially when dealing with image details and small targets, and cannot be as effective as convolutional neural networks [33].

2.3 Attention Mechanisms

The attention mechanism is an important technique to improve the performance of deep learning models [34]. The core idea of the attention mechanism is to improve overall performance by giving different weights to different parts of the input data, allowing the model to focus more on important features [35]. Attention mechanisms can be divided into several categories according to the application scenario and structure, including self-attention [36], soft attention [37], and hard attention [38]. The self-attention mechanism is widely used in Transformer models to capture long-distance dependencies by calculating the associations between elements in the input sequence [28]. Soft attention generates a weighted average by weighting all parts of the input [37]; Hard attention is dealt with by selecting some key areas [38]. Attention mechanisms have shown significant performance improvements in many visual tasks, such as image classification, object detection, and image generation, and have become one of the indispensable components of modern deep learning models [39,40].

3 METHODS

As a method for identifying environmental targets in mountainous fields, the Multi-scale Feature Fusion Network (MFFNet) proposed in this paper can effectively obtain the local spatial information and global semantic representation information of environmental images of mountainous fields at different scales. The MFFNet model is based on the Global Feature Module with the local feature module extracts the global and local feature information of the mountain field environment image in parallel, and fuses the features of different scales through the adaptive feature fusion module and the downsampling step, and finally obtains the classification results. In the following sections, this article first introduces the overall structure of the MFFNet model, and then introduces the global feature module, local feature module, and adaptive feature fusion module.

3.1 MODEL ARCHITECTURE

In order to improve the accuracy of target recognition in mountainous field environment, this paper proposes a parallel network structure (MFFNet) that integrates multi-scale local features and global features. The overall framework of MFFNet is shown in Figure 1 and aims to enhance the recognition performance of the model by fusing multi-scale features. MFFNet consists of three main branches: the local branch is used to extract the local features of the image, the global branch is used to extract the global semantic representation of the image, and the feature fusion branch is used to integrate the multi-scale feature information extracted by the local branch and the global branch. The three-branch design of the network structure retains the local features and global representations to the greatest extent, so that the local branches and global branches can extract features independently. In addition, the multi-branch parallel structure enhances the feature characterization ability of the MFFNet model, and has good robustness to the extraction of multi-scale features. Table 1 describes the network parameters of MFFNet.

