MAFNet: A Multiangle Attention Fusion Network for Land Cover Classification

Guoying Miao, Huiqin Wang, and Enwei Zhang

***Abstract*—The classification of land cover types is an important task for monitoring land use. Moreover, with the continuous application of high resolution remote sensing images, the time and space span is becoming larger, which greatly increases the difficulty of classification of target types. And there are few results that can effectively deal with deep land category information. Furthermore, the extraction and fusion of deep features still need to be improved. In this article, a multiangle attention fusion network is proposed for the classification of land cover types. The network uses a 50-layer residual network as a feature extraction network, and an adaptive special-shaped window attention module is added to the deep layer of the network to extract deep semantic information, building the connection between global information. In addition, the multiangle interactive attention fusion module is used to fuse feature maps at different levels, and an interactive attention mechanism is established at different angles. Finally, a new decoder module is proposed for the decoder to adjust the fused feature map again. Through experiments on three datasets, the results show that the method proposed in this article is more accurate than the previous network for the segmentation results of different categories. It can realize the accurate division of land types in remote sensing images, and has good generalization ability.**

***Index Terms*—Deep learning, fusion, multiangle, remote sensing images.**

I. Introduction

LAND cover type classification refers to the process of feature recognition in the remote sensing image collection obtained by satellite or UAV. It aims to classify different land use types and obtain final land cover images containing different classes. At present, the maturity of remote sensing technology makes it possible for people to obtain higher resolution ground remote sensing images. High resolution ground remote sensing images are widely used in cloud/snow classification [1], [2], [3], [4], urban and rural land description [5], [6], and change detection [7], [8]. The accurate division of land cover types is one of the effective methods for monitoring land use, which can intuitively represent different land cover types. The resulting classification map plays an important role in many fields, such as precision agriculture, urban mapping, and environmental monitoring [9]. Land cover analysis and land use monitoring are important for land planning and management. With the improvement of technical means, high-resolution remote sensing images can be collected, and detailed land use data can be obtained by analyzing remote sensing data [10]. High-resolution remote sensing images are very important raw materials for land cover classification. However, they contains some disturbance information, so it will bring great difficulties to the classification of land cover types. At present stage, due to the continuous construction of buildings, roads, bridges, and other series, the difficulty of classifying land types in remote sensing images is also increasing [11].

Traditional methods mainly use feature-based methods to classify land types, which requires manual extraction of features and then, classification of land types based on these features. However, manually extracted features are usually not accurate. It does not analyze the association between data, where these problems have led researchers to notice the importance of data analysis in land cover tasks. In the past ten years, many researchers combined machine learning with remote sensing images, using low-level information based on spectral features and geometric features to realize classification, such as decision tree [12], random forest [13], [14], [15], maximum likelihood method [16], [17], [18], etc. Initially, most researchers choose the maximum likelihood method to calculate the mean and variance of each category in the region of interest in the image, and classify each pixel through the classification function [17]. With the improvement of the resolution of remote sensing images, the spectral features become more abundant, and the classification relying on spectral features gradually shows its limitations. The above methods are only suitable for the case of fewer spectral features. To overcome the problem, researchers began to explore other machine learning algorithms. Since there exists a drawback that the use of prior probability in the maximum likelihood, decision tree [19], [20] classification algorithm has been considered to have great potential in land cover tasks, where it has successfully improved the results of land cover classification. However, although decision trees improve the accuracy of land cover classification to some extent, researchers still want to further improve the classification results. In the following time, as the exploration of new machine learning algorithms continued, support vector machine (SVM) [21], [22], [23], [24] gradually came into people's field of vision. SVM can fit an optimal hyperplane between different categories, and it can get high accuracy without too many training samples. This is a great advantage for cases with less training data. In

Manuscript received 25 May 2023; revised 27 July 2023; accepted 6 September 2023. Date of publication 13 September 2023; date of current version 27 September 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62073169 and in part by Jiangsu Province "333 Project" under Grant BRA2020067. *(Corresponding author: Guoying Miao.)*

The authors are with the School of Automation, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: mgyss66@163.com; 20211249155@nuist.edu.cn; 20211249116@nuist.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3314847

addition, random forest is also a popular method in land cover classification tasks. For example, comparing the random forest method with other multiremote sensing and geographic data integration methods, Goel et al. [13] discussed the application of random forest classifier in land cover classification. Eisavi et al. [14] used a random forest classifier to study the accuracy and efficiency of land cover classification in and around Nagad city, West Azerbaijan province, northwest Iran.

Both traditional methods and machine learning-based methods rely heavily on prior knowledge. Moreover, their operation is complex and limited, so they cannot adapt to current needs. For land cover classification tasks, accurate identification of different land use types is the primary goal. However, with the development of remote sensing technology and the continuous improvement of image resolution, the time and space span is becoming larger, and the content is becoming more complex, so the segmentation of the target object meets great difficulties. The occupation of land by different objects become more complex. For example, the size of the land covered by buildings is very different, and the shadow of high-rise buildings on the low-rise buildings will confuse the low-rise buildings from the background. For obstacles such as water, bridge, ship and their projections, it will cause errors in image segmentation details. Moreover, the low-rise buildings, trees, and vegetation in the villa area will shade the pond, which increases the difficulty of feature extraction. At the same time, the increasing resolution of image and complexity of spectral features also limit the accuracy of traditional methods.

Deep learning is a branch of machine learning. In the 21st century, SermaNet [25] proposed convolutional neural network, which uses convolution to extract feature information in images. Through a large number of training data, the algorithm adaptively learns the abstract features and the correlation between different features. Therefore, analyzing texture details and spatial environment features in high-resolution remote sensing images is much better than traditional methods. At present, the land type cover classification of remote sensing images based on deep learning has achieved certain results, but there are still many problems to be solved. Convolutional neural networks generally use downsampling to extract features in images, so that the high-resolution detail information is easy to lose, resulting in final segmentation inaccuracy. With the continuous improvement of the resolution of remote sensing images, the spatial information and detail information among them increase sharply. Unfortunately, most existing models have defects in the utilization of deep features. If the networks use pure convolution to process the feature information in the image, the limitations of convolution also cause the model to fail to establish the relationship between global features. Some researchers tried to use transformers in image tasks or combined them with convolution, which can effectively focus on important information, but it cannot balance the relationship between accuracy and parameters. For land cover classification tasks, there is still a need for further improvement in the extraction and fusion of deep features.

To solve these problems, this article proposes a multiangle attention fusion network (MAFNet) for land cover classification. This network uses ResNet-50 [26] as the feature extraction network to extract information at different levels in the image. In terms of processing deep features, we proposed an adaptive special-shaped window attention module (ASWA) to make up for the shortage of convolution focusing only on local features, which can establish the connection between long-distance pixels, and the use of attention mechanism makes the model more efficient in the use of deep feature information. Multiangle interactive attention fusion module (MIAF) is applied to the deep layer of the network, which is utilized to fuse information between different feature layers and focus on the relationship between global information in different directions in the feature graph. A new decoder module is proposed for the secondary processing of the fused feature map in order to restore the high-resolution feature map.

In general, this article has the following contributions.

1) A MAFNet is proposed for land cover classification. The network can accurately segment different land types in the image and has better detection accuracy than the current model.
2) An ASWA is proposed to make up for the problem of insufficient processing of deep feature information in the current model, which can establish the connection between global information and extract important information effectively.
3) A MIAF is proposed to fuse feature maps of different levels, and an interactive attention mechanism is established in different directions, which can accurately extract the information in the feature map and avoid the influence of various interference information.
4) A new decoder module is proposed to process the fused feature map, which is of great significance for recovering the information in the high-resolution image.

II. Related Works

At present, remote sensing data gradually presents the trend of high spectral resolution, high time-phase resolution, and high spatial resolution. Traditional machine learning methods have been unable to meet the requirements of land cover types. With the rise of deep learning, it is gradually applied in the field of image. The method of deep learning can handle a variety of visual tasks and avoid the deficiency of manual feature extraction. Compared with traditional methods, deep learning-based methods do not need artificial feature extraction and can achieve accurate pixel-level classification. It also has stronger anti-interference ability and strong generalization ability in the face of complex and changing environment, so it is more suitable for the classification of land cover types in high-resolution remote sensing images.

The initial approach is to divide the large-scale image into small image patches, and then use different models to classify the image patches. For example, Heryadi et al. [27] used convolutional neural networks as classifiers to study the classification of land cover types in Semarang Area, Indonesia. By dividing the input image into different parts, extracting the color, hue and other features of each part and sending them to the convolutional neural network for classification, the land cover classification of each part is realized. Dai et al. [28] studied the method of combining multiple classifiers to classify images, and the final category was obtained by integrating the results of multiple classifiers through voting. These methods are based on the classification of each small block area, which cannot obtain fine division results. Compared with the pixel-level classification, this classification method is rough.

Since 2005, Long et al. [29] proposed the fully convolutional network (FCN), which realizes end-to-end pixel-level classification for the first time, a large number of subsequent segmentation networks have been proposed and used in pixel-level classification tasks. Ronneberger et al. [30] proposed a U-shaped network structure (UNet) that can obtain both context information and location information. Then it achieves better training effect with fewer training samples. The pyramidal pool module proposed by Zhao et al. [31] can aggregate context information of different regions, thus improving the ability to obtain global information. Yu et al. [32] adopted a multibranch structure consisting of a detail branch and a semantic branch, as well as a guided aggregation layer to fuse the features of the two branches. Chen et al. [33] designed an ASPP module to solve the multiscale problem of the object, which is common in the segmentation task. Because deep learning has shown unique advantages in pixel-level classification tasks, researchers have widely applied it to classify land cover types. Pang et al. [34] proposed a real-time semantic segmentation framework (SGBNet) to solve the common problems of detail loss and edge blurring in real-time semantic segmentation of land cover. The proposed network framework has extremely fast inference speed for tasks with fewer categories, and its performance for classification tasks with more categories remains to be considered. In order to reduce the weight of the model, Gao et al. [35] studied the role of multibranch network in the land cover type classification task. Different branches focus on different feature information for extraction, and the three-way parallel structure of the network can significantly improve the performance, while the algorithm complexity only slightly increases. However, the effect of this method is not ideal when facing more categories or more complex datasets. When the buildings are in different shooting conditions from different angles, it is difficult to ensure that the prediction results are perfectly combined with the actual situation. Shen et al. [36] proposed a multiscale aggregation network to solve the problem of information loss and resolution degradation in the downsampling process of traditional convolutional neural networks, which assembled regional context information from different fields and solved the problem of incomplete information in traditional convolutional neural networks at different scales. This method is beneficial to multiscale targets, but it has certain limitations for the performance in the presence of more interference factors. Mehrotra et al. [37] proposed a model based on U-Net for the segmentation of surface rivers and land. The network performs well on SAR remote sensing images, but the classification of more categories in optical remote sensing images needs to be verified. Li et al. [38] developed the first land classification dataset together optical with SAR remote sensing, and proposed a joint semantic segmentation framework for multimodal images of land use classification. Most of the above methods adopt the form of convolution, which do not pay attention to global information, resulting in weak processing ability for different categories of information in particularly complex situations.

