Multipath Multiscale Attention Network for Cloud and Cloud Shadow Segmentation

Guowei Gu, Liguo Weng, Min Xia, *Member, IEEE*, Kai Hu, and Haifeng Lin

***Abstract*— The segmentation task of cloud and cloud shadow has always been one of the important tasks in remote sensing image processing. At present, cloud detection based on deep learning methods lacks generalization, which is easy to cause the loss of space and detail information, and missed detection and false detection occur from time to time. Aiming at the above problems, this article proposes a multipath multiscale attention (MMA) network. In this network, multiscale overlapping patch (MSP) embedding is introduced to extract multiscale semantic information with multipath pyramid vision transformer (PVT), and strip convolution is used to supplement the spatial detail information of the image, so as to realize the effective aggregation of fine and rough features at the same feature level. To aggregate the global information, the multiscale global aggregation module (MGAM) is used for deep feature extraction to supplement the high semantic information. In the decoding stage, aiming at the problem of small target cloud detection, the attention-guided fusion module (AGFM) is proposed to focus on the important information of the image, remove the network noise, and increase the detection accuracy of small targets. Contextual information fusion (CIF) decoding method is proposed for the coarse segmentation boundary problem, which fully integrates the context information and effectively helps to restore the image. The experimental results on Biome 8 Dataset, HRC-WHU Dataset, and SPARCS Dataset confirm that our method is superior to the current cutting-edge cloud and cloud shadow detection technology.**

***Index Terms*— Multipath, multiscale, remote sensing image, segmentation.**

I. Introduction

SEMANTIC segmentation of remote sensing images based on deep learning is to assign each pixel in remote sensing images to different semantic categories. For example, Biome 8 Dataset is to assign each pixel to the four categories of cloud, cloud shadow, thin cloud, and background. Many factors can affect the accurate discrimination of pixels, such as the size and number of datasets, small edge information, scene information of each pixel, light, and computing resources. How to effectively distinguish the category of each pixel under different interference and prevent missed detection and false detection is a major challenge in the current remote sensing image semantic segmentation task.

Although the traditional threshold method [1], [2], [3] is suitable for cloud detection tasks in most scenes, it can be flexibly adjusted according to different scenes and needs, and it is easy to optimize. However, it is susceptible to factors such as illumination and brightness. In complex scenes, such as encountering problems such as the junction of thin clouds and the ground, cloud occlusion, etc. the traditional threshold method is difficult to achieve better results.

Since its launch, deep convolutional neural networks (DCNNs) have had a huge impact on the field of computer vision. Based on convolutional neural network (CNN) [4], [5], end-to-end training and optimization can automatically learn and extract features to achieve accurate classification. Long et al. [6] proposed a fully convolutional networks (FCNs) to achieve pixel-level segmentation of images through techniques such as fully convolutional layers, deconvolution layers, and skip layers, and achieved great breakthroughs. Ronneberger et al. [7] proposed UNet to achieve high-precision segmentation of images and achieved excellent results in processing small samples. Chen et al. [8] proposed DeepLab, using atrous spatial pyramid pooling (ASPP) to expand the receptive field, multiscale capture information, greatly enriching high semantic information. Zhao et al. [9] proposed the pyramid scene parsing network (PSPNet), which uses the pyramid pooling module (PPM) to aggregate contextual information at different scales, thereby improving the ability to obtain global information. Yu et al. [10] proposed that bilateral segmentation network (BiseNet) uses a bilateral segmentation architecture, and different branches extract different information to achieve an effective combination of detailed information and semantic information. HRvit [11] and HRNet [12] use multiscale at the same feature level to improve the robustness of the model and effectively improve the generalization ability of the model.

Cloud detection based on deep learning network has been widely used. Mohajerani and Saeedi [13], Chai et al. [14], and Jiao et al. [15] used FCN, SegNet, and UNet to segment cloud and cloud shadow, respectively, but there are more or less defects. When dealing with clouds, FCN may have the problem of incomplete segmentation. When facing the complex background of clouds and cloud shadows, such as the distribution of clouds and cloud shadows interacting with ground features and other targets, it is prone to misclassification and missed classification. SegNet is often prone to

Manuscript received 25 January 2024; accepted 14 March 2024. Date of publication 18 March 2024; date of current version 29 March 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 42075130. *(Corresponding author: Liguo Weng.)*

Guowei Gu, Liguo Weng, Min Xia, and Kai Hu are with the Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: 002311@nuist.edu.cn).

Haifeng Lin is with the College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China.

Digital Object Identifier 10.1109/TGRS.2024.3378970

missed detection and false detection due to poor extraction of global semantic information. UNet can extract multiscale information due to its unique structure, but the shape and texture of clouds and cloud shadows are easily affected by meteorological conditions. Irregularity makes UNet easy to ignore boundary information when dealing with fuzzy objects such as thin clouds.

In recent years, transformer [16] has achieved great success in the field of natural language processing (NLP). With the introduction of vision transformer (VIT) [17], Swin transformer [18], it has a huge impact on the field of computer vision. VIT is a transformer specially designed for image classification. Its excellent global perception ability has a good performance in classification tasks. Wang et al. [19] proposed pyramid VIT (PVT), which introduces a pyramid structure into the transformer. Like many CNN architectures, it uses different scale feature maps at different feature layers, which is very effective for dense prediction tasks. Wu et al. [20] proposed introducing convolutions to VITs (CVT), which is the first time to apply convolution to the self-attention mechanism. While focusing on global information, convolution is used to extract important details, which improves the applicability of the transformer model in visual tasks. Swin transformer proposes a hierarchical design and moving window method, which has higher efficiency in processing images and is more flexible in dealing with multiscale tasks.

The traditional CNN can only obtain the information in the receptive field through the filter, and can only establish the dependency relationship in the local area. The transformer model can consider the whole input sequence at the same time, so as to obtain the global context information, but the grasp of the detail information is insufficient. Under the influence of [21], [22], and [23], this article proposes a multipath multiscale attention (MMA) network. We use PVT with multiscale receptive fields and stripe convolution as the backbone of the network, and use convolution to extract local deep information. With the transformer's ability to model long-distance dependence, the advantages of both are fully utilized, and the multiscale global aggregation module (MGAM) is used to aggregate global information, aggregate global semantic information, and improve the detection ability of thin clouds and strip clouds. In the decoding stage, the attention-guided fusion module (AGFM) is used to make the model pay more attention to the important information in the encoder, improve the detection accuracy of the model for small targets, and make the positioning of clouds and cloud shadows more accurate. The contextual information fusion (CIF) decoding method makes the boundary segmentation more detailed. Our main contributions are as follows.

1) Multiscale patch embedding provides different sizes of receptive fields for transformer module, enhances the acquisition of global information, uses convolution branch to extract detailed information to supplement global information, deepens the ability of network to extract information, and builds transformer and CNN interaction platform.
2) A MGAM is proposed, which can aggregate feature maps with different receptive fields to further improve the receptive field of high semantic information. To solve the problem of coarse segmentation of cloud and cloud shadow, this article also adds the AGFM, which makes the network pay more attention to the important information of the image while removing some noise interference, strengthens the detection ability of small targets, and makes the positioning of cloud and cloud shadow more accurate.
3) The CIF decoding method is used to make the network fully integrate different levels of information, improve the utilization efficiency of context information, refine the segmentation of image edges, and improve the stability and reliability of the network.