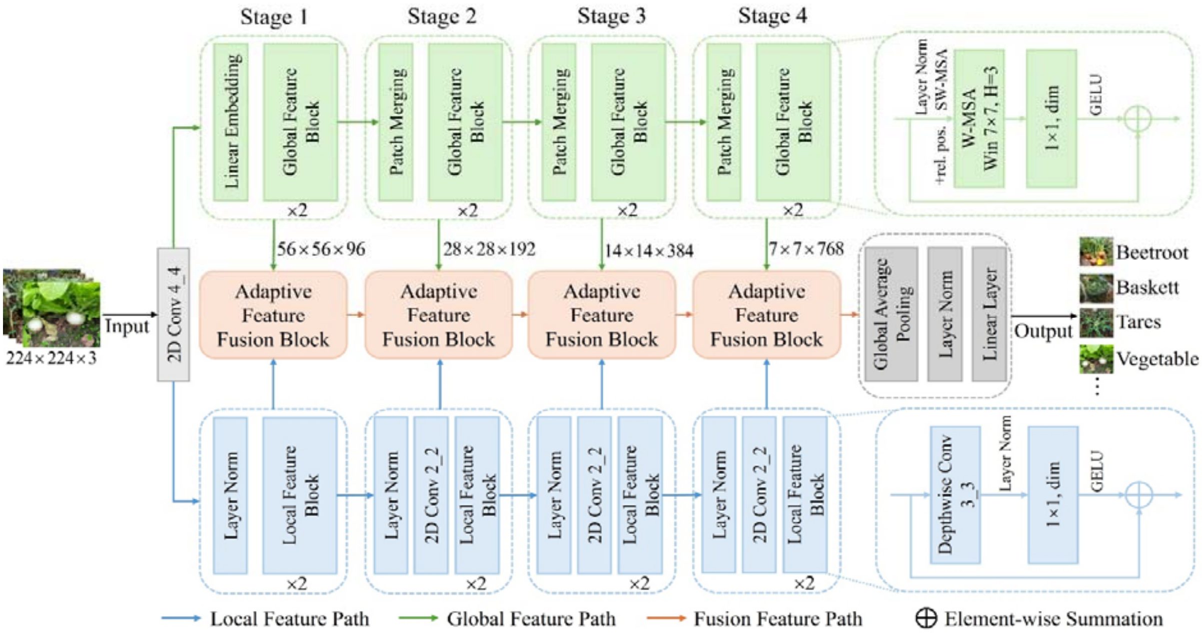

FIGURE 1 THE OVERALL ARCHITECTURE OF THE MFFNET MODEL

TABLE 1 PARAMETERS OF THE MFFNET NETWORK

stage	Output size	Local branches	Feature fusion branches	Global branches
0	56×56×96	Nucleus 4×4, channel 96, step 4	-	Nucleus 4×4, channel 96, step 4

<table>
<tr><td rowspan="3">1</td><td rowspan="3">56×56×96</td><td rowspan="3">Depth 3×3, channel 96
Nuclei 1×1, channel 96</td><td rowspan="3">×C</td><td>Spatial attention</td><td>Channel attention</td><td rowspan="3">Long attention, window 7×7, quantity 3
Relative position deviation, sliding</td><td rowspan="3">×C</td></tr>
<tr><td colspan="2">Depth 3×3, channel 96</td></tr>
<tr><td colspan="2">Nucleus 1×1, channel 384</td></tr>
</table>

				Nuclei 1×1, channel 96	window attention Nuclei 1×1, channel 96	
2	28×28×192	Depth 3×3, channel 192 Nucleus 1×1, channel 192	×C	Spatial attention | Channel attention Depth 3×3, channel 192 Nucleus 1×1, channel 768 Nucleus 1×1, channel 192	Long attention, window 7×7, quantity 6 Relative position deviation, sliding window attention Nucleus 1×1, channel 192	×C
3	14×14×384	Depth 3×3, channel 384 Nucleus 1×1, channel 384	×C	Spatial attention | Channel attention Depth 3×3, channel 384 Nucleus 1×1, channel 1536 Nucleus 1×1, channel 384	Long attention, window 7×7, quantity 12 Relative position deviation, sliding window attention Nucleus 1×1, channel 384	×C
4	7×7×768	Depth 3×3, channel 768 Nucleus 1×1, channel 768	×C	Spatial attention | Channel attention Depth 3×3, channel 768 Nucleus 1×1, channel 3072 Nucleus 1×1, channel 768	Long attention, window 7×7, quantity 12 Relative position deviation, sliding window attention Nucleus 1×1, channel 768	×C

Firstly, the image in the mountain field environment is divided into 4 patches by convolution operation with a kernel size of 4 and a step size of 4, and the height and width of each patch are 56 (that is, the resolution of the original image is reduced to 1/4 of the original). These patches are then preprocessed with Linear Embedding for global branches and Layer Norm for local branches, respectively. The global branch is then downsampled by Patch Merging and the Global Features module is fed into the Global Features module to extract the global features. At the same time, the local branches are downsampled by a convolution with a kernel size of 2 and a step size of 2. Subsequently, the feature map of each stage is extracted by four different global feature modules and local feature modules. At each stage, the Adaptive Feature Fusion Module (FFB) is used to fuse the features extracted by the local and global branches, while connecting with the output of the previous FFB. Finally, the fused features are fed into the Global Average Pooling and Layer Norm layers, and then the target is identified by a linear classifier.

In addition, different variants of the MFFNet model are constructed, named MFFNet-Tiny, MFFNet-Small, and MFFNet-Base. Each variant uses a different number of global and local feature modules in each stage, as shown in Table 2.