Later, it was found that the attention mechanism in transformer [39] can also produce good effects in visual tasks. So, Wu et al. [40] introduced convolution into visual transformer, which not only maintains the features of convolution (invariability of translation, scaling and rotation), but also maintains the advantages of transformers (dynamic attention, global context, and better generalization). Wang et al. [41] introduced pyramid structure into transformer, so that a gradually shrinking pyramid can be used to reduce the calculation amount in the training process of the model. Chen et al. [42] combined Swin Transformer and convolution, made full use of its ability to capture global information, and proposed a two-branch model for land cover type classification. Although this method uses the attention mechanism, the final result can achieve good accuracy. But the structure is not optimized, which leads to no advantage in the weight of the model.

III. Methodology

With the continuous improvement of the resolution, the complexity of the information contained in the remote sensing images we obtain is also getting higher. How to obtain the information we need more accurately from the complex content has become a problem to be solved. Therefore, how to improve the conduction process of information flow in network feature extraction has become a key to improve the accuracy of the model. At present, the effective use of deep-level information is not very good. With the deepening of network layers, the complexity and abstractness of information also increase. Furthermore, the application of deep features needs to be improved.

In this article, we propose a novel convolutional neural network for land cover classification, which focuses on the application of deep information and a more efficient transmission mode. This section first introduces the architecture of MAFNet. Then, we elaborate our network from the selection of backbone network, ASWA, MIAF, and DE.

A. Network Architecture

We propose a new network for the classification of land types. The network architecture is shown in Fig. 1. As the whole, we use an encoder-decoder structure to learn from low-resolution feature maps and recover high-resolution feature maps, which can effectively alleviate the loss of details in the extraction process of high-resolution image details. Compared with the method of using and saving high-resolution feature maps in the whole network [43], the hardware requirements of the encoder-decoder structure are relatively low, and the training inference can also be performed on ordinary devices.

The network structure of encoder and decoder enables our network to be divided into two parts: 1) feature extraction; and 2) original information recovery. We use ResNet50 [26] as our backbone network to extract information from images.

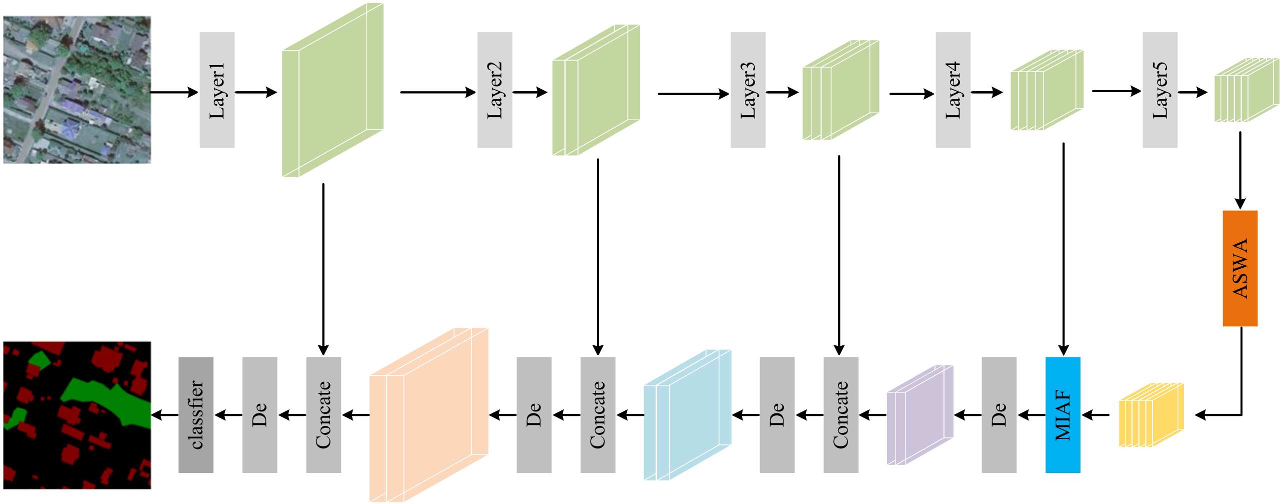

Fig. 1. Overall structure of the network.

In order to make full use of deep features and improve the transmission efficiency of information, we designed the adaptive anomalous window attention module (ASWA) to extract the deep semantic information in the deep layer. Meanwhile, we also designed an MIAF to integrate different levels of information in the deep layer of the network, improve the utilization rate and transmission efficiency of information, and make full use of the rich category information in the deep layer. This can effectively improve the final classification accuracy of the network. In the decoder part, a new decoder module (De) is added in this article. Because U-shaped network will produce information overactivity after the fusion of feature maps of different levels, a new decoder module is used here to refilter the fused feature maps to gradually recover the information in high-resolution remote sensing images. Finally, a new classfier is used in the output section to refine the final result.

B. Backbone

The extraction of feature information plays an important role in the segmentation task. Backbone network is utilized to extract rich feature information in remote sensing images, which can effectively extract the feature information of different objects in the image. The information extracted from the shallow network is not sufficient, and the available new information of the network is too little, which makes it impossible to make accurate judgment on the final result. However, the network with too deep layers will produce a lot of unnecessary information, resulting in excessive information, which will interfere with the final judgment result of the model. Therefore, the network with appropriate depth is very important for the feature extraction process of the whole network. We choose a residual network with 50 layers as the backbone network of this article for feature extraction. The residual connection mode in the residual network can effectively avoid gradient explosion, network degradation and other problems under the condition of increasing the depth of the network. In addition, after experimental comparison, we will finally prove that the 50-layer residual network can have the best effect for our task.

C. Adaptive Anomalous Window Attention Module (ASWA)

Owing to the increase of the number of network layers, more feature information is extracted, and the types are suddenly increased. Thus, effective processing means play an important role in the final result of the whole network. Traditional convolution windows are square structures of $k \times k$, which can pay attention to the local information of a certain area in the image. In addition, it can determine the relationship between each pixel and the surrounding pixels in a certain area. However, in practical applications, the relationship between global information is ignored, where the relationship between long distance pixels cannot be established.

To solve the above problems, we added ASWA in the deep layer, as shown in Fig. 2. The whole module adopted a multibranch structure, and the convolutional window in the trunk branch was different from the traditional $k \times k$ form. We adjusted the window size adaptively with the width and height of the feature map, and the window showed a rectangular shape. We used a total of two window forms in different directions to extract the longitudinal information and horizontal information in the image, respectively. In this trunk branch, the input feature map is first split and the information is extracted based on different directions. Convolution is followed by a maximum pooling layer. The research shows that pooling can further reduce the computation while preserving the main features. Moreover, it prevents overfitting and increases the translation invariance of image features, where it enhances the robustness of image to translation, rotation and scale transformation. The reason why maximum pooling is used instead of average pooling is that we hope to mainly reduce the deviation of estimated mean caused by the parameter error of the previous convolutional layer, retaining more texture information in the feature map, and filtering out useless interference information. In order to make

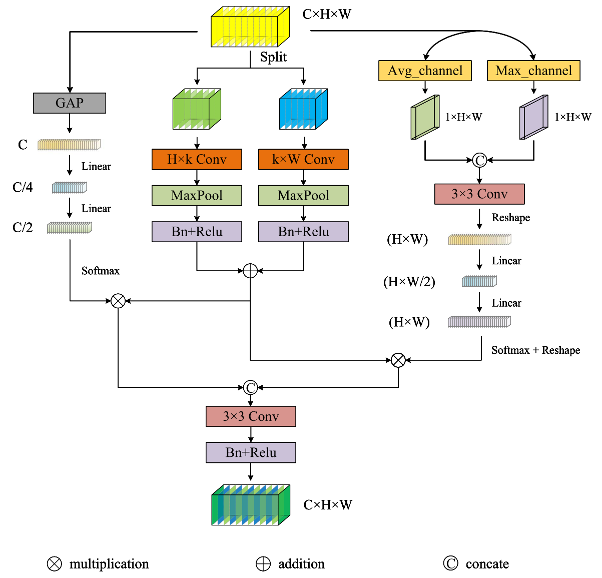

Fig. 2. Structure of adaptive anomalous window attention module.

the output more stable and increase the nonlinear performance of the model, batch normalization layer and nonlinear activation function are added to the end of each convolution layer, which is effective to speed up the learning speed of the model and restrain overfitting to a certain extent. Finally, the output features extracted in different directions are added and fused to obtain the output features of trunk branches. The calculation process is as follows:

$$X_1, X_2 = Split(X) \tag{1}$$

$$F_1 = \sigma(Bn(MaxPool(Conv_{H\times K}(X_1)))) \tag{2}$$

$$F_2 = \sigma(Bn(MaxPool(Conv_{K\times W}(X_2)))) \tag{3}$$

where $X \in R^{C\times H\times W}$ represents the input characteristics of the entire module, $F_1 \in R^{\frac{C}{2}\times H\times W}$ and $F_2 \in R^{\frac{C}{2}\times H\times W}$ represents the output feature graph obtained after operation based on different directions, $Split(\cdot)$ represents a split operation based on channel dimensions, $Conv_{n\times m}(\cdot)$ represents a convolution operation with window size $n \times m$, $MaxPool(\cdot)$ indicates the maximum pooling operation, $Bn(\cdot)$ and $\sigma(\cdot)$ represents the batch normalization layer and the nonlinear activation function ReLu.