II. Methodology

In recent years, the combination of multiple CNN and transformer networks has been effectively applied to remote sensing image processing [24], [25]. Considering the excellent performance of transformer in capturing global information, CNN performs better in processing local feature extraction. Combining the advantages of the two can better handle the segmentation of clouds and cloud shadows in complex backgrounds.

Therefore, this article proposes a multipath, multiscale architecture combining transformer and convolution, which has a good effect on the positioning of clouds and cloud shadows, thin clouds and other blurred objects, the detection ability of small targets, and image edge refinement. The network architecture is shown in Fig. 1. Fig. 2 introduces the specific structure of CNN module and transformer module in Fig. 1, respectively. PVT with pyramid structure shows excellent performance in dense prediction, such as semantic segmentation tasks. In this article, PVT is used as the backbone, multiscale overlapping patch (MSP) is used to extract features of different scales, and different spatial scales are integrated into the transformer sequence. Aiming at the problems of insufficient extraction of detailed information and missed detection and false detection of strip clouds and small targets, strip convolution is added to improve the positioning ability of the model for irregular clouds, and the information interaction platform of transformer and CNN is built. At the end of encoder, adding MGAM can effectively solve the problem of semantic gap and positioning error in the feature map, and enrich the high semantic information of the network. In the decoder part, AGFM can adaptively capture the information on the channel and spatial dimension, redistribute the weight, increase the focus of the model on the key area, and reduce the interference of noise. At the same time, CIF decoding method is used to make full use of context semantic information to compensate for the semantic gap of the picture.

A. MSP Embedding

The shape and size of clouds and cloud shadows are irregular. Different weather and light intensity will affect the segmentation effect of clouds and cloud shadows. For the accurate positioning of multiscale clouds and cloud shadows, affected by [12] and [26], this article proposes a method

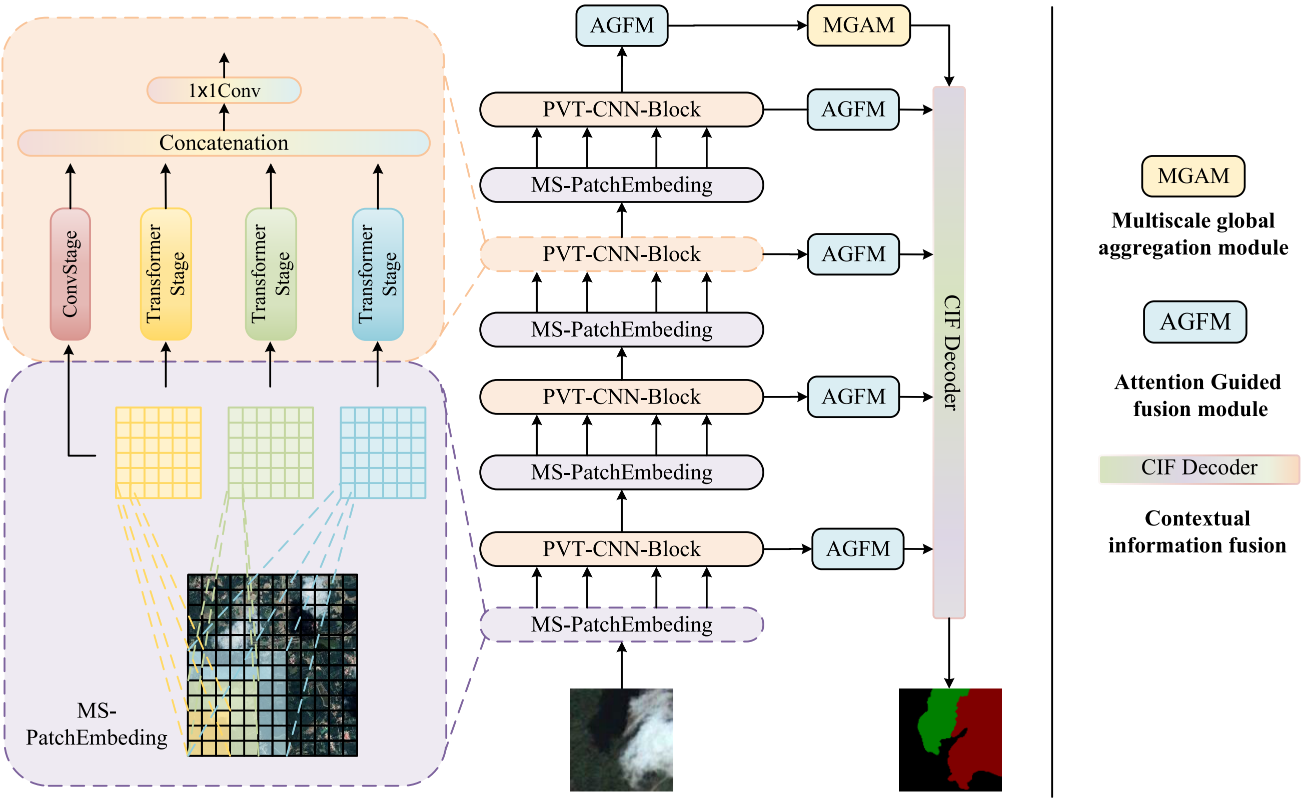


Fig. 1. MMA network structure. The model consists of a multipath PVT and strip convolution combined backbone and two auxiliary modules and CIF Decoder.

of MSP embedding. Specifically, we use PVTv2 [27] to use the method of expanding the patch window, so that the adjacent window can overlap half of the area in the process of extracting information. In addition, to ensure the same resolution, zero filling is used to supplement. The specific expression of zero-filled convolution to achieve overlapping patch embedding is as follows:

$$y_j^i = \begin{cases} Relu\left(\mathrm{BN}\left(\mathrm{Conv}_{3\times 3}\left(x_j\right)\right)\right), & i=0,\ j\neq 1 \\ Relu\left(\mathrm{BN}\left(\mathrm{Conv}_{11\times 11}\left(x_j\right)\right)\right), & i=0,\ j=1 \\ Relu\left(\mathrm{BN}\left(\mathrm{Conv}_{5\times 5}\left(x_j\right)\right)\right), & i=1 \\ Relu\left(\mathrm{BN}\left(\mathrm{Conv}_{7\times 7}\left(x_j\right)\right)\right), & i=2 \end{cases} \tag{1}$$

where, y_j^i represents the output matrix of the ith of the jth layer ($j = 1, 2, 3, 4$) of MSP, x represents the input matrix, $i = 0, 1, 2$, y_j^0, y_j^1, y_j^2 are the input of the next layer transformer, y_j^0 is the input of the next layer convolution, $\mathrm{Conv}_{s\times v}(\cdot)$ represents the convolution kernel size of $s \times v$, $\mathrm{BN}(\cdot)$ represents the batch normalization layer, and $\mathrm{Relu}(\cdot)$ represents the Relu activation function.

We change the size of each PVT-CNN-Block input feature map by changing the size of the stride. Using different scales of convolution at the same feature level can help the network better capture features with different information. The size of the convolution kernel determines the size of the receptive field. The features with different receptive fields are introduced into the transformer, and multiscale information is integrated into the sequence, which can help the model better locate clouds and cloud shadows in complex backgrounds. The structure diagram is shown in the left part of Fig. 1, and the specific parameters are shown in Table I.