TABLE 2 CONFIGURATIONS OF DIFFERENT MFFNET VARIANTS

name	quantity			
	Phase 1	Phase 2	Stage 3	Stage 4
MFFNet-Tiny	2	2	2	2
MFFNet-Small	2	2	6	2
MFFNet-Base	2	2	18	2

3.2 *GLOBAL CHARACTERISTICS MODULE*

Mountain field environment images often exhibit significant intra-class variation and inter-class similarity, so global semantic representation information is particularly critical in this environment. In this paper, a global feature module based on Window Multi-Head Attention Mechanism (W-MSA) and Sliding Window Multi-Head Attention Mechanism (SW-MSA) is proposed. At each stage, the feature map first passes through the Layer by incorporating the image patch into the global feature module. The Norm layer is processed into the W-MSA module and then passes through a linear layer with a Gaussian error linear element activation function (GELU), as shown in the upper right corner of Figure 1. A common self-attention mechanism is in feedforward neural networks. network (FNN) with the aim of introducing nonlinearity. Usually, nonlinear activation is performed on the extended channel dimension generated by the linear layer to enhance the representation of the model. Activation is typically performed on the extended channel dimension generated by the linear layer to enhance the representation of the model. In order to further reduce the computational cost, this paper replaces the FNN with the GELU activation function, only performs nonlinear transformation, and compensates in the Adaptive Feature Fusion Module (AFF).

Residual connections are applied after each module and a relative position offset (rel. pos.), introducing SW-MSA in the next module. This process can be represented by the following formula:

$$F_g = F_{1\times1}(\text{W-MSA}(\text{LN}(F_G^{i-1}))) + F_G^{i-1} \qquad （1）$$

$$F_G^i = F_{1\times1}(\text{SW-MSA}(\text{LN}(F_g))) + F_g \qquad （2）$$

where F_G^i and $F_g$represent the global feature modules SW-MSA and W respectively,

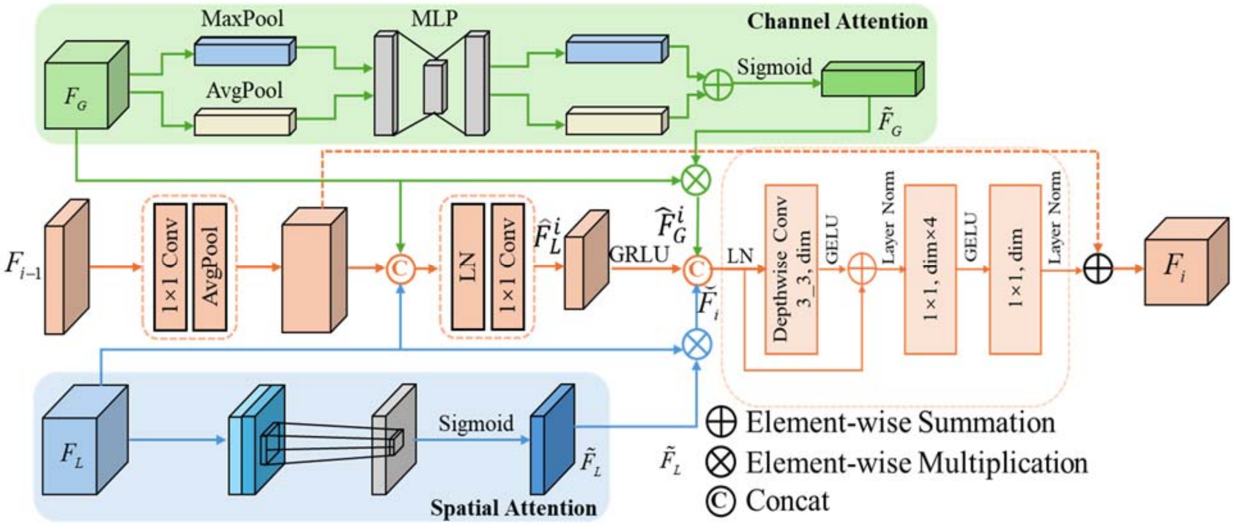

FIGURE 2 ADAPTIVE FEATURE FUSION MODULE

The output characteristic of the MSA, representing a convolution with a convolution kernel of $1\mathcal{F}_{1\times1}(\cdot)\times1$, "LN" is a layer-level normalization operation.