With the deepening of the number of layers in the network, the amount of information generated will continue to increase, and how to make effective use of these information is also crucial. Simple convolution pays the same attention to each pixel in the image, which makes the model unable to focus on the target area. Currently, attention mechanisms are commonly used in visual tasks because they allow our models to focus on the meaningful parts of the image and ignore the distractions of useless information. Therefore, two branches are mapped here, which are, respectively, applied to extract the weight of channel information and spatial information contained in the deep features. The calculation process is as follows:

$$Ch\_attn = \text{Soft max}(Linear(GAP(X))) \tag{4}$$

$$Spatial = Cat(Avg(X), Max(x)) \tag{5}$$

$$Sp\_attn = \text{Soft max}(Linear(Conv_{3\times 3}(Spatial))) \tag{6}$$

where Linear(.) indicates a fully connected operation, GAP(.) indicates the global average pooling operation, Avg(.) and Max(.) represents average and maximum pooling operations based on channel dimensions, Cat(.) represents concatenation based on channel dimensions, $Conv_{3\times 3}(\cdot)$ denotes a convolution operation with a convolution kernel size of 3 × 3, Softmax(.) is the nonlinear activation function Softmax. In the channel attention information extraction branch, we first apply global average pooling to extract the global information. In order to effectively calculate the channel attention, the dimension of the feature map is adjusted, and the dimension of the feature map is compressed through two layers of full connection to filter out useless information. After the nonlinear activation function Softmax, the channel information weight in the original feature map is obtained. In the spatial information extraction branch, we adopt average pooling and maximum pooling operations along the direction of channel dimension, and then combine the two obtained feature maps to form an effective representation feature map of spatial information. In order to make the representation of spatial information more efficient, we also use a layer of convolution and two layers of full connection to adjust the feature maps. The filtering process of convolution makes the information representation more effective. Therefore, the compression and amplification of the feature dimensions can filter out many useless representations. Furthermore, the spatial information weight attention diagram in the original feature graph is obtained after the dimension of the feature graph is adjusted by the nonlinear activation function Softmax.

Finally, we multiply the obtained attention diagram of channel information weight and spatial information weight with the output feature map of the trunk branch, respectively, obtaining two feature maps containing channel information concern and spatial information concern. Since the addition operation may destroy the position information, we chose to concatenate the two feature graphs in a cascading way, so as to retain the spatial position information to the maximum extent. Then, a layer of 3 × 3 convolution is used to adjust the channel to get the final result, which is calculated as follows:

$$ch = Ch\_attn \otimes (F_1 + F_2) \tag{7}$$

$$sp = Sp\_attn \otimes (F_1 + F_2) \tag{8}$$

$$Output = \sigma(Bn(Conv_{3\times 3}(Cat(ch, sp)))) \tag{9}$$

where Cat(.) represents concatenation based on channel dimensions, $Conv_{3\times 3}(\cdot)$ denotes a convolution operation with a convolution kernel size of 3 × 3, $Bn(\cdot)$ and $\sigma(\cdot)$ represents the batch normalization layer and the nonlinear activation function ReLu.

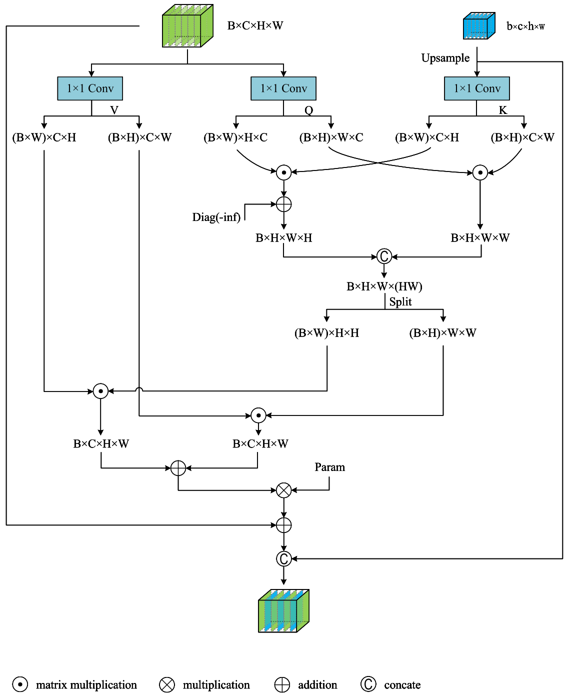

Fig. 3. Structure of multiangle interactive attention fusion module.

D. Multiangle Interactive Attention Fusion Module (MIAF)

In the task of land cover segmentation, the resolution of aerial remote sensing image is getting higher, and the span of time and space is also getting larger. Objects of the same category may produce different spectral phenomena in different areas, so it is not enough to accurately segment target objects by relying solely on spectral characteristics. In remote sensing images, the amount of information contained is very large, and the occupation of land by various objects is very complex, which greatly increases the difficulty of the model to distinguish different types of targets. For example, buildings of different heights will cast different shadows at different angles, and the targets in the shadows are easy to be misclassified as background. On the other hand, woodland and undergrowth have similar characteristics and attributes, which is easy to be confused. Complex background interference can also have a big impact on the final result.

In view of the above problems, it is far from enough to focus only on the association of local features in the image by relying on convolution, but also on the association in the global scope of the image. We propose an MIAF for blending feature maps at different levels, using deep features to guide classification of shallow features, and focusing on the relationship between global information in different directions in feature maps. Fig. 3 shows the internal structure of MIAF. The whole module is based on the attention mechanism. First, the upsampling operation is carried out on the deep feature graph to change it into the same dimension size of the shallow feature. Then, three 1 × 1 convolution operations are used to obtain the corresponding query (Q), key (K), and value (V), in which Q, K, V $\in R^{B\times C\times H\times W}$ are transformed in different directions. We believe that the deep feature map contains more category information. Therefore, keys in different directions are extracted from deep features and multiplied with query in shallow features to form an attention weight matrix. Different from the traditional attention mechanism that calculates the attention of the whole image at one time, we calculate the attention weights in the vertical and horizontal directions of the image, respectively, which can significantly reduce the amount of computation. However, separate horizontal or vertical calculation will cause the problem that other positions cannot be related. Therefore, we will finally fuse the weight feature graphs calculated from different angles to make up for this problem.

In order to make the final results more accurate, we add a learnable parameter matrix (Param) to adjust the results. The splicing operation at the end of the final process can keep some important information in the deep features. In this way, the category information in the high-level features can be used to strengthen the extraction ability of the model, so as to avoid the mutual interference between different categories and focus on the connection between the global scope. At the same time, the rich semantic information in the high-level features can be preserved, which is beneficial for the model to distinguish the difference between different objects. The calculation process of the module is as follows:

$$Q_{\mathrm{i}} = f_i(Conv_{1\times1}(X_1)), i = 1, 2 \tag{10}$$

$$K_i = f_i(Conv_{1\times1}(Up(X_2))), i = 1, 2 \tag{11}$$

$$V_i = f_i(Conv_{1\times1}(X_1)), i = 1, 2 \tag{12}$$

$$attn = Cat(Q_1 \odot K_1, Q_2 \odot K_2) \tag{13}$$

$$attn1, attn2 = Split(attn) \tag{14}$$

$$attn\_out = (attn1 \odot V_1 + attn2 \odot V_2) \otimes Param \tag{15}$$

$$Outout = Cat(Up(X_2), (X_1 + attn\_out)) \tag{16}$$

where X_1 and X_2 represents shallow feature and deep feature, Up(.) represents the upsampling operation, $Conv_{1\times1}(\cdot)$ represents a convolution operation with a convolution kernel of size 1 × 1, $f_i(\cdot)$ is stretching of dimensions along both vertical and horizontal lines, Cat(.) denotes concatenation based on channel dimensions, $Split(\cdot)$ represents a split operation based on channel dimensions, Param(.) is a learnable parameter matrix.

E. Decoder Module (De)

We first downsample the image to extract feature information, and then upsample the reduced feature map to restore the original size. In this process, the same feature fusion method as UNet can effectively avoid the loss of high-resolution details. However, if the fused features are not effectively processed, there will still be a lot of information interference. At the same time, in the task of land cover classification, there are certain requirements for the efficiency of feature extraction of the model. Here, we propose a new decoder module for the secondary processing of the fused feature map.

Fig. 4 shows the structure of the decoder module proposed in this article. The overall structure is residual structure, in

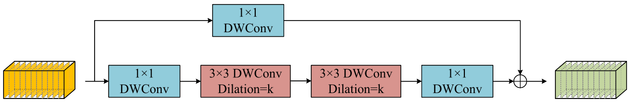

Fig. 4. Structure of decoder module.

which we add a two-layer cavity convolution module. And we stack multiple cavity convolution layers, increasing the perception field. Voidness convolution can increase the receptive field without increasing the parameters. A larger receptive field means that more information is received, which is beneficial for the high-resolution image recovery process. In order to further reduce the complexity of the model, detachable convolution in depth direction is widely used. Here, we use deep separable convolution to replace traditional convolution methods. Deep separable convolution can reduce the number of parameters to some extent, especially in deep networks. If only one kind of attribute is extracted, the influence of deep separable convolution is small, even worse than that of ordinary convolution. However, with the increase of the types of attributes extracted, depth-separable convolution gradually shows its advantages. In the multiclassification tasks solved in this article, detachable convolution can give full play to its advantages and extract more details better while reducing the amount of computation. The overall calculation process of decoder module is as follows:

$$F_{(X_i)} = DWConv_{1\times1}(D(D(DWConv_{1\times1}(X_i)))) \tag{17}$$

$$X_{i+1} = F_{(X_i)} + DWConv_{1\times1}(X_i) \tag{18}$$

where $DWConv_{1\times1}(\cdot)$ represents a deeply separable convolution operation with a convolution kernel size of 1 × 1, D(.) represents a depth-separable convolution operation with void convolution, X_i is the input of the current module, X_{i+1} denotes the output of the current module.