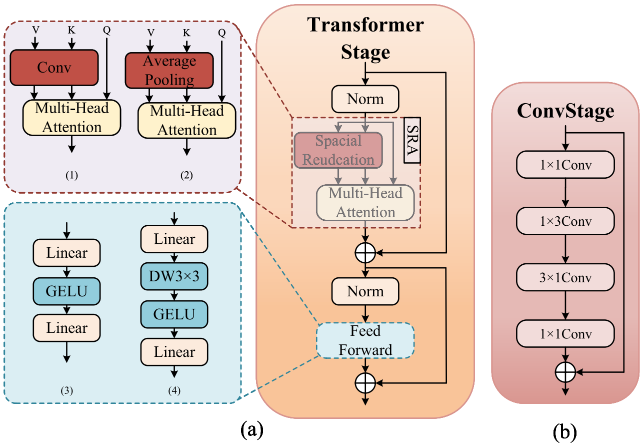


Fig. 2. Network structure of the Conv and transformer stage. (a) Composition of transformer stage. (b) Composition of convolution stage.

B. PVT-CNN-Block

It is well known that transformer has self-attention and can better capture contextual information when processing global information. PVT introduces the pyramid structure into transformer. Although it is based on the backbone of pure transformer, it has excellent performance in many visual tasks with a multiscale architecture similar to CNN. Convolution has local correlation and translation invariance. Strip

TABLE I
ARCHITECTURE OF MSP EMBEDDING

Layer	Input size	Patch Embeding1	Patch Embeding2	Patch Embeding3	Output size
L1	1, 3	$5\times 5conv$ $stride = 4$ $padding = 2$	$7\times 7conv$ $stride = 4$ $padding = 3$	$11\times 11conv$ $stride = 4$ $padding = 5$	1/4, 64
L2	1/4, 64	$3\times 3conv$ $stride = 2$ $padding = 1$	$5\times 5conv$ $stride = 2$ $padding = 2$	$7\times 7conv$ $stride = 2$ $padding = 3$	1/8, 128
L3	1/8, 128	$3\times 3conv$ $stride = 2$ $padding = 1$	$5\times 5conv$ $stride = 2$ $padding = 2$	$7\times 7conv$ $stride = 2$ $padding = 3$	1/16, 192
L4	1/16, 192	$3\times 3conv$ $stride = 2$ $padding = 1$	$5\times 5conv$ $stride = 2$ $padding = 2$	$7\times 7conv$ $stride = 2$ $padding = 3$	1/32, 256

TABLE II
ARCHITECTURE OF PVT-CNN-BLOCK

Layer	Input size	CNN Branch	Transformer Branch	dim	Output size
L1	1/4, 64	$1\times 1conv, stride = 1$ $1\times 3conv, stride = 1$ $3\times 1conv, stride = 1$ $1\times 1conv, stride = 1$	dim = 56 $head = 1$ $mlp\_ratio = 4$ $sr\_ratio = 8$	3	1/4, 64
L2	1/8, 128	$1\times 1conv, stride = 1$ $1\times 3conv, stride = 1$ $3\times 1conv, stride = 1$ $1\times 1conv, stride = 1$	dim = 128 $head = 2$ $mlp\_ratio = 4$ $sr\_ratio = 4$	4	1/8, 128
L3	1/16, 192	$1\times 1conv, stride = 1$ $1\times 3conv, stride = 1$ $3\times 1conv, stride = 1$ $1\times 1conv, stride = 1$	dim = 192 $head = 4$ $mlp\_ratio = 4$ $sr\_ratio = 2$	6	1/16, 192
L4	1/32, 256	$1\times 1conv, stride = 1$ $1\times 3conv, stride = 1$ $3\times 1conv, stride = 1$ $1\times 1conv, stride = 1$	dim = 256 $head = 8$ $mlp\_ratio = 4$ $sr\_ratio = 1$	3	1/32, 256

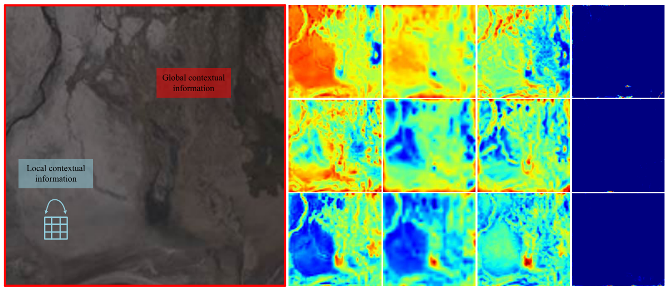
convolution is used to extract image detail information and position information in the convolution module. Through local receptive field, prior information can be effectively used to model local fine-grained information [28], which makes up for transformer's neglect of detail features. To accurately locate the location information of clouds and cloud shadows and make full use of multiscale information, we built a transformer and CNN information interaction platform, and constructed four feature extraction stages to process feature maps of different scales. The structure of transformer stage and CNN Stage is shown in Fig. 2. The specific parameters of PVT-CNN-Block are shown in Table II. After MSP downsampling, there are four PVT-CNN-Blocks for deep understanding of images.

The expression of transformer branch is as follows:

$$T_{1j}^{i} = \text{Drop}\left(\text{Norm}\left(\text{Flatten}\left(y_j^i\right) + \text{Pose\_embed}_j\right)\right) \quad (2)$$

$$T_{2j}^{i} = T_{1j}^{i} + \text{Drop}\left(\text{SRA}\left(\text{Norm}\left(T_{1j}^{i}\right)\right)\right) \quad (3)$$

$$x_{tj}^{i} = \text{Reshape}\left(T_{2j}^{i} + \text{Drop}\left(\text{MLP}\left(\text{Norm}\left(T_{2j}^{i}\right)\right)\right)\right) \quad (4)$$



(a) (b) (c) (d) (e)

Fig. 3. Thermal comparison diagram of transformer and CNN scene information acquisition ability. (a) Test image. (b) Method with ResNet. (c) Method with PSPNet. (d) Method with DeepLabV3. (e) Method with PVT.

where y_j^i represents the jth layer of transformer, the ith input matrix, x_{tj}^i represents the jth layer of transformer, the ith output, Flatten($\cdot$) represents flattening the image from 2-D data to 1-D data, Pose_embed$_j$ represents adding position coding, Norm($\cdot$) represents layer normalization, Drop($\cdot$) represents dropout regularization, SRA($\cdot$) represents space reduction attention operation, MLP($\cdot$) represents multilayer perceptron operation, and Reshape($\cdot$) represents restoring 1-D data to 2-D data. The expression of the convolution branch is as follows:

$$\text{Conv}(x) = \text{Conv}_{1\times 1}(\text{Conv}_{3\times 1}(\text{Conv}_{1\times 3}(\text{Conv}_{1\times 1}(x)))) \quad (5)$$

$$x_{cj}^{0} = Relu\left(\text{BN}\left(\text{Conv}\left(y_j^0\right)\right)\right) + y_j^0 \quad (6)$$

$$x_{j+1} = \text{Conv}_{1\times 1}\left(\text{Concat}\left(x_{cj}^0, x_{tj}^i\right)\right) \quad (7)$$

where y_j^0 represents the input matrix for the jth convolution, x_{cj}^0 represents the output matrix for the jth convolution, x_{j+1} represents the output of the jth PVT-CNN-Block, x_{cj}^0 represents the output of the jth convolution part, and x_{tj}^i represents the output of the transformer part.