3.3 LOCAL FEATURE MODULE

In the image of the mountain field environment, local spatial features are also crucial. The Local Features module shown in Figure 1 uses a 3×3 Depthwise Separable Convolution to extract local features. Subsequently, the information exchange between features is realized through the linear layer. Finally, the extracted local features are fed into the Adaptive Feature Fusion Module (AFF) for further processing. The specific process is as follows:

$$F_L^i = F_{1\times1}(\text{LN}(\text{F}_d(F_L^{i-1}))) + F_L^{i-1} \qquad （3）$$

where F-L-i. represents the output feature with fast local features, and F-d. represents the deep separation convolution operation.

From a macro point of view, the structure of global and local branches is similar, and the number of channels and hierarchical design are maintained in different stages, which lays a solid foundation for the fusion of global and local coding features at different scales.

3.4 ADAPTIVE FEATURE FUSION MODULE

The Adaptive Feature Fusion Module (AFF) can adaptively fuse the local features and global representations of different stages and the semantic information after the fusion of the previous stage according to the input features, as shown in Figure 2. It represents the feature matrix generated by the global module, the feature matrix that represents the output of local features, the feature matrix generated by the adaptive feature fusion module in the previous stage, and the feature matrix generated by the adaptive feature fusion module in the current stage$F_G^i F_L^i F_{i-1} F_i$

Feature matrix. The calculation formula of the Adaptive Feature Fusion Module (AFF) proposed in this paper is as follows:

$$\hat{F}_G^i = CA(F_G^i) \otimes F_G^i \qquad （4）$$

$$\hat{F}_L^i = SA(F_L^i) \otimes F_L^i \qquad （5）$$

$$\tilde{F}_i = \text{Avgpool}(\text{F}_{1\times1}(F_{i-1})) \qquad （6）$$

$$\hat{F}_i = F_{1\times1}(LN(\text{Concat}[F_G^i, F_L^i, \tilde{F}_i])) \qquad （7）$$

$$\check{F}_i = \text{Concat}[\hat{F}_G^i, \hat{F}_L^i, \hat{F}_i] \qquad （8）$$

$$F_i = F_{1\times1}(F_{1\times1}(F_d\left(\text{LN}(\check{F}_i)\right) + \check{F}_i)) + \tilde{F}_i \qquad （9）$$

where ⊗ represents element-by-element multiplication, CA(·) represents channel attention, SA(·) represents spatial attention.-G-i. is a global semantic feature generated by channel attention combination.-L-i. is a local feature generated by spatial attention combination F.-i. is generated by down sampling in the previous adaptive feature fusion module stage,,,F.-i. is the result of the fusion of global-local features and the previous stage,, F-i. is the last multi-scale fusion feature generated.

The attention mechanism is an adaptive selection process, which selects the feature regions that are more conducive to recognition by assigning different weights to different features of the input. Among them, the channel attention mechanism allows the network to selectively focus on more important channel features and objects, while spatial attention pays more attention to spatial regions. The self-attention mechanism in the global feature module can capture the global information in space to a certain extent, but ignores the adaptability of the channel dimension. Therefore, the adaptive feature fusion module inputs global features into Channel Attention (CA), which takes advantage of the interdependence between channel graphs to improve the feature representation of specific semantics. Local features are input into spatial attention (SA) to selectively enhance important areas and suppress irrelevant information, so as to better retain more important details. The main processes of channel attention and spatial attention are as follows:

$$\begin{cases} \mathrm{CA}(x) = \sigma(\mathrm{MLP}(\mathrm{AvgPool}(x)+\mathrm{MLP}(\mathrm{MaxPool}(x))) \\ \mathrm{SA}(x) = \sigma(\mathrm{F}_{7\times7}(\mathrm{Concat}[\mathrm{AvgPool}(x),\mathrm{MaxPool}(x)])) \end{cases} \quad (10)$$

where σ denotes the Sigmoid function, F-7×7. (·) denotes a convolution operation with a convolution kernel of 7×7.

4 Experimental Results And Analysis

In this chapter, we mainly introduce the experimental data set, evaluation indexes, experimental details, and comparison methods of self-built mountain field environment, as well as the experimental results of the MFFNet model and comparison method proposed in this paper.