Inspired by the "bottleneck module," channel compression is first carried out on the fused feature map, which can greatly reduce the calculation amount of the model. Then, hierarchical calculation is carried out on the feature map through the two-layer depth separable cavity convolution to retain the important information in it. Finally, channel restoration is carried out. The operation of compression and restoration of the channel means that the computation of the model can be greatly reduced while the accuracy is maintained, which is meaningful for deep networks.

F. Experimental Method

All contents of this experiment are based on Pytorch deep learning framework with version 1.10.0 and Python version 3.8.12. We trained our model with an NVIDIA GeForce RTX 3090 card with 24 GB of video memory. In the training process, the StepLR adjustment strategy was adopted, and the learning rate was gradually reduced with the increase of training times to achieve good training effect. Among them, the initial value of learning rate at the beginning of the training was set as 0.0005, the attenuation coefficient was 0.98, and the learning rate was updated every three rounds of training, a total of 300 rounds of training. The formula for calculating the learning rate of each training round is as follows:

$$lr_N = lr_0 \cdot \beta^{N/s} \tag{19}$$

where lr_N is the learning rate of the N training, lr_0 is the initial learning rate, β is the attenuation coefficient, and *s* is the renewal interval. Loss function cross entropy loss function is selected, and the calculation formula is as follows:

$$\begin{aligned} Loss(x, class) &= -\log\left(\frac{e^{x[clas]}}{\sum_i e^{x[i]}}\right) \\ &= -x[class] + \log\left(\sum\nolimits_i e^{x[i]}\right) \end{aligned} \tag{20}$$

where x is the output tensor of the network and class is the real label. Adam algorithm [44] is a stochastic objective function first-degree optimization algorithm based on adaptive estimates of lower-order moments. Due to its simple implementation and high computational efficiency, Adam is very suitable for problems with large data volume or nonstationary targets and coefficient gradient, so we choose Adam as our existing optimizer.

In order to evaluate the actual performance of the model, we used category average pixel accuracy (MPA), F1, weighted crossover ratio (FWIOU), and average crossover ratio (MIOU) as indicators to evaluate the performance of the model. The corresponding calculation formula is as follows:

$$P = \frac{p_{ii}}{p_{ii} + p_{ij}} \tag{21}$$

$$R = \frac{p_{ii}}{p_{ii} + p_{ji}} \tag{22}$$

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{23}$$

$$MPA = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{24}$$

$$FWIoU = \frac{1}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}}\sum_{i=0}^{k}\frac{\sum_{j=0}^{k} p_{ij}\, p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{25}$$

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{26}$$

where *P* is the Precision, representing the probability that a certain category in the prediction result is predicted correctly; *R* is the Recall, denoting the probability that a certain category in the real value is predicted correctly; *k* is the number of cloud and cloud shadow (excluding background). p_{ii} represents the number of pixels belonging to class *i* and predicted to be class *i*; p_{ij} is the number of pixels belonging to class *i*, but predicted to be class *j*; p_{ji} represents the number of pixels belonging to category *j*, but predicted to be category *i*.

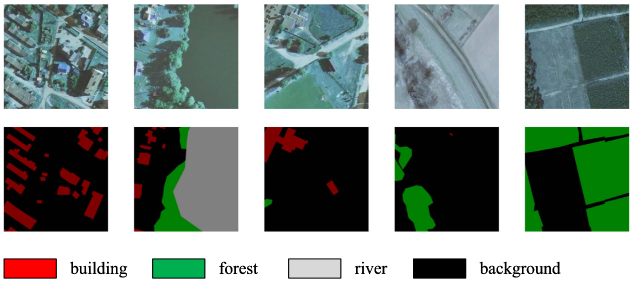

Fig. 5. Part of the training data and its labels.

IV. Experimental

A. Datasets

1) LandCover Dataset: This dataset [45] is derived from aerial photos in Poland, a Central European country, covering a total of 216.27 square kilometers of land, featuring high resolution and multitemporal distribution. The dataset contains 33 images with a resolution of 25 cm and 8 images with a resolution of 50 cm. The entire dataset is manually labeled into four categories: buildings (red), forest land (green), water (gray), and background (black). Due to the limitation of computer hardware, we clipped the dataset and uniformly clipped all the pictures to 512 × 512 size. If the clipped pictures only contain a class of objects, we removed them and then, randomly divided all the pictures into the training set and the verification set according to the ratio of 8:2. As shown in Fig. 5, we show part of the training data and its corresponding labels.

2) Postdam Dataset: This is a land cover classification dataset obtained from airborne sensors [46]. The challenging part of this dataset is that buildings, cars, and other objects have very heterogeneous appearance in high-resolution images. This results in high intraclass variance and low inter-class variance.

The dataset was collected in Potsdam, a typical historical city with very large buildings, narrow streets, and dense residential structures. The datasets are manually divided into six categories as follows:

1) impervious surface;
2) building;
3) low vegetation;
4) tree;
5) clutter/background;
6) car.

Among them, clutter / background includes water bodies and other objects that look very different from other categories (such as containers, tennis courts, swimming pools, etc.), as given in Fig. 6 , which shows part of the training data in the dataset and their corresponding label images.

3) Wuhan Dense Labeling Dataset (WHDLD): WHDLD [47] is the third intensively labeled dataset we used. It was cropped from a large volume remote sensing image of Wuhan urban area and manually labeled into six categories as follows:

1) building;
2) road;

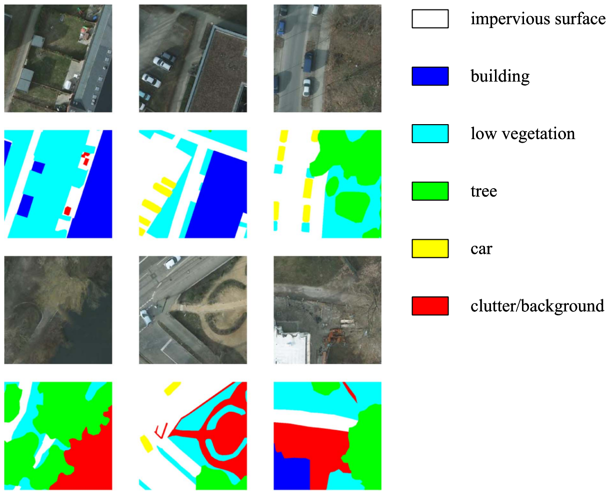

Fig. 6. Part of the training data and its labels.

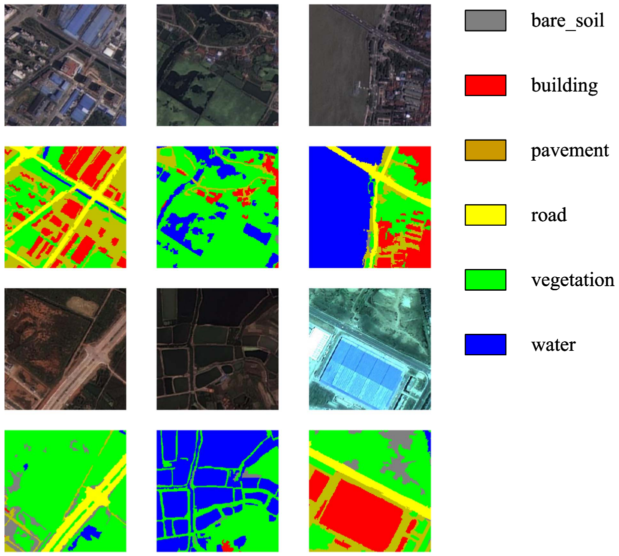

Fig. 7. Part of the training data and its labels.

3) pavement;
4) vegetation;
5) bare soil;
6) water.

The WHDLD contains a total of 4940 color images with a resolution of 2 m and a size of 256 × 256. Fig. 7 shows some of the images in the dataset and their corresponding labels.

B. Experimental Results

1) Ablation Experiments: In this part, we test the influence of feature extraction networks of different depths on the final effect of the model, and conduct ablation experiments on our model on three datasets, aiming to show the actual influence of each module more clearly. Here, MPA, MIOU, and FWIOU are used as evaluation indicators to evaluate the final results.

TABLE I
IMPACT OF DIFFERENT BACKBONE NETWORKS ON MODEL ACCURACY (BEST RESULTS ARE SHOWN IN BOLD)

Backbone	MPA(%)	MIOU(%)	FWIOU(%)
resnet18	91.346	86.000	89.713
resnet34	91.952	86.761	90.036
resnet50	91.864	**87.128**	**90.290**
resnet101	**92.109**	87.005	90.180

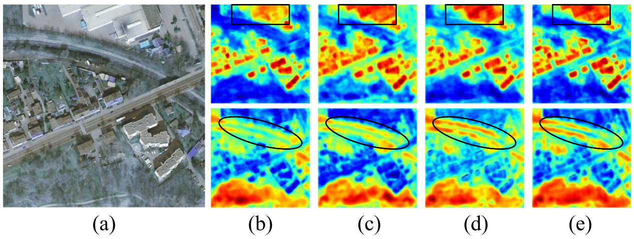

Fig. 8. Visual comparison of different module combinations. The first row is the thermal map of the building. The second row is the heat map of forest land. (a) Image; (b) Backbone; (c) Backbone+De; (d) Backbone+De+ASWA; (e) Backbone+De+ASWA+MIAF.

We first replaced the feature extraction network (backbone network) of the model with residuals of different depths and conducted experiments on landcover dataset. Table I shows the final results. As you can see from the table, more layers is not always better. When the number of layers reaches 50, it has the best effect. And as the number of layers continues to deepen, the effect shows a regressive phenomenon. After experiments, we finally choose a 50-layer residual network as our backbone network for feature extraction operation.