The ability of transformer to extract local detail information is limited [29]. A good encoder is the key to realize semantic segmentation. The self-attention mechanism in transformer can obtain the dependence of long-distance images. In the actual segmentation scene, scene information is very important. For example, in the water scene, the color of the water area and the shadow are similar, and it is difficult to distinguish. The shadow is generally attached to the cloud. At this time, the transformer can better analyze the distribution of the shadow through the global information, and the CNN model that only focuses on the local information will cause a higher probability of mutual false detection of the water area and the shadow. In addition, the accurate segmentation of strip ridges and shadows, snow and thin clouds, and shadows and cloud shadows of mountains is also extremely dependent on scene factors. The transformer model with global receptive field has excellent scene perception ability and can better deal with such problems.

In Fig. 3, a set of thermal contrast maps of transformer and CNN scene information acquisition are visualized. From top to bottom, the attention to cloud, thin cloud, and cloud shadow is in turn. It can be seen from the graph that the CNN-based model lacks the ability to capture long-distance information and cannot accurately identify and detect objects. For example, the snow area is more concerned when detecting clouds, and the gully area is more concerned when detecting shadows. In the graph, the snow area is red when detecting clouds, and the gully area is red when detecting cloud shadows. PSPNet and DeepLabV3 with PPM and ASPP increase the receptive field with pooling and dilated convolution, respectively, which reduces the attention to some similar objects to some extent. The PVT model with the global receptive field effectively utilizes the scene information, and has a good performance in the detection of clouds, thin clouds, and cloud shadows. The effect in the figure is that the attention is very light, which is dark blue.

While obtaining these information, how to effectively extract the edge detail information of the cloud and solve the problem of missed detection and false detection of thin clouds, strip clouds, cloud, and cloud shadow boundary information has become the key to affect the segmentation effect. Thanks to the local receptive field of CNN, CNN has a good effect in dealing with complex cloud textures. Affected by [30], strip convolution is better than ordinary convolution segmentation in dealing with strip like targets, such as strip clouds, tributaries, and straight targets, and strip has a smaller amount of calculation. Therefore, this article introduces strip convolution into multiscale transformer to build an interactive platform between CNN and transformer, so that global information and local information are deeply integrated, enriching the ability of the network to extract information.

C. Multiscale Global Aggregation Module

In the actual remote sensing image, there are a large number of small target clouds and cloud shadows in the image. Considering the previous part, there is no effective extraction for small targets, and the problem of fuzzy objects, thin clouds and ground segmentation has not been effectively solved. The semantic information of the context is not fully aggregated, which will cause a large number of false detections and missed detections in the actual prediction. ESPNetv1 [31] uses the improved PPM to extract the context information of multiscale feature mapping using the average pooling of different convolution kernel sizes. Finally, the obtained context multiscale features are spliced with the original features. Although it can effectively reduce the problem of pixel-level positioning of clouds and cloud shadows, missed detection, and false detection problems, such as thin clouds and residual cloud information may be lost, which is extremely unfavorable for correct segmentation. DeepLab uses dilated convolution to increase the receptive field and fully extract high semantic information, but it will cause the model to be insensitive to small-scale targets and prone to missed and false detection. Therefore, this article proposes a MGAM. The network structure is shown Fig. 4.

First, based on the Backbone of convolution and multiscale transformer interaction, high semantic information with rich category information and location information has been extracted. It generates five sets of features, respectively and maps them to five strip convolution branches. Multiscale strip convolution deeply aggregates and refines high semantic information. Finally, we stitch these information together and fuse them with the original features, and use high-level features to guide low-level features for classification. At the same time, low-level features retain more detailed information, which can help high-level feature segmentation more refined, and also avoid gradient disappearance and gradient explosion. Using the method of multiscale strip convolution, the information of high semantic features is fused and increased, so as to improve the detection effect of the model on small targets, reduce the missed detection and false detection, and reduce the semantic gap, so as to make the positioning of clouds and cloud shadow more accurate. The formula is as follows:

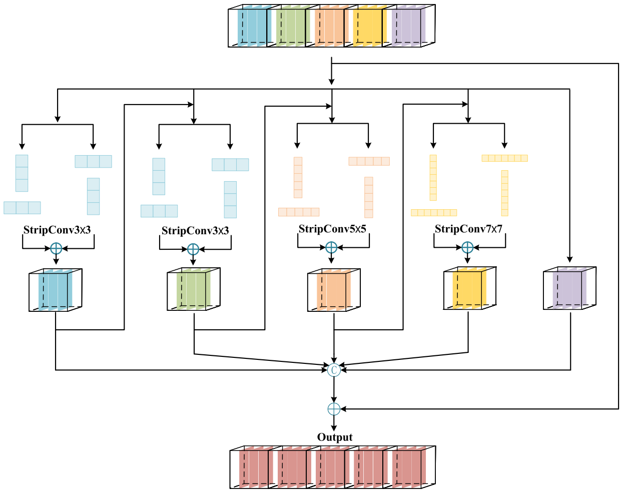


Fig. 4. Structure of MGAM.

$$d_i = x_4, \quad i = (0, 1, 2, 3, 4) \tag{8}$$

$$f_i = \begin{cases} \mathrm{Conv}_{1\times j}\big(\mathrm{Conv}_{j\times 1}(d_i)\big) + \\ \mathrm{Conv}_{j\times 1}\big(\mathrm{Conv}_{1\times j}(d_i)\big), & i = (0, 1, 2, 3) \\ d_i, & i = 4, \end{cases} \tag{9}$$

$$f_{\mathrm{out}} = \mathrm{Re}lu(\mathrm{BN}(\mathrm{Concat}(f_i) + x_4)) \tag{10}$$

where x_4 represents the output of the fourth layer PVT-CNN-Block, and Concat$(\cdot)$ represents concatenation.