4.1 Experimental Datasets

The experimental datasets constructed in this study were collected online and offline, covering a variety of targets in the mountainous field environment of Guizhou, including the following 33 categories: spinach, cauliflower, sweet potato, cucumber, pepper, pepper, eggplant, celery, lettuce, garlic sprouts, green onions, potatoes, bitter gourd, coriander, Chinese cabbage, oilseed lettuce, tomatoes, cabbage, white radish, carrots, corn stalks, red cabbage, chicken hair, water spinach, Shanghai greens, weeds, people, flower baskets, knives, soil, and trees. In order to further enrich the dataset, this paper performs a variety of data augmentation processing on the images, including rotation, left and right mirroring, and adding salt and pepper noise. Finally, this dataset contains 11,742 images, as detailed in Table 3.

TABLE 3 DESCRIPTION OF THE DATASET OF MOUNTAINOUS FIELDS IN GUIZHOU

Sample category		Sample size
vegetable	spinach	201
	cauliflower	373
	sweet potato Cucumber	201 194
	bell pepper Peppers	200 196
	eggplant	203
	celery	188
	lettuce	199
	Garlic sprouts	201
	scallions	401
	potato	196
	Momordica charantia	198
	coriander	201
	cabbage	200
	Oily lettuce	201
	tomato	195
	cabbage	398
	White radish	262
	carrot	188
	Corn stalks	201
	Red cabbage	201
	Chicken hairy vegetables	187
	Water spinach	200
	Shanghai green	200
weed	amaranth	276
	Gray cabbage	

person	259
flower	3671
basket	661
animal	400
cutting tool	360
soil	200
tree	330
Total	11742

4.2 EVALUATION INDICATORS

To evaluate the trained model, four key metrics are used to comprehensively measure the model's performance: Accuracy, Precision, Recall, and F1-score. These metrics can be defined using a confusion matrix. In the confusion matrix, each column represents the predicted category and each row represents the actual category. TP (True Positive) refers to the number of samples correctly predicted to be positive by the model, TN (True Negative) refers to the number of samples correctly predicted by the model to be negative, and FP (False Positive) refers to the number of samples that were incorrectly predicted as positive by the modelFN (False Negative) refers to the number of negative samples that were not correctly identified by the model.

Accuracy, precision, recall, and F1 score are defined as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{11}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{12}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{13}$$

$$\text{F1-score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{14}$$

4.3 EXPERIMENTAL DETAILS

In the experiment, the image preprocessing steps of different methods are distinguished. For the convolutional neural network comparison method, all images are scaled to 256×256, while in the method in this paper and the Transformer-based comparison method, the images are uniformly scaled to 224×224. At the same time, the dataset is divided into a training set, a validation set, and a test set according to 7:1:2. In the training process, both the method and the comparison method in this paper perform random horizontal flipping and rotation data enhancement operations on the training set. The base learning rate for all methods was set at 1×10-4, the batch size was 32, and the learning rate was dynamically adjusted using a cosine annealing learning rate strategy. All methods were subjected to 100 training cycles to ensure adequate training and comparison under comparable conditions. In the output layer of the model, the softmax function is used to obtain the category prediction score, and the classification cross-entropy is used as the loss function to calculate the loss value of the model.

4.4 EXPERIMENTAL RESULTS

On the self-built Guizhou mountain field environment dataset, this paper selects three types of comparison methods: convolutional neural networks (such as VGG, ResNet, DenseNet), lightweight convolutional neural networks (such as MobileNet-v2, MobileNet-v3, ShuffleNetv2), and Transformer-based methods (such as ViT, Swin-Transformer, ConvNeXt, PerViT, UniFormer）. These methods are tested under different indicators with the MFFNet proposed in this paper, and the specific test results are shown in Table 4.

TABLE 4 RESULTS OF MFFNET AND COMPARISON METHODS

method	Accuracy	precision	Recall	F1 score
VGG-16	76.98	76.09	76.09	76.17
ResNet-34	78.76	78.54	78.54	78.40
DenseNet-121	74.51	74.20	74.04	74.04
DenseNet-201	75.10	74.58	74.58	74.52
MobileNet-v2	73.93	73.65	73.65	73.56
MobileNet-v3-s	67.10	66.50	66.50	66.43
MobileNet-v3-l	70.30	69.65	69.65	69.63
ShuffleNetv2-1	69.16	68.64	68.64	68.68
ShuffleNetv2-2	73.61	73.11	73.11	73.06
ViT-T-16	62.03	61.45	61.45	61.31
ViT-B-16	59.02	58.71	58.71	58.66
Swin-T-4-7	74.29	73.95	73.95	73.78
Swin-S-4-7	74.19	73.65	73.65	73.67
Swin-B-4-7	72.88	72.64	72.64	72.46
ConvNeXt-T	68.21	67.93	67.93	67.66
ConvNeXt-B	69.96	69.11	69.11	69.00