As shown in Table II, the backbone network acts as a feature extraction network, so other modules are first removed or replaced with the simplest connection structure. In the deep part of the network, the ASWA module is directly removed, and the MIAF module is replaced by the simplest linear splicing operation for fusion. The De is also removed at the decoding side, leaving only the backbone part as the benchmark to test the evaluation score of the model at this time. Then, the three modules De, ASWA, and MIAF are successively added into the benchmark network for testing. It can be concluded from Table II that for the land division task, the three models proposed are effective in improving the final accuracy. We extracted a picture from the landcover dataset for visualization experiment, as shown in Fig. 8, which shows the thermal map generated by the combination of different modules. From the black solid line area in the figure, it can be seen that with the continuous addition of modules, the attention to the target is more concentrated. When only the benchmark network is available, the attention of the model is scattered.

Ablation experiment of De: Restoring the deep feature map obtained by network downsampling is an important step in the remote sensing image segmentation task. An effective decoder can restore the details of the high-resolution image as much as possible, and eliminate the interference of useless information in the process of restoring the original image, which is effective to improve the accuracy of the model. We build a new decoder module, in which dilated convolution makes it have a larger receptive field, and depthwise separable convolution has better results than ordinary convolution for the multiclassification task studied in this article. It can be seen from the table that after adding our proposed decoder module, the segmentation accuracy of the model has a certain improvement. As can be seen from Fig. 8, after adding the De module, the attention to useless information is less than the situation without adding it.

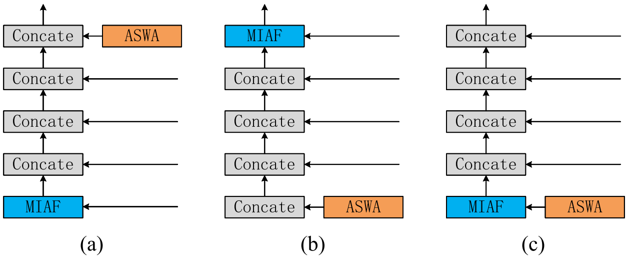

Fig. 9. Display of MIAF and ASWA in different stages of the decoder. (a) ASWA is located in the shallow layer of the network. (b) MIAF is located in the shallow layer of the network. (c) ASWA and MIAF are located in the deep layer of the network.

Ablation experiments for ASWA: Aiming at the problem of insufficient processing of deep feature information in current methods, we propose the ASWA module, which adopts a larger convolution window to focus on global feature information, which is proved to be effective by experiments. It can be seen from Fig. 8 that after adding the ASWA module, the attention to the background is significantly reduced, and the model can pay more attention to the buildings we want to extract. The scores on the three datasets in Table II also indicate the effectiveness of the ASWA module. In Fig. 9, (a) and (c) show the situation of this module in different stages at the decoder side. In Table III, (a) and (c) are the corresponding evaluation scores of the model on the Landcover dataset when the module is placed in different stages. When the ASWA module is in a shallow level, there will be a sudden increase in the number of parameters and it will be at a low level in all indicators. This shows that placing ASWA in the deep layer of the network can effectively use the deep information to improve the network performance.

Ablation experiments of MIAF: In order to make full use of the deep feature information of different levels, we propose MIAF to fuse the feature maps of the last two deep layers of the network, because we believe that these two layers contain the most information. The effective fusion of deep features can avoid the mutual interference between different categories. The results shown in Fig. 8 show that after adding the MIAF module, information such as the position and shape of the target can be accurately extracted from the surrounding similar background. Without processing the deep information, the model is easy to be disturbed by the similar environment around it and make wrong judgments on the target. Table II shows that after adding all the modules, the overall network achieves the highest accuracy and has the best effect on the three datasets. In Fig. 9, (b) and (c) show the different stages of this module at the decoder side. In

TABLE II
ABLATION EXPERIMENTS ON DIFFERENT DATA SETS (THE HIGHEST SCORE IS BOLDLY DISPLAYED)

Datasets	Methods	MPA(%)	MIOU(%)	FWIOU(%)
Landcover	Backbone	91.016	85.030	89.126
	Backbone+De	91.235	86.283	89.740
	Backbone+De+ ASWA	91.668	86.770	90.143
	Backbone+De+ASWA+MIAF	**91.864**	**87.128**	**90.290**
Postdam	Backbone	81.703	73.380	81.965
	Backbone+De	82.910	73.690	81.557
	Backbone+De+ASWA	83.338	74.363	81.602
	Backbone+De+ASWA+MIAF	**84.057**	**74.756**	**82.280**
WHDLD	Backbone	76.300	63.748	81.920
	Backbone+De	75.570	63.770	82.616
	Backbone+De+ASWA	76.351	63.881	82.070
	Backbone+De+ASWA+MIAF	**76.412**	**64.050**	**82.286**

TABLE III
ABLATION EXPERIMENTS OF MIAF AND ASWA AT DIFFERENT STAGES OF THE DECODER

Module	Stage	MPA(%)	MIOU(%)	FWIOU(%)	Params(M)	Flops(G)
ASWA	a	89.237	86.384	90.174	73.87	15.384
	c	**91.864**	**87.128**	**90.290**	50.372	11.276
MIAF	b	90.906	85.986	89.347	88.985	20.48
	c	**91.864**	**87.128**	**90.290**	50.372	11.276

Table III, (b) and (c) are the corresponding evaluation scores of the model on the Landcover data set when the module is placed in different stages. Due to the attention mechanism used in MIAF to focus on global information, its computational complexity and parameter amount are closely related to the feature map size. Being close to the output incurs a lot of extra computational overhead, which is not beneficial. At the same time, according to the final evaluation results of the two methods, the fusion effect near the output is significantly worse than that when it is placed in the deep layer of the network.

2) Experiments on the LandCover Dataset: In this section, we compare the proposed method with other advanced networks, including classic segmentation networks and the best segmentation networks in recent years. In order to verify the advantages of this article in the land cover type classification task, networks that performs well in this task in recent years is also added for comparison. We use F1 score to evaluate the classification results of our model and other networks between different categories. Table IV shows the comprehensive scores of different models on this dataset, where MPA, MIOU, and FWIOU are used as evaluation indicators for evaluation. Compared with other types of models, the network proposed in this article has the highest comprehensive score, in which the MIOU value is 87.133%, where it is 1.132% higher than the best network, and the comprehensive performance is far better than the existing segmentation networks.

In order to verify the actual performance of the proposed model, as shown in Fig. 10, images located in different environments are picked for actual prediction. At the same time, for comparison, networks with different strategies are selected for prediction, and their prediction results are compared with ours. For example, UNet adopts the same encoder-decoder structure as this article, PSPNet adds a pyramid pooling module to aggregate context information in different regions to mine global context information, PVT _ m introduces the pyramid structure into transformer, and uses a gradually reduced pyramid to reduce the calculation of large feature maps. ACFNet proposes the concept of class center, and uses the attention class feature module (ACF) to combine different class centers according to the adaptation of each pixel.

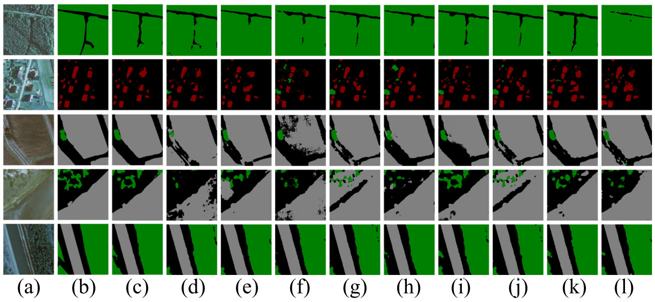

Fig. 10. Prediction effects of different models on landcover dataset. (a) Image; (b) Label; (c) Ours; (d) UNet; (e) PSPNet; (f) PVT _ m; (g) ACFNet; (h) LEDNet; (i) FCN8s; (j) BiseNetV2; (k) DeepLabV3Plus; (l) SegNet.

In the selected images, various scene categories are included. The first picture is taken from a dense forest area, and there will be gaps at the junction of the number of different regions. This part does not belong to forest land, but because the area is too small, most models will not detect it and misclassify it as forest land, such as SegNet, PSPNet, and other models. The second picture shows a building group. The low green vegetation next to the building has similar color attributes to trees, which is easy to be misclassified as woodland. For example, this phenomenon occurs in models such as PVT _ m and ACFNet. The third and fourth images are located next to the water area, and the overall

TABLE IV
EVALUATION RESULTS ON THE LANDCOVER DATASET (BOLD CHARACTERS REPRESENT THE HIGHEST SCORES, THE SECOND BEST SCORE IS UNDERLINED)

Methods	F1(%)				MPA(%)	MIOU(%)	FWIOU(%)
	void	building	forest	river			
SegNet [48]	93.074	68.859	91.047	91.815	90.799	76.996	85.264
DFANet [49]	93.156	68.493	90.780	92.382	91.185	77.058	85.252
ESPNetv2 [50]	94.259	67.091	92.744	93.629	**94.415**	78.528	87.687
GhostNet [51]	93.927	68.597	92.011	93.922	90.456	78.624	86.960
CVT [40]	93.687	69.822	92.086	93.733	88.279	78.824	86.761
DeepLabV3Plus [33]	92.705	77.186	89.889	92.563	87.713	79.260	84.421
ShuffleNetV2 [52]	93.190	77.977	90.967	90.617	92.421	79.356	85.298
BiSeNetV2 [32]	93.987	76.560	92.204	93.237	91.038	80.886	87.153
DFN [53]	94.162	76.155	92.029	94.027	91.247	81.106	87.318
PAN [54]	94.056	76.606	92.395	93.680	90.665	81.210	87.399
ENet [55]	94.377	77.184	92.788	93.455	92.155	81.615	87.940
UNet [30]	94.174	80.179	92.132	93.835	94.226	82.426	87.425
FCN8s [29]	94.508	78.398	92.781	94.539	91.890	82.559	88.228
LEDNet [56]	94.540	79.297	92.801	94.703	91.092	82.963	88.308
ACFNet [57]	94.519	80.949	92.974	93.533	91.689	83.081	88.263
DDRNet [58]	94.568	79.380	92.859	95.162	92.525	83.237	88.436
OCRNet [59]	94.701	79.467	92.902	94.930	92.739	83.240	88.565
ERFNet [60]	94.573	80.067	92.631	95.045	93.139	83.324	88.293
PVT_s [41]	94.256	81.395	92.254	94.883	91.129	83.412	87.740
PVT_l [41]	94.350	81.364	92.584	94.681	91.679	83.495	88.011
PVT_m [41]	94.357	81.694	92.526	94.524	91.998	83.520	87.965
DenseASPP [61]	94.902	78.755	93.373	95.479	93.143	83.543	89.128
CCNet [62]	94.954	80.532	93.422	95.462	93.162	84.194	89.235
DABNet [63]	94.862	81.069	93.283	95.819	93.316	84.444	89.112
PSPNet [31]	95.215	82.193	93.637	95.779	93.249	85.143	89.703
CGNet [64]	95.268	82.892	93.702	95.988	93.692	85.546	89.838
MSFANet [36]	94.942	83.867	92.673	95.833	93.325	85.021	90.035
MLNet [35]	94.143	82.674	92.933	95.354	92.582	84.938	89.412
DBPNet [42]	95.769	83.243	93.864	95.001	93.796	86.001	90.162
SGBNet [34]	94.757	82.734	92.434	94.348	92.971	85.479	89.374
Ours	**95.485**	**86.258**	**93.955**	**96.230**	94.214	**87.133**	**90.307**