D. Attention-Guided Fusion Module

The combination of CNN and transformer can effectively transform remote sensing images from multichannel color space into a deeper advanced feature space [32]. The semantic segmentation network based on encoder–decoder will inevitably appear semantic dilution phenomenon in the decoding stage. UNet simply splices the low-level semantic features with the high-level semantic features, which will result in a large amount of information redundancy, and the effective information features cannot be obtained accurately in the upsampling process. In addition, the shape of cloud and cloud shadow is not fixed, the target feature is not obvious, the texture distribution is similar, the color depth is easily affected by weather and other factors, and it is very easy to be disturbed by background noise in the actual segmentation. Therefore, based on CAM and SAM, we add the AGAM in the decoding stage. Different from the simple stacking scheme of CAM and SAM in CBAM [33], we add a branch with original features while using CAM and SAM branches. Using CAM and SAM branches, the weight of channel and space is redistributed, and the pixels of important channels are adaptively selected. To fully integrate the information of

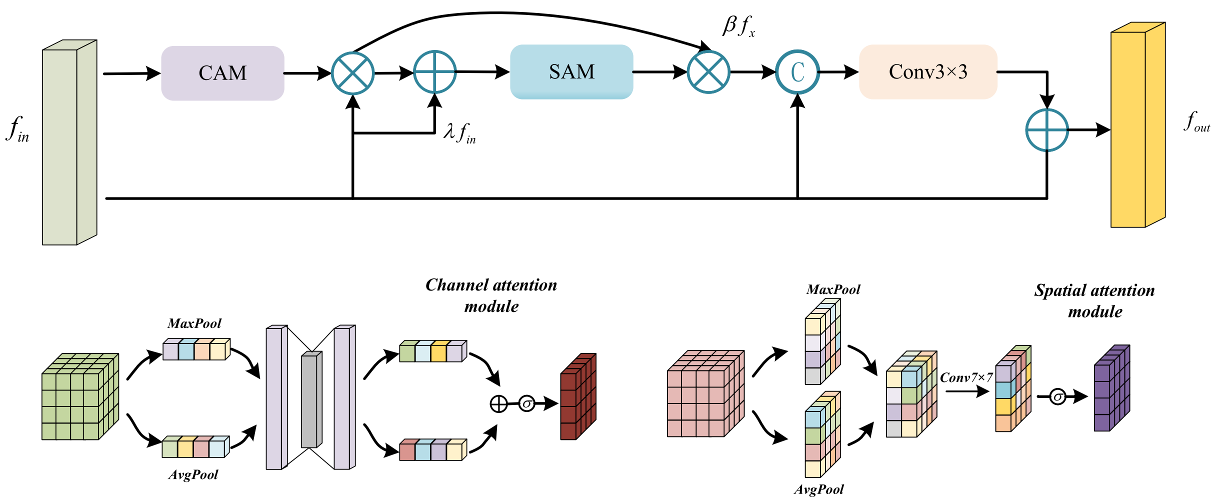


Fig. 5. Structure of AGFM.

the two branches, the skip connection is used to enhance the mutual guidance ability of the two branches. Moreover, we also introduce learnable parameters λ and β to more effectively realize the redistribution of weights, and further calibrate the importance of each pixel in the spatial dimension and channel dimension, so as to improve the discriminative ability of features and improve the performance of neural networks. The network structure is shown in Fig. 5.

This decoding based on attention feature fusion can fully integrate the features extracted by convolution and transformer, automatically mine the semantic information of each pixel, adaptively capture the transformation of information in spatial dimension and channel dimension, effectively assign attention to edge detail information, and enhance the network's attention to boundary information and small targets. The expression of the attention stage is as follows:

$$\mathrm{CAM}(x) = \sigma(\mathrm{MLP}(\mathrm{AvgPool}(x)) + \mathrm{MLP}(\mathrm{MaxPool}(x))) \tag{11}$$

$$\mathrm{SAM}(x) = \sigma(\mathrm{Conv}_{7\times 7}(\mathrm{Concat}(\mathrm{Avgpool}(x), \mathrm{Maxpool}(x)))) \tag{12}$$

$$f_x = \mathrm{CAM}(f_{\mathrm{in}}) \otimes f_{\mathrm{in}} \tag{13}$$

$$f_z = \mathrm{SAM}(f_x + \lambda f_{\mathrm{in}}) \otimes \beta f_x \tag{14}$$

$$f_{\mathrm{out}} = \mathrm{Conv}_{3\times 3}\left(\mathrm{Concat}\left(f_{z,} f_{\mathrm{in}}\right)\right) + f_{\mathrm{in}} \tag{15}$$

where CAM$(\cdot)$ and SAM$(\cdot)$ represent channel attention module and spatial attention module, respectively. AvgPool$(\cdot)$ represents average pooling operation, MaxPool$(\cdot)$ represents maximum pooling operation, MLP$(\cdot)$ represents multilayer perceptron, σ represents the activation function, λ, β represents the learned parameters, f_{in} represents the input tensor, and f_{out} represents the output tensor.

E. CIF Model

In the semantic segmentation task, the feature map needs to be sampled to the original image size. At this time, the decoder plays a key role. At present, decoder has many architectures, as shown in Fig. 6(a), which directly concatenate context information of different scales to restore the resolution of the original image. This operation may integrate a large amount of low semantic information, resulting in the loss of effective information. In Fig. 6(b), the same as UNet, the jump connection is used to connect the high semantic information step by step with the lower level feature map in Encoder, which helps to retain the low-level features and detail information of the image. However, too much integration of low semantic noise may reduce the effect of high semantic information and cause the loss of depth information. The Decoder method proposed in this article is shown in Fig. 6(c). After fully integrating the high and low semantic information, it is combined with the information in MGAM, which not only effectively retains the high semantic information, but also compensates the high semantic information with a certain resolution, so as to achieve a more effective segmentation effect. The expression of CIF is as follows:

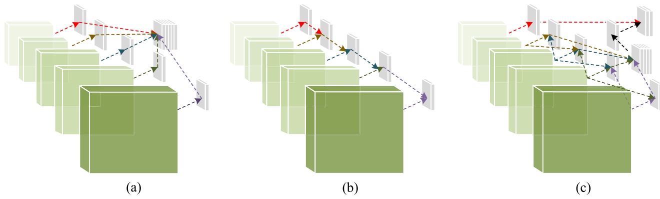


Fig. 6. Structure of CIF. Decoding method of (a) FCN. (b) UNet. (c) Our.

$$f_j = \mathrm{Upsample}(\mathrm{Conv}_{1\times 1}(f_i)), \quad i = (1, 2, 3, 4) \tag{16}$$

$$f_k^j = \begin{cases} f_j + f_{j+1}, & j = 1 \\ f_{j-1} + f_j + f_{j+1}, & j = 2, 3 \\ f_{j-1} + f_j, & j = 4 \end{cases} \tag{17}$$

$$f_{\mathrm{out}} = \mathrm{Concat}\left(\mathrm{Conv}_{1\times 1}\left(\mathrm{Concat}\left(f_k^j\right)\right), f_5\right) \tag{18}$$

where Upsample$(\cdot)$ represents bilinear interpolation upsampling.

III. Experiments

A. Datasets

1) Biome 8 Dataset: Biome 8 Dataset [34] is a dataset for us to verify the validity of the semantic segmentation model. It is a public dataset consisting of 96 Landsat8 scenes and manually generated GT, including 11 bands of information. Among

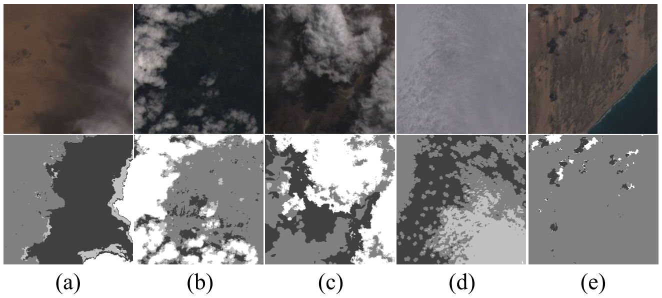


Fig. 7. Biome 8 Dataset. The first line of the picture is a visual picture after cutting, from left to right. (a) Barren. (b) Forest. (c) Shrubland. (d) Snow. (e) Water scene, and the second row is label.

them, for the study of cloud and cloud shadow, only 32 scenes with cloud shadow information were selected this time, including cloud, thin cloud, clear, shadow, and empty five categories. Limited by GPU memory, we cut the original remote sensing image dataset to a size of 224 × 224, and after screening (removing images with empty categories), we retained 25 340 images. To reduce the contingency caused by the selection of a single training set and validation set, we use a fivefold cross-validation method to divide the dataset into five subsets. The model is trained on four of the subsets, and then verified on the remaining subset. This process is repeated five times. MeanMIOU is used as the main evaluation index in the experiment (MeanMIOU is the average of the five MIOU obtained by the experiment). Using a combination of multiple different training sets and validation sets, the performance of the model can be evaluated more comprehensively. Finally, the dataset uses cloud (white), thin cloud (light gray), cloud shadow (black), and background (dark gray) semantic labels to test the performance of the model in different scenarios. Fig. 7 shows the visual sample.