PerViT-T	73.58	73.78	73.78	73.44
B-PerViT-B	76.58	76.35	76.35	76.14
PerViT-M	75.74	75.29	75.29	75.13
UniFormer-S	76.87	76.77	76.77	76.49
UniFormer-B	72.33	72.47	72.47	72.07
MFFNet-T	76.25	75.63	75.63	75.60
MFFNet-S	78.64	78.37	78.37	78.22
MFFNet-B	79.82	79.46	79.46	79.29

As can be seen from Table 4, MFFNet-B has an accuracy rate of 79.82%, a precision of 79.46%, a recall rate of 79.46%, and an F1 score of 79.29%. From these results, it can be seen that traditional convolutional neural networks such as VGG, ResNet, and DenseNet series perform well in various metrics. Specifically, the DenseNet series, such as DenseNet-121 and DenseNet-201, excels in accuracy and F1 scores, thanks to its denser network structure, which can effectively extract deep features and improve the overall performance of the model. However, the MFFNet series proposed in this paper (including MFFNet-T, MFFNet-S, and MFFNet-B) are about 4% more accurate than DenseNet and about 5% better than DenseNet. This shows the advantages of MFFNet when dealing with more complex tasks.

For lightweight convolutional neural networks, the MobileNet and ShuffleNet series still exhibit high recognition accuracy while keeping the model lightweight. In particular, MobileNet-v2 and ShuffleNetv2 have a balanced performance in terms of accuracy and recall, which is suitable for resource-constrained real-world use cases. However, due to the limitations of their lightweight design, these models have slightly lower F1 scores than standard convolutional neural networks in complex scenarios. Comparatively, the MFFNet family improved by about 6% on each of these metrics, showing greater robustness in complex contexts, significantly outperforming these lightweight models.

Among the Transformer-based approaches, the ViT, Swin-Transformer, ConvNeXt, PerViT, and UniFormer series demonstrate powerful performance in vision tasks, especially in global semantic information extraction. Among them, the UniForme and PerViT series excelled in F1 scores. The MFFNet series (MFFNet-T, MFFNet-S, and MFFNet-B) proposed in this paper combine the advantages of multi-scale features to improve the accuracy by about 3% compared with the best Transformer model, and improve the accuracy, recall and F1 score by about 3%, showing its strong potential in environmental target recognition in mountainous fields.

In order to more fully demonstrate the recognition effect of the MFFNet model proposed in this paper on different targets in a mountainous field environment, the ROC curve and confusion matrix are plotted, as shown in Figures 3 and 4. These graphs provide us with a visual assessment of the model's performance and further validate the accuracy and reliability of MFFNet in different target classification tasks.

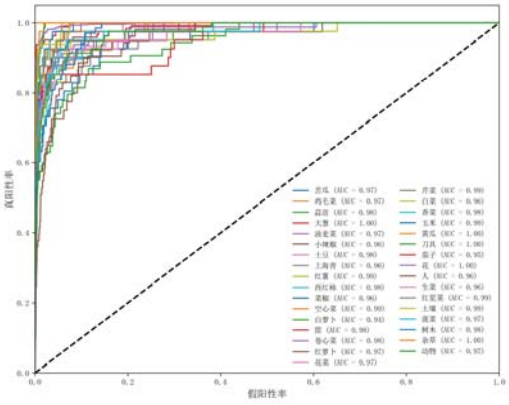

FIG.3 ROC CURVES OF THE MFFNET MODEL

The closer the ROC curve is to the upper left corner, the better the classification effect of the model. As can be seen in Figure 3, the MFFNet model performs very well in most categories, with the curve almost close to the upper left corner of the coordinates, indicating that the model has a high true positive rate and a low false positive rate in these categories. In addition, the AUC (Area Under the Curve) value under the curve is close to 1, which further verifies the excellent performance of the model.