color attribute is dark, which leads to confusion between the water area and other parts and is difficult to distinguish. It can be seen from the figure that although most of the models can locate the position of the water area, the edge of the water area is easy to be confused with the background, so the segmentation effect of the boundary is very rough. Our model can accurately identify the water area, and the segmentation effect of the edge is in line with the actual situation, which can avoid the influence of the surrounding environment.

From the final prediction results, other models are more or less affected by interference factors due to improper processing of deep information. Due to the addition of ASWA and MIAF modules to filter deep feature information, the model proposed in this article can effectively filter out interference factors in the picture, so that the model can focus on more important information. The ability of our network to distinguish between different categories is much better than other models, such as the distinction between woodland and low vegetation, the distinction between water and similar background, etc. This is because the MIAF module effectively extracts category information and different layers of the network usually contain different degrees of information. The MIAF module can fully integrate the category information contained in different layers, and the correlation between global information can be paid attention to through the attention mechanism in different directions, which is useful to avoid the interference of similar features between different categories. It can also be seen from the figure that the network in this article has a significantly better ability to recover the target edge than other models, which is due to the fact that we also use a new classifier after the last upsampling to refine the final result. Not only that, it can be seen from the figure that our model has good detection results for objects of different scales, which can be seen in the prediction results of dense buildings in the second picture and scattered trees in the third and fourth pictures in Fig. 10.

3) Experiments on the Postdam Dataset: In order to further prove the effectiveness of the model proposed in this article, this part carries out comparative experiments on the Postdam dataset. We evaluate the F1 score of different models for different targets. In addition, we also evaluate the comprehensive performance scores of each model, as shown in Table V, and we use MPA, MIOU, FWIOU as evaluation metrics to calculate the comprehensive scores of different models on this dataset. It can be seen from the table that our model has the highest MIOU value and FWIOU value, indicating that the method proposed in this article has good performance both for single object classification and for the comprehensive classification ability of the whole image.

TABLE V
EVALUATION RESULTS ON THE POSTDAM DATASET (BOLD CHARACTERS ARE THE HIGHEST SCORES, THE SECOND BEST SCORE IS UNDERLINED)

Methods	F1(%)						MPA(%)	MIOU(%)	FWIOU(%)
	Surface	Building	Vegetable	Tree	Car	Clutter			
CVT	87.532	91.660	82.577	67.276	75.557	35.568	74.671	60.965	72.436
DFANet	89.288	91.580	83.409	65.176	80.521	50.506	80.815	64.363	73.832
GhostNet	87.885	91.398	84.310	74.226	74.672	52.407	77.484	64.921	74.527
PVT_s	88.547	91.411	84.000	69.698	85.248	43.576	78.802	65.280	74.186
PVT_m	89.151	92.050	84.590	70.528	86.496	41.407	79.314	65.963	75.120
DFN	88.632	92.391	84.965	75.012	83.543	44.121	81.535	66.560	75.756
PVT_l	89.608	92.052	85.155	71.623	85.975	47.135	80.152	67.103	75.865
ESPNetv2	90.172	93.842	84.896	73.494	83.797	55.669	82.889	68.839	77.199
ShuffleNetV2	90.748	92.616	86.254	77.724	84.801	51.686	83.532	69.528	78.228
SegNet	91.405	94.424	85.797	75.383	87.506	47.590	83.737	69.706	78.757
LEDNet	91.361	94.034	87.111	76.339	85.217	53.340	84.809	70.391	79.300
ENet	91.567	94.328	86.200	75.836	87.013	53.364	83.793	70.657	79.152
CGNet	91.227	94.097	86.379	77.615	86.062	53.563	82.591	70.713	79.196
FCN8s	91.395	93.718	85.936	77.506	87.939	53.071	82.589	70.924	79.002
ERFNet	91.266	93.900	86.688	76.931	88.003	53.565	82.581	71.101	79.202
DeepLabV3Plus	91.553	95.296	86.452	78.025	87.489	50.344	83.151	71.157	79.940
PAN	91.273	94.269	86.635	78.335	86.356	54.351	84.716	71.203	79.537
DenseASPP	92.121	95.509	86.434	77.083	87.267	52.127	83.974	71.380	80.243
DABNet	91.813	95.525	86.680	77.459	88.028	53.592	83.517	71.870	80.269
BiSeNetV2	91.811	94.946	87.261	78.594	87.132	53.666	84.597	71.875	80.434
OCRNet	90.717	92.885	87.181	78.339	87.674	**60.018**	85.177	72.054	79.039
UNet	91.791	94.815	86.966	78.833	89.016	56.450	84.374	72.750	80.406
DDRNet	91.944	95.108	87.264	77.336	89.083	59.824	84.823	73.201	80.560
PSPNet	92.336	**95.976**	87.570	80.093	87.137	57.195	**85.366**	73.328	81.633
ACFNet	92.089	95.836	87.269	78.425	87.712	59.805	84.226	73.339	81.089
CCNet	92.256	95.868	87.432	80.122	87.436	58.711	84.979	73.571	81.532
MSFANet	91.254	95.783	87.326	77.853	88.936	58.642	84.652	73.121	81.382
MLNet	92.493	94.050	86.748	78.671	87.635	57.753	85.253	72.643	80.682
DBPNet	91.280	94.782	86.693	79.138	89.397	58.993	83.561	72.231	80.236
SGBNet	91.759	93.982	87.162	77.263	88.735	57.652	84.629	71.472	81.273
Ours	**92.510**	95.558	**88.485**	**81.367**	**89.691**	58.893	84.941	**74.757**	**82.276**

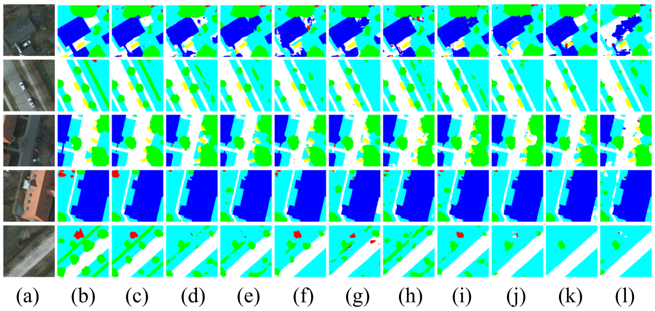

Fig. 11. Prediction effects of different models on Postdam dataset. (a) Image; (b) Label; (c) Ours; (d) ACFNet; (e) PSPNet; (f) UNet; (g) BiseNetV2; (h) DeepLabV3Plus; (i) FCN8s; (j) LEDNet; (k) SegNet; (l) PVT _ m.

In Fig. 11, we selected pictures in different environments for testing experiments, such as concentrated building areas, rural roads, suburbs, etc., where buildings contain houses of different scales and trees are distributed in different densities. The low-rise vegetation in the picture has similar texture features with trees, which will also cause some interference to the classification results of the model. As can be seen from the figure, our model has excellent segmentation results for different categories, where it can accurately locate the target area, and the restoration of the shape of the target is in line with the actual situation. Other networks such as PVT _ m and SegNet have the worst prediction results, and the segmentation of the edge of the target is very rough. In the first image, PVT _ m cannot accurately segment the shape of the house at all, and there are a lot of false detection phenomena. Although LEDNet can detect the building area, the segmentation result of the building edge details is not ideal and the detection of vehicles is affected by the surrounding environment. Therefore, it cannot accurately restore the shape of the vehicle in the picture. There are many kinds of trees and vegetation in rural areas, so it is a challenging task to distinguish between them. For example, there is a forest belt composed of trees on the right of the second picture. However, many models are affected by the surrounding vegetation, and all of them fail to accurately distinguish the relationship between trees and vegetation. The same problem occurs in the last figure, with generally poor classification results for trees and vegetation. Looking at our proposed model, whether it is for the shape restoration of buildings, vehicles and other targets, or the discrimination between trees and vegetation, it has incomparable advantages to other models.