2) High-Resolution Cloud Coverage Validation (HRC-WHU) Dataset: To evaluate the accuracy and robustness of the model more accurately, we use the HRC-WHU dataset to evaluate the performance index of the model. The HRC-WHU dataset was created by researchers from Wuhan University. The high-resolution cloud coverage validation dataset was created by Li et al. [35]. There are 150 high-resolution remote sensing images in the original data. The resolution of each image is 1280 × 720, including red, green, and blue channel data. To facilitate training, we cut the data into a small image of 224 × 224 for training. The dataset contains complex backgrounds such as snow, water, city, desert, and plants. To prevent overfitting, we disrupted the training dataset in the loading dataset, and performed image enhancement, random rotation, flipping, and adding noise. Fig. 8 shows a typical sample. The dataset uses cloud (white), background (black) semantic classification labels to test the performance of the model in different scenarios. This dataset uses the same fivefold cross-validation method as the Biome 8 Dataset.

3) SPARCS Dataset: The SPARCS dataset is also derived from the Landsat8 satellite, and the spatial procedures for automated removal of cloud and shadow (SPARCS) [36] is generated by processing and analyzing Landasat8 OLI sensor data. Data details are shown in Table III.

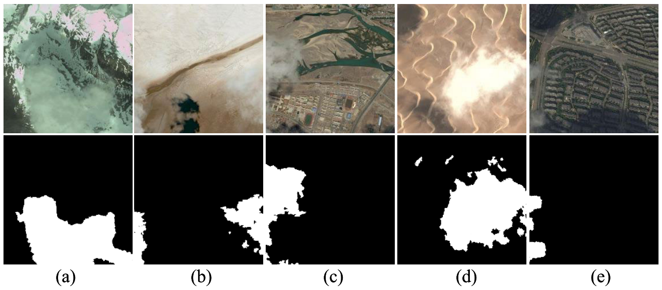


Fig. 8. HRC-WHU dataset. The first row is the cropped image, from left to right. (a) Snow. (b) River. (c) City. (d) Desert. (e) Woodland scene and the second row is label.

TABLE III
SPARCS Datasets

Satellite	band	Wavelength(nm)	Resolution
Landsat-8	1(Coastal)	430-450	30
	2(Blue)	450-515	30
	3(Green)	525-600	30
	4(Red)	630-680	30
	5(NIR)	845-885	30
	6(SWIR-1)	1560-1660	30
	7(SWIR-2)	2100-2300	30
	8(PAN)	503-676	30

Cloud validation mask dataset is a set of remote sensing data masks used to identify and distinguish clouds, shadows, and other stray objects. The dataset can be used to help researchers remove cloud and shadow cover from Landsat8 satellite data. The SPARCS dataset contains a set of high-quality, high-resolution remote sensing images, including 80 images of 1000 × 1000 pixels. On the one hand, the SPARCS dataset contains a large amount of complex scene information. Compared with 224-sized images, 256-sized images can retain more scene and detail information. On the other hand, the cloud detection network experiment compared in this article is to cut the SPARCS dataset into 256-sized images. To maintain the consistency of the experiment as much as possible, we cut the dataset into 256 × 256 images and obtained 1280 images. Due to the small number of images, we performed vertical flip, horizontal flip, and random rotation operations on the dataset. Then we divided the dataset into a training set and a verification set according to 8:2. The dataset contains multiple scenes, such as hills, woodlands, snow, waters, fields, and deserts. It is a five-category dataset, including cloud, cloud shadow, snow/ice, water, and land. Fig. 9 shows the typical sample. The dataset uses cloud (white), cloud shadow (black), water (dark blue), snow/ice (light blue), and land (gray) semantic classification labels to test the performance of the model in different scenarios. This dataset uses the same fivefold cross-validation method as the Biome 8 Dataset.

B. Experimental Details

The experiments in this article are based on the PyTorch platform, and the GPU uses RTX3070. In this article, Adam optimizer is used to optimize the experimental parameters,

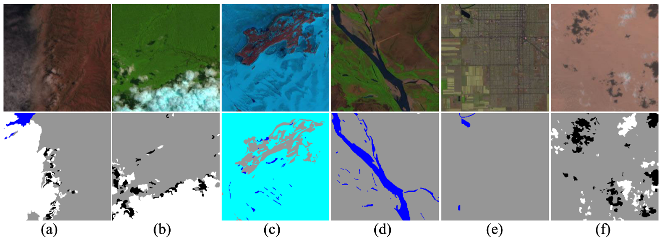


Fig. 9. SPARCS dataset. The first line selects images from (a) hills, (b) woodlands, (c) snow, (d) waters, (e) fields, (f) desert, and the second line is the label of the previous line.

where β_1 is set to 0.9, β_2 is set to 0.999, and the learning rate strategy is step learning rate schedule (StepLR). The formula is as follows, where the benchmark learning rate is set to 0.001, the adjustment multiple is 0.95×, the adjustment interval is 3, and the number of iterations is 300×. The loss function used is the cross-entropy loss function, which is limited by the memory capacity. The experiment in this article is carried out under batchsize equal to 16

$$lr_{\text{new}} = lr_{\text{initial}} \times \gamma^{\frac{\text{epoch}}{\text{stepsize}}}. \tag{19}$$

We choose precision (P), recall (R), F_1(BF) score [37], pixel accuracy (PA), mean intersection over union (MIOU), and frequency weighted intersection over union (FWIOU) to evaluate the performance of different models in cloud and cloud shadow segmentation. The formula of the above indicators is as follows:

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{20}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{21}$$

$$F_1(\text{BF}) = 2 \times \frac{P \times R}{P + R} \tag{22}$$

$$\text{PA} = \frac{\sum_{i=0}^{k} P_{i,j}}{\sum_{i=0}^{k} \sum_{j=0}^{k} P_{i,j}} \tag{23}$$

$$\text{MPA} = \frac{1}{k} \sum_{i=0}^{k} \frac{P_{i,j}}{\sum_{j=0}^{k} P_{i,j}} \tag{24}$$

$$\text{MIoU} = \frac{1}{\text{k} + 1} \sum_{i=0}^{k} \frac{P_{i,j}}{\sum_{j=0}^{k} P_{i,j} + \sum_{j=0}^{k} P_{j,i} - P_{i,i}} \tag{25}$$

$$\text{FWIOU} = \frac{1}{\sum_{i=0}^{k} \sum_{j=0}^{k} P_{i,j}} \sum_{i=0}^{k} \frac{\sum_{j=0}^{k} P_{i,j} \times P_{i,i}}{\sum_{j=0}^{k} P_{i,j} + \sum_{j=0}^{k} P_{j,i} - P_{i,i}} \tag{26}$$

where $P_{i,i}$, $P_{j,i}$, and $P_{i,j}$ represent the number of pixels in the category $i(j,\ i)$ but predicted as the category $i(i,\ j)$, respectively, TP represents the number of pixels in the correctly predicted category (cloud, cloud shadow), FN represents the number of pixels in the incorrectly predicted category (cloud, cloud shadow), and k represents the number of categories (excluding the background). In this article, we represent the F_1(BF) calculated by a single image as BF, and the average F_1(BF) of all images in the entire dataset as Avg.BF.