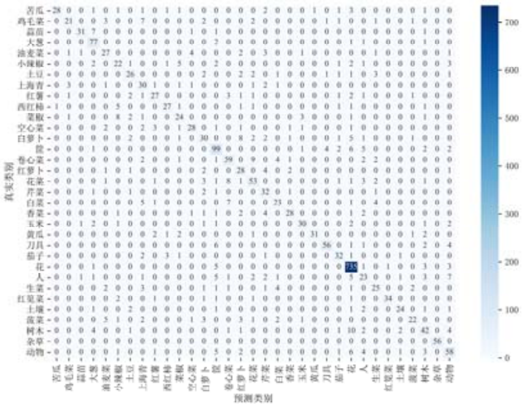

FIG. 4 CONFUSION MATRIX OF THE MFFNET MODEL

As can be seen from the confusion matrix of the model4, the MFFNet model has a higher classification accuracy in most categories, especially in the main categories such as weeds, flowers, green onions, knives, and cucumbers. On the contrary, for some categories with large intra-class differences or high similarity with other categories, such as white radish and pepper, some misclassification phenomena occurred in the confusion matrix, which may be due to the strong inter-class similarity or intra-class differences of such samples in the dataset.

In summary, it can be seen that the MFFNet model has shown strong ability in dealing with target recognition tasks in mountainous field environments, especially in the process of extracting global and local features, and has achieved high-precision classification of different types of targets in complex backgrounds. However, there is still room for improvement in the model when dealing with some hard-to-distinguish categories, and future research can consider further optimizing the feature extraction and fusion strategies to improve the recognition accuracy of these categories.

5 Conclusion

The Multi-scale Feature Fusion Network (MFFNet) proposed in this paper has shown significant progress in target recognition tasks in mountain field environments. By innovatively combining parallel global and local feature extraction modules, as well as feature fusion strategies based on channel attention and spatial attention, MFFNet can effectively capture and fuse multi-scale features. This method not only overcomes the limitations of traditional convolutional neural networks and Transformer architectures, but also significantly improves the accuracy and efficiency of object recognition in complex environments. Extensive testing on carefully collected and annotated mountain field image datasets further validates the excellent performance of MFFNet, and proves the effectiveness and practicability of the proposed method.

Funding

Zunhongkeheshizi [2022]06 ; Zunshi Kehe HZ Word [2023] No.152;Zunshi XM [2023] No.1-02; Zunshi kehe HZ Word [2023]No.161 ; Zunshi CXY [2024] 08.

References

[1] Zeng Jiexian, Ji Kang. Multi-view aircraft target recognition algorithm based on multi-feature fusion[J].Journal of Nanchang Hangkong University(Natural Science Edition),2016,30(02):8-15.)

[2] Logothetis N K, Sheinberg D L. Visual object recognition[J]. Annual review of neuroscience, 1996, 19: 577-621.

[3] Forsyth D A, Mundy J L, di Gesú V, et al. Object recognition with gradient-based learning[J]. Shape, contour and grouping in computer vision, 1999: 319-345.

[4] Zhou Yanan, Chen Hui, Liu Hongbin. Transactions of the CSAE,2022,38(23):213-222.)

[5] Kang Mengzhen, Wang Xiujuan, Hua Jing, et al. Parallel Agriculture: Intelligent Technology Towards Smart Agriculture[J]. Chinese Journal of Intelligent Science and Technology, 2019, 1(2): 107-117.

[6] Yao Z, Yang X, Wang B, et al. Multidimensional beta-diversity across local and regional scales in a Chinese subtropical forest: The role of forest structure[J]. Ecology and Evolution, 2023, 13(10): e10607.

[7] Wang Q J, Zhang S Y, Dong S F, et al. Pest24: A large-scale very small object data set of agricultural pests for multi-target detection[J]. Computers and Electronics in Agriculture, 2020, 175: 105585.

[8] Wang Q, Huang W, Xiong Z, et al. Looking closer at the scene: Multiscale representation learning for remote sensing image scene classification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 33(4): 1414-1428.

[9] Arkin E, Yadikar N, Muhtar Y, et al. A survey of object detection based on CNN and transformer[C]//2021 IEEE 2nd international conference on pattern recognition and machine learning (PRML). IEEE, 2021: 99-108.