TABLE VI
EVALUATION RESULTS ON THE WHDLD DATASET (BOLD CHARACTERS REPRESENT THE HIGHEST SCORES, THE SECOND BEST SCORE IS UNDERLINED)

Methods	F1(%)						MPA(%)	MIOU(%)	FWIOU(%)
	bare_soil	building	pavement	road	vegetation	water			
GhostNet	53.198	64.465	40.406	57.446	90.383	88.934	67.657	51.991	73.986
CVT	54.490	66.390	42.533	65.142	91.429	91.131	69.219	55.062	76.267
DFANet	55.100	72.024	46.388	70.438	92.314	92.209	71.692	58.357	78.383
ESPNetv2	57.451	73.308	48.668	73.154	93.099	94.156	74.281	60.674	80.240
DenseASPP	56.638	74.626	50.285	73.625	92.910	93.987	72.472	61.048	80.150
PVT_l	56.967	73.915	50.134	73.335	93.333	94.377	74.723	61.108	80.677
SegNet	57.236	73.239	50.372	73.125	93.320	94.876	75.356	61.150	80.772
PVT_m	58.063	73.916	50.268	73.626	93.229	94.165	74.087	61.276	80.548
BiSeNetV2	58.246	74.361	49.033	73.750	93.305	94.266	74.745	61.296	80.666
ShuffleNetV2	60.337	74.063	49.895	73.514	93.193	94.113	74.103	61.584	80.565
CGNet	59.514	74.102	50.584	75.330	93.245	94.278	74.161	62.003	80.748
PVT_s	58.124	74.917	50.864	74.926	93.393	94.657	74.105	62.056	81.055
DFN	59.165	74.120	51.346	74.735	93.346	94.690	74.241	62.089	80.989
LEDNet	58.977	74.105	51.158	75.315	93.407	94.484	74.753	62.106	81.000
DABNet	59.426	73.824	50.446	75.729	93.385	94.672	75.049	62.155	81.012
DDRNet	59.390	74.764	49.073	75.904	93.369	94.672	75.024	62.177	81.028
PSPNet	59.147	74.690	49.464	76.086	93.492	94.697	74.957	62.261	81.182
PAN	58.158	74.642	50.630	75.592	93.567	95.005	75.796	62.267	81.354
CCNet	**60.731**	74.193	50.504	75.724	93.501	94.536	75.324	62.455	81.182
ENet	57.336	75.152	52.632	75.067	93.684	95.305	75.488	62.556	81.665
ACFNet	58.944	75.091	51.626	76.398	93.694	95.034	76.268	62.864	81.650
FCN8s	60.112	75.518	52.821	74.890	94.013	95.625	75.823	63.284	82.276
DeepLabV3Plus	60.210	75.556	52.732	76.171	93.760	95.444	76.060	63.441	81.982
OCRNet	59.044	76.143	53.186	76.267	93.953	95.609	75.802	63.569	82.291
UNet	59.078	75.863	54.390	**76.522**	94.056	95.529	**77.104**	63.763	82.420
ERFNet	60.524	75.859	53.897	76.187	93.940	95.561	75.984	63.832	82.314
MSFANet	60.182	75.832	53.758	75.265	93.487	95.623	74.672	62.833	82.181
MLNet	59.701	76.035	53.982	74.872	94.232	95.582	75.721	63.836	81.792
DBPNet	58.972	75.638	**54.823**	75.217	93.238	**95.753**	75.836	63.219	81.345
SGBNet	58.263	75.342	53.656	75.468	**94.347**	94.897	74.836	62.673	81.154
Ours	60.244	**76.401**	54.424	76.138	93.943	95.714	76.027	**64.022**	**82.430**

4) Experiments on the WHDLD: In this section, we use another multiclass land cover classification dataset (WHDLD) for testing, where the performance scores of different models are calculated. The overall evaluation scores of different models on this dataset are shown in Table VI, where the best scores are highlighted in bold. As can be seen from the table, our model also has good classification ability on this dataset, which has the highest score in the comprehensive evaluation. The MPA, MIOU and FWIOU values are 76.027 %, 64.022%, and 82.430%, respectively, which are higher than the scores of all other models.

In order to test the actual segmentation performance of the model, several pictures are also selected to show the prediction effect, and the prediction effect is shown in Fig. 12. The displayed images contain different kinds of distribution features, including dense buildings, narrow roads, indistinct water areas, etc. In this dataset, there is a strict requirement to distinguish between roads and paved roads. Since the distinction between roads and roads is not obvious in most datasets under normal circumstances, this is highly convincing to evaluate the actual performance of a model. As shown in the last image in Fig. 12, there are many long and narrow roads in the image. Our model can effectively identify the road area, but other models such as LEDNet and BiseNetV2 misclassify the road surface as a road. In the pictures to be predicted, the roofs of some buildings have similar texture features with the bare land, so it is easy to cause misclassification between the two. For example, in the first and second pictures, due to the influence of the surrounding ground, most models cannot accurately distinguish the contours of the buildings, especially for the small and scattered housing buildings. DeepLabV3Plus, BiseNetV2, and other networks have missed detection. Our model can accurately locate the location of the building, and the segmentation effect of

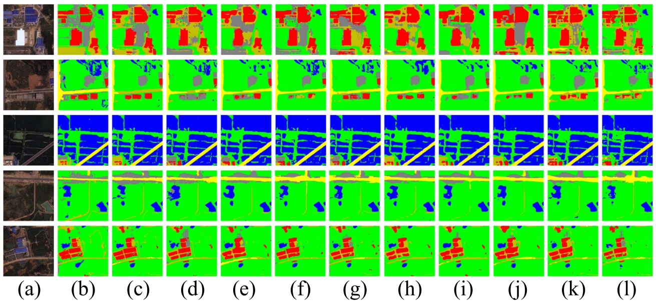

Fig. 12. Prediction effects of different models on WHDLD dataset. (a) Image; (b) Label; (c) Ours; (d) UNet; (e) LEDNet; (f) DeepLabV3Plus; (g) FCN8s; (h) ACFNet; (i) PSPNet; (j) BiseNetV2; (k) PVT _ m; (l) SegNet.

the building edge is also more in line with the actual situation. In the segmentation results of the narrow and long pavement, most of the other models have serious fracture phenomena, which cannot completely restore the area of the pavement, and even misclassify the pavement as vegetation. However, the method proposed in this article can recover the shape of the pavement completely and avoid the influence of the surrounding environmental factors. The prediction results show that the proposed method also has excellent classification ability for different land types on this dataset.

V. Discussions

This article mainly focuses on the fact that the current methods do not properly deal with the deep features of the network in the land cover classification task, which makes the model susceptible to the interference of different factors in the image and confusion between different categories. Therefore, ASWA and MIAF modules are proposed to deal with the deep feature information. The above experimental results based on different datasets show that the proposed model has a certain improvement effect on land cover classification tasks, and it has better land classification ability than the current model. The prediction results in Figs. 10–12 show that the comparison method is very rough for the classification of land cover types, and the processing of boundaries between different categories is not fine enough. In the classification process, targets with similar characteristics are easy to interfere with each other. As shown in Fig. 10, the narrow area between forest and forest in the figure is easy to miss detection, and low shrubs will interfere with forest, resulting in misjudgment. In Fig. 12, road and pavement have similar superior reflection characteristics, and many models cannot strictly distinguish them. Existing methods add attention mechanism to the network to focus on important features and prevent missing, however, it will lead to a sharp increase in the number of model parameters. Moreover, our proposed method also adopts the idea of attention mechanism and improves the structure, through the interaction between global information in different directions, it avoids the sharp increase in the amount of parameters while maintaining the original role. At the same time, we pay more attention to the interaction between the deep information of the network, which plays a certain role in containing the mutual interference between different categories.

VI. Conclusion

The classification of land cover types is an important branch of high-resolution remote sensing image processing, which has important guiding significance for the study of land use. In this article, we propose a multiview attention fusion network for land cover classification task. The proposed model is helpful to analyze the characteristics of land use types and biodiversity, and provides a basis for environmental quality, land planning, and land use development. This article uses the method of convolutional neural network to construct a new network, and most of the networks do not make full use of the information of deep features, which leads to the problem that the land type cannot be classified accurately. Furthermore, ASWA and MIAF modules are proposed. The proposed ASWA adopts a multibranch structure, and replaces the traditional convolution window with a rectangular window, which can focus on the global features in the nonpassing directions. MIAF is used to fuse the feature maps of different levels in the deep layer of the network, and make full use of the category information to guide the classification of the model. The new decoder module improves the efficiency of information transmission and also makes the final segmentation result more refined. The final experimental results show that the proposed model achieves the best classification performance on Landcover, Postdam, and WHDLD datasets, and the MIOU scores on the three datasets are 87.133%, 74.757%, and 64.022%, respectively. It also has strong generalization ability.

References

[1] Y. Qu, M. Xia, and Y. Zhang, “Strip pooling channel spatial attention network for the segmentation of cloud and cloud shadow,” *Comput. Geosci.*, vol. 157, 2021, Art. no. 104940. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0098300421002259

[2] S. Miao, M. Xia, M. Qian, Y. Zhang, J. Liu, and H. Lin, “Cloud/shadow segmentation based on multi-level feature enhanced network for remote sensing imagery,” *Int. J. Remote Sens.*, vol. 43, no. 15/16, pp. 5940–5960, 2022.

[3] E. Zhang, K. Hu, M. Xia, L. Weng, and H. Lin, “Multilevel feature context semantic fusion network for cloud and cloud shadow segmentation,” *J. Appl. Remote Sens.*, vol. 16, no. 4, 2022, Art. no. 046503. [Online]. Available: https://doi.org/10.1117/1.JRS.16.046503

[4] K. Hu, E. Zhang, M. Xia, L. Weng, and H. Lin, “McaNet: A multi-branch network for cloud/snow segmentation in high-resolution remote sensing images,” *Remote Sens.*, vol. 15, no. 4, 2023, Art. no. 1055. [Online]. Available: https://www.mdpi.com/2072-4292/15/4/1055

[5] J. Gao, L. Weng, M. Xia, and H. Lin, “MLNet: Multichannel feature fusion lozenge network for land segmentation,” *J. Appl. Remote Sens.*, vol. 16, no. 1, pp. 1–19, 2022.

[6] Z. Ma, M. Xia, L. Weng, and H. Lin, “Local feature search network for building and water segmentation of remote sensing image,” *Sustainability*, vol. 15, no. 4, 2023, Art. no. 3034. [Online]. Available: https://www.mdpi.com/2071-1050/15/4/3034

[7] B. Chen, M. Xia, M. Qian, and J. Huang, “Manet: A multi-level aggregation network for semantic segmentation of high-resolution remote sensing images,” *Int. J. Remote Sens.*, vol. 43, no. 15/16, pp. 5874–5894, 2022.