TABLE IV
PVT BACKBONE SELECTION COMPARISON TABLE

Method	PA(%)	MPA(%)	MeanMIOU(%)	Flops(G)	Paras(M)
PVTv1(Fig. 2(1)(3))	88.84	82.68	73.82	4.03	23.41
PVTv2(Fig. 2(2)(4))	89.01	82.78	74.14	3.95	22.36
ours(Fig. 2(1)(4))	**89.68**	**84.38**	**75.60**	4.47	28.92

TABLE V
TRANSFORMER BACKBONE SELECTION COMPARISON TABLE

Method	PA(%)	MPA(%)	MeanMIOU(%)	Flops(G)	Paras(M)
CVT	87.04	78.70	69.84	6.82	54.32
Swin	88.30	80.71	72.26	13.05	127.28
ours	**88.88**	**82.92**	**73.90**	**4.47**	**28.92**

C. Network Backbone Selection

1) PVT Backbone Selection: PVTv2 has recently made great improvements to v1. First, the downsampling method of overlapping patch embedding is introduced to expand the window captured during each downsampling, and the overlapping block embedding method is used for serialization. To improve the detection of irregular clouds and cloud shadows, we use multiscale overlapping blocks to enrich the captured information. Second, PVTv2 uses DWConv3 × 3, fully connected layer and GELU to construct a feedforward network, so that the transformer model can have the ability to extract local features similar to CNN. However, the use of average pooling instead of Conv in spatial reduction attention reduces the amount of calculation to a certain extent, but also brings the loss of calculation accuracy. Table IV compares the effects of different modules (Based on Biome 8 Dataset).

Among them, PVTv1 and PVTv2 represent the backbone, PVTv1 and PVTv2 are selected, and the corresponding combination is (1)(3), (2)(4) combination of Fig. 2. Ours represents the feedforward network constructed by adding DWConv3 × 3, full connection layer and GELU on the basis of PVTv1, and the corresponding (1)(4) combination of Fig. 2. It can be seen from the experimental results that the experimental results are the best when using ours, so we choose this scheme as the backbone.

2) Transformer Backbone Selection: To select the appropriate backbone network, we replace the transformer backbone with Swin transformer and CVT for comparative experiments. Limited by GPU memory, we use batchsize 4 for training, and other parameters are set to default values. Table V records the experiment(Based on Biome 8 Dataset). The experimental results show that the MeanMIOU and other indicators of PVT are better than the other two networks when the parameters and computational complexity are small.

D. Network Fusion Experiment

To further improve the effectiveness of model fusion, we conducted comparative experiments on the transformer and CNN fusion stage and the decoding fusion stage. The

TABLE VI
NETWORK FUSION EXPERIMENT

Method	PA(%)	MPA(%)	MeanMIOU(%)
Fig. 6(a)	89.26	83.09	74.40
Fig. 1(Pass)	89.27	83.53	74.72
Fig. 6(b)	89.37	83.79	74.92
Fig. 6(c)	**89.68**	**84.38**	**75.60**

TABLE VII
ABLATION FOR DIFFERENT MODULES IN THE MODEL

Method	Avg.BF(%)	MeanMIOU(%)
PVT	77.61	72.35
PVT+MSP	79.54	**73.81**(**1.46**↑)
PVT+MSP+CNN	79.72	**74.46**(**0.65**↑)
PVT+MSP+CNN+PPM	79.74	74.52
PVT+MSP+CNN+ASPP	79.83	74.92
PVT+MSP+CNN+MGAM	80.38	**75.19**(**0.73**↑)
PVT+MSP+CNN+MGAM+CBAM	80.43	75.34
PVT+MSP+CNN+MGAM+AGFM	80.59	**75.60**(**0.41**↑)

experimental information is recorded in Table VI (Based on Biome 8 Dataset).

Fig. 1 (Pass) represents that the original concat connection is changed to pass connection in the transformer and CNN fusion stage in Fig. 1. Fig. 6(a)–(c) represent the three decoder methods in Fig. 6, respectively. From the table, it can be seen that whether the addition fusion is used in the transformer and CNN fusion stage or the (a)(b) connection is used in the decoding stage, the MeanMIOU model in this article has the best effect.

E. Ablation Experiments on Biome 8 Dataset

To better understand the role of each module and accurately evaluate the role of each module, we conducted ablation experiments on Biome 8 Dataset. PVT with CIF Decoder is used as the baseline, by gradually eliminating or adding some modules to achieve a better understanding of the role of model performance, it helps us to study the working principles and mechanisms within the model and the role of different components. All parameters are set to default values, and the experimental results of ablation experiments are shown in Table VII. We mainly judge the quality of our model by MeanMIOU parameters.

1) Ablation for MSP: To make the model have richer contextual semantic information, MSP embedding is used to extract sequences with different receptive fields using convolution kernels of different sizes at the same feature level, and fine rough information is fused. The experimental results in Table VII show that the MeanMIOU is 1.46% higher than the simple PVT, which fully illustrates that the performance of the model can be greatly improved by fusing rough and fine features.

2) Ablation for CNN: Due to the insufficient ability of transformer to extract local information, we add strip convolution in the process of information extraction. On the one hand, we increase the model's mining of detail information, on the other hand, we strengthen the detail extraction ability of some strip and irregular clouds. The experimental results in Table VII show that MeanMIOU can be increased to 74.46% by adding strip convolution, indicating that the model of convolution and transformer interaction is very effective for cloud and cloud shadow semantic segmentation.

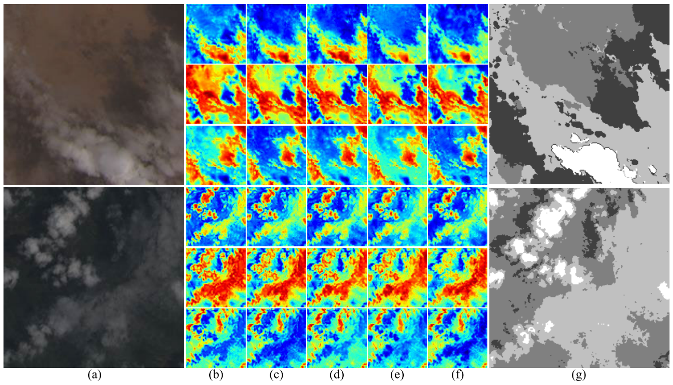


Fig. 10. Heatmaps of Cloud and Cloud shadow for different modules. (a) Test image. (b) Method with PVT. (c) Method with PVT+MSP. (d) Method with PVT+MSP+CNN. (e) Method with PVT+MSP+CNN+MGAM. (f) Method with PVT+MSP+CNN+MGAM+AGFM. (g) Label.