[10][10] Maurício J, Domingues I, Bernardino J. Comparing vision transformers and convolutional neural networks for image classification: A literature review[J]. Applied Sciences, 2023, 13(9): 5521.

[11] Khan S, Naseer M, Hayat M, et al. Transformers in vision: A survey[J]. ACM computing surveys (CSUR), 2022, 54(10s): 1-41.

[12] Xu Y, Zhang Q, Zhang J, et al. Vitae: Vision transformer advanced by exploring intrinsic inductive bias[J]. Advances in neural information processing systems, 2021, 34: 28522-28535.

[13] Zhang Y, Liu H, Hu Q. Transfuse: Fusing transformers and cnns for medical image segmentation[C]//Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, Part I 24. Springer International Publishing, 2021: 14-24.

[14] Wang H, Chen X, Zhang T, et al. CCTNet: Coupled CNN and transformer network for crop segmentation of remote sensing images[J]. Remote Sensing, 2022, 14(9): 1956.

[15] Alzubaidi L, Zhang J, Humaidi A J, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions[J]. Journal of big Data, 2021, 8: 1-74.

[16] Bo ZHANG. Artificial intelligence is entering the post deep-learning era[J]. Chinese Journal of Intelligent Science and Technology, 2019, 1(1): 4-6.

[17] Li Y, Chen L, Huang Zhaohong, et al. Plant Leaf Detection Technology Based on Multi-scale Convolutional Neural Network Feature Fusion[J]. Chinese Journal of Intelligent Science and Technology, 2021, 3(3): 304-311.

[18] Simonyan K. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[19] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.

[20] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[21] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25.

[22] Zhao X, Wang L, Zhang Y, et al. A review of convolutional neural networks in computer vision[J]. Artificial Intelligence Review, 2024, 57(4): 99.

[23] Khan A, Sohail A, Zahoora U, et al. A survey of the recent architectures of deep convolutional neural networks[J]. Artificial intelligence review, 2020, 53: 5455-5516.

[24] Lindsay G W. Convolutional neural networks as a model of the visual system: Past, present, and future[J]. Journal of cognitive neuroscience, 2021, 33(10): 2017-2031.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[26] Maurício J, Domingues I, Bernardino J. Comparing vision transformers and convolutional neural networks for image classification: A literature review[J]. Applied Sciences, 2023, 13(9): 5521.

[27] Han K, Wang Y, Chen H, et al. A survey on vision transformer[J]. IEEE transactions on pattern analysis and machine intelligence, 2022, 45(1): 87-110.

[28] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[29] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.

[30] Li Y, Miao N, Ma L, et al. Transformer for object detection: Review and benchmark[J]. Engineering Applications of Artificial Intelligence, 2023, 126: 107021.

[31] Li X, Ding H, Yuan H, et al. Transformer-based visual segmentation: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.

[32] Han K, Wang Y, Chen H, et al. A survey on vision transformer[J]. IEEE transactions on pattern analysis and machine intelligence, 2022, 45(1): 87-110.

[33] Zhu W, Sun J, Wang S, et al. Identifying field crop diseases using transformer-embedded convolutional neural network[J]. Agriculture, 2022, 12(8): 1083.

[34] Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning[J]. Neurocomputing, 2021, 452: 48-62.

[35] Hassanin M, Anwar S, Radwan I, et al. Visual attention methods in deep learning: An in-depth survey[J]. Information Fusion, 2024, 108: 102417.

[36] Zhao H, Jia J, Koltun V. Exploring self-attention for image recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10076-10085.

[37] Chaudhari S, Mithal V, Polatkan G, et al. An attentive survey of attention models[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2021, 12(5): 1-32.

[38] Papadopoulos A, Korus P, Memon N. Hard-attention for scalable image classification[J]. Advances in Neural Information Processing Systems, 2021, 34: 14694-14707.

[39] Guo M H, Xu T X, Liu J J, et al. Attention mechanisms in computer vision: A survey[J]. Computational visual media, 2022, 8(3): 331-368.

[40] Brauwers G, Frasincar F. A general survey on attention mechanisms in deep learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 35(4): 3279-3298.