[8] L. Song, M. Xia, L. Weng, H. Lin, M. Qian, and B. Chen, “Axial cross attention meets CNN: Bibranch fusion network for change detection,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 32–43, 2023.

[9] H. Alemohammad and K. Booth, “Landcovernet: A global benchmark land cover classification training dataset,” 2020, *arXiv:2012.03111*.

[10] M. Li, S. Zang, B. Zhang, S. Li, and C. Wu, “A review of remote sensing image classification techniques: The role of spatio-contextual information,” *Eur. J. Remote Sens.*, vol. 47, no. 1, pp. 389–411, 2014.

[11] R. Pradhan, M. P. Pradhan, A. Bhusan, R. K. Pradhan, and M. K. Ghose, “Land-cover classification and mapping for eastern himalayan state Sikkim,” 2010, *arXiv:1003.4087*.

[12] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson, “Random forests for land cover classification,” *Pattern Recognit. Lett.*, vol. 27, no. 4, pp. 294–300, 2006.

[13] P. K. Goel, S. O. Prasher, R. M. Patel, J.-A. Landry, R. Bonnell, and A. A. Viau, “Classification of hyperspectral data by decision trees and artificial neural networks to identify weed stress and nitrogen status of corn,” *Comput. Electron. Agriculture*, vol. 39, no. 2, pp. 67–93, 2003.

[14] V. Eisavi, S. Homayouni, A. M. Yazdi, and A. Alimohammadi, “Land cover mapping based on random forest classification of multitemporal spectral and thermal images,” *Environ. Monit. Assessment*, vol. 187, pp. 1–14, 2015.

[15] J. D. T. De Alban, G. M. Connette, P. Oswald, and E. L. Webb, “Combined landsat and l-band SAR data improves land cover classification and change detection in dynamic tropical landscapes,” *Remote Sens.*, vol. 10, no. 2, 2018, Art. no. 306. [Online]. Available: https://www.mdpi.com/2072-4292/10/2/306

[16] J. R. Otukei and T. Blaschke, “Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms,” *Int. J. Appl. Earth Observ. Geoinformation*, vol. 12, pp. S27–S31, 2010.

[17] R. Manandhar, I. O. Odeh, and T. Ancev, “Improving the accuracy of land use and land cover classification of Landsat data using post-classification enhancement,” *Remote Sens.*, vol. 1, no. 3, pp. 330–344, 2009.

[18] S. Gul et al., “Monitoring of land use and land cover changes using remote sensing and geographic information system,” 2022.

[19] M. A. Friedl and C. E. Brodley, “Decision tree classification of land cover from remotely sensed data,” *Remote Sens. Environ.*, vol. 61, no. 3, pp. 399–409, 1997.

[20] D. McIver and M. Friedl, “Using prior probabilities in decision-tree classification of remotely sensed data,” *Remote Sens. Environ.*, vol. 81, no. 2/3, pp. 253–261, 2002.

[21] G. M. Foody and A. Mathur, “A relative evaluation of multiclass image classification by support vector machines,” *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 6, pp. 1335–1343, Jun. 2004.

[22] S. Yousefi et al., “Image classification and land cover mapping using sentinel-2 imagery: Optimization of SVM parameters,” *Land*, vol. 11, no. 7, p. 993, 2022.

[23] G. H. Halldorsson, J. A. Benediktsson, and J. R. Sveinsson, “Support vector machines in multisource classification,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp. Proc.*, 2003, vol. 3, pp. 2054–2056.

[24] C. Huang, L. Davis, and J. Townshend, “An assessment of support vector machines for land cover classification,” *Int. J. Remote Sens.*, vol. 23, no. 4, pp. 725–749, 2002.

[25] P. Sermanet, S. Chintala, and Y. LeCun, “Convolutional neural networks applied to house numbers digit classification,” in *Proc. 21st Int. Conf. Pattern Recognit.*, 2012, pp. 3288–3291.

[26] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[27] Y. Heryadi and E. Miranda, “Land cover classification based on sentinel-2 satellite imagery using convolutional neural network model: A case study in semarang area, Indonesia,” *Intell. Inf. Database Syst.: Recent Develop.*, vol. 11, pp. 191–206, 2020.

[28] L. Dai and C. Liu, “Multiple classifier combination for land cover classification of remote sensing image,” in *Proc. IEEE 2nd Int. Conf. Inf. Sci. Eng.*, 2010, pp. 3835–3839.

[29] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[30] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Med. Image Comput. Comput.-Assisted Interv.: 18th Int. Conf.*, Munich, Germany, 2015, pp. 234–241.

[31] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

[32] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, “Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation,” *Int. J. Comput. Vis.*, vol. 129, pp. 3051–3068, 2021.

[33] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[34] K. Pang, L. Weng, Y. Zhang, J. Liu, H. Lin, and M. Xia, “SGBNet: An ultra light-weight network for real-time semantic segmentation of land cover,” *Int. J. Remote Sens.*, vol. 43, no. 15/16, pp. 5917–5939, 2022.

[35] J. Gao, L. Weng, M. Xia, and H. Lin, “MLNet: Multichannel feature fusion lozenge network for land segmentation,” *J. Appl. Remote Sens.*, vol. 16, no. 1, pp. 016513–016513, 2022.

[36] X. Shen, L. Weng, M. Xia, and H. Lin, “Multi-scale feature aggregation network for semantic segmentation of land cover,” *Remote Sens.*, vol. 14, no. 23, 2022, Art. no. 6156.

[37] M. M. Pai, V. Mehrotra, S. Aiyar, U. Verma, and R. M. Pai, “Automatic segmentation of river and land in SAR images: A deep learning approach,” in *Proc. IEEE 2nd Int. Conf. Artif. Intell. Knowl. Eng.*, 2019, pp. 15–20.

[38] X. Li et al., “McaNet: A joint semantic segmentation framework of optical and SAR images for land use classification,” *Int. J. Appl. Earth Observ. Geoinformation*, vol. 106, 2022, Art. no. 102638.

[39] A. Vaswani et al., “Attention is all you need,” *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[40] H. Wu et al., “CVT: Introducing convolutions to vision transformers,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 22–31.

[41] W. Wang et al., “Pyramid vision transformer: A versatile backbone for dense prediction without convolutions,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 568–578.

[42] J. Chen, M. Xia, D. Wang, and H. Lin, “Double branch parallel network for segmentation of buildings and waters in remote sensing images,” *Remote Sens.*, vol. 15, no. 6, 2023, Art. no. 1536.

[43] K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep high-resolution representation learning for human pose estimation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5693–5703.

[44] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, *arXiv:1412.6980*.

[45] A. Boguszewski, D. Batorski, N. Ziemba-Jankowska, T. Dziedzic, and A. Zambrzycka, “Landcover. AI: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1102–1110.

[46] F. Rottensteiner, “ISPRS test project on urban classification and 3D building reconstruction: Evaluation of building reconstruction results,” Technical report, Tech. Rep., 2013.

[47] Z. Shao, W. Zhou, X. Deng, M. Zhang, and Q. Cheng, “Multilabel remote sensing image retrieval based on fully convolutional network,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 318–328, 2020.

[48] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SEGNet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE Trans. Pattern Anal. Mach Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[49] H. Li, P. Xiong, H. Fan, and J. Sun, “Dfanet: Deep feature aggregation for real-time semantic segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9522–9531.

[50] S. Mehta, M. Rastegari, L. Shapiro, and H. Hajishirzi, “ESPNetv2: A light-weight, power efficient, and general purpose convolutional neural network,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9190–9200.

[51] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, “Ghostnet: More features from cheap operations,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1580–1589.

[52] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, “Shufflenet v2: Practical guidelines for efficient CNN architecture design,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 116–131.

[53] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, “Learning a discriminative feature network for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1857–1866.

[54] W. Wang et al., “Efficient and accurate arbitrary-shaped text detection with pixel aggregation network,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8440–8449.

[55] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, “ENet: A deep neural network architecture for real-time semantic segmentation,” 2016, *arXiv:1606.02147*.

[56] Y. Wang et al., “Lednet: A lightweight encoder-decoder network for real-time semantic segmentation,” in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 1860–1864.

[57] F. Zhang et al., “ACFNet: Attentional class feature network for semantic segmentation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6798–6807.

[58] Y. Hong, H. Pan, W. Sun, and Y. Jia, “Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes,” 2021, *arXiv:2101.06085*.

[59] Y. Yuan, X. Chen, X. Chen, and J. Wang, “Segmentation transformer: Object-contextual representations for semantic segmentation,” 2019, *arXiv:1909.11065*.

[60] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, “ERFNet: Efficient residual factorized convnet for real-time semantic segmentation,” *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 263–272, Jan. 2018.

[61] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, “Denseaspp for semantic segmentation in street scenes,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3684–3692.

[62] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, “CCNet: Criss-cross attention for semantic segmentation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 603–612.

[63] G. Li, I. Yun, J. Kim, and J. Kim, “Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation,” 2019, *arXiv:1907.11357*.

[64] T. Wu, S. Tang, R. Zhang, J. Cao, and Y. Zhang, “CGNet: A light-weight context guided network for semantic segmentation,” *IEEE Trans. Image Process.*, vol. 30, pp. 1169–1179, 2020.

Guoying Miao received the B.S. degree in mathematics and M.S. degree in system analysis and integration from Qufu Normal University, Qufu, China, in 2006 and 2009, respectively, and the Ph.D. degree in control theory from the Nanjing University of Science and Technology, Nanjing, China, in 2013.

From 2013, she has been with the School of Automation at Nanjing University of Information Science and Technology, Nanjing, China. Her current research interests include time delay systems, cooperative control, and stochastic systems.

Huiqin Wang is currently working toward the master’s degree in electronic information with the Nanjing University of Information Science and Technology, Nanjing, China.

Her research interests include machine learning and remote sensing image analysis.

Enwei Zhang is currently working toward the master’s degree in electronic information with the Nanjing University of Information Science and Technology, Nanjing, China.

His research interests include machine learning and its application.