3) Ablation for MGAM: The shape of cloud and cloud shadow is not fixed. Thin cloud and strip cloud have always been a major difficulty in cloud and cloud shadow segmentation. To fully integrate global semantic information, increase information extraction ability, and reduce missed detection and false detection, we use multiscale strip convolution pairs to extract fusion context information. Experiments in Table VII show that adding MGAM can increase MeanMIOU by 0.73%. At the same time, compared with ASPP and PPM, which also extract multiscale information, they perform better on MeanMIOU than the above two.

4) Ablation for AGFM: To make up for the shortcomings of cloud and cloud shadow coarse segmentation and improve the positioning ability of small target objects, we add an improved CBAM attention mechanism in the decoding stage, and use channel and spatial attention to focus on important information to improve segmentation accuracy. It can be seen from Table VII that MeanMIOU increased to 75.60%, at the same time, we added a comparison with the CBAM module in the ablation experiment, and the performance of MeanMIOU was higher than that of CBAM, which fully illustrates the effectiveness of the attention mechanism on the model.

More intuitively, Fig. 10 shows the heatmaps of different periods of ablation in two complex scenes, and each scene shows the heatmaps for clouds, thin clouds, and cloud shadows from top to bottom. Fig. 10(b) shows the prediction effect of PVT. The network based on self-attention mechanism can effectively locate clouds, cloud shadows, and thin clouds, but the problem is also obvious. The edge false detection rate is high. It can be seen from the figure that the edge division is not obvious enough, and the gap between the edge division and Label is large. Avg.BF in Table VII also achieved the lowest value of 77.61%. In addition, the accurate segmentation of thin clouds has always been one of the difficult problems in cloud and cloud shadow segmentation. Due to the lack of local information extraction ability, the PVT-based model in

TABLE VIII
COMPARISON OF EVALUATION METRICS OF DIFFERENT MODELS ON BIOME 8 DATASET

	Cloud			Cloud Shadow			Thin Cloud				
Method	P(%)	R(%)	Avg.BF(%)	P(%)	R(%)	Avg.BF(%)	P(%)	R(%)	Avg.BF(%)	Param(M)	FLOPs(G)
ExtremeC3Net[38]	92.80	95.05	91.67	59.73	82.12	64.32	55.38	75.70	58.01	0.13	0.04
BiseNet[10]	91.05	95.99	91.66	63.77	79.10	64.59	69.32	69.04	59.89	2.45	3.62
CVT[20]	95.94	94.54	92.68	61.33	86.45	67.93	61.25	76.85	61.82	1.56	74.13
PSPNet[9]	92.81	94.76	91.40	70.06	73.93	63.84	68.50	72.28	61.98	33.93	48.94
CGNet[39]	94.65	93.46	91.05	65.44	77.91	64.65	67.42	76.71	64.72	0.56	0.35
PAN[40]	92.89	96.06	92.65	58.10	87.59	66.57	72.41	73.86	64.75	4.11	23.65
LinkNet[41]	93.26	96.82	93.54	64.48	83.40	67.82	73.83	69.00	61.48	1.96	11.53
CMT[42]	94.98	95.10	92.77	62.93	86.04	68.64	70.59	72.84	63.24	3.07	26.37
HRNet[12]	95.12	95.04	92.78	63.41	86.62	69.30	75.51	67.29	60.67	17.85	65.85
PVT[19]	95.33	95.32	93.15	66.60	83.12	68.78	72.52	70.82	62.44	2.12	63.24
Hrvit[11]	93.28	96.01	92.81	60.08	86.63	66.72	75.72	74.71	66.72	1.7	35.7
DeepLabv3[8]	95.12	95.92	93.62	**75.01**	74.48	66.26	63.61	81.53	66.11	34.85	54.92
DBNet[24]	96.31	96.08	94.34	59.89	**89.66**	68.96	70.19	76.73	65.98	17.84	95.29
Unet[7]	96.69	95.55	94.41	64.83	87.37	70.63	67.17	79.45	66.53	23.96	13.4
DBPNet[43]	95.71	**97.24**	95.17	67.81	82.57	69.04	68.74	**82.10**	**69.18**	17.84	95.29
MMA(Our)	**97.49**	97.11	**95.92**	69.49	83.98	**70.90**	**78.88**	75.87	68.88	4.47	28.92

the first image is not very fine on the edge, and some areas that are not thin clouds are given higher weights. To this end, we propose MSP to build a multiscale self-attention platform, and the heatmap effect is shown in Fig. 10(c). Compared with PVT of single receptive field, MSP provides multiscale receptive field for PVT. On the one hand, the positioning accuracy of cloud, cloud shadow, and thin cloud is obviously improved. It can be seen from the figure that the area given by high weight is obviously reduced, the prediction effect is closer to label, and the boundary segmentation is obviously improved. According to the Avg.BF in Table VII, Avg.BF has the highest increase of 1.93%. The addition of MSP makes the effective combination of fine and rough features at the same feature level, and the prediction weight distribution is more reasonable, which is reflected in the figure that the nontarget area is bluer and the target area is redder. In Fig. 10(d), the strip CNN module is added, and the model boundary segmentation ability is further improved. The Avg.BF also rises to 79.72%, which proves the CNN's ability to obtain local information. MGAM is added to Fig. 10(e). It can be clearly seen from the figure that the detection ability of small targets has been improved, and the detection ability of edges has been further improved. Avg.BF has also increased by 0.66%. At the same time, MeanMIOU has been greatly improved after the addition of CNN module and MGAM module, but with the improvement of local information acquisition ability, noise interference has increased. To this end, this article proposes the AGFM module, which uses space and channel attention to reduce noise interference, increase weight for the target area, and reduce weight for background noise. It is reflected in the heatmap that the red area in Fig. 10(f) is more focused, the blue area is assigned less weight, and the yellow segmentation line of the boundary becomes clearer.

F. Comparative Experiments on Different Datasets

1) Comparison Test on Biome 8 Dataset: To fully understand the actual performance of our model, we designed comparative experiments on Biome 8 Dataset. The experiments compared the current cutting-edge semantic segmentation techniques. To ensure the objectivity of the experiment, we set the parameters of the experiment as default values. In this experiment, Flops and Paras are calculated on an image with a size of 224 × 224 × 3. Tables VIII and IX records the detailed data of this experiment.

TABLE IX
COMPARISON OF EVALUATION METRICS OF DIFFERENT MODELS ON BIOME 8 DATASET

Method	PA(%)	MPA(%)	Fwiou(%)	MeanMIOU(%)
ExtremeC3Net[38]	86.04	75.96	77.81	67.54
BiseNetV2[10]	86.60	79.32	77.62	69.53
CVT[20]	87.41	78.41	79.57	70.37
PSPNet[9]	87.01	80.93	78.16	70.47
CGNet[39]	87.72	80.35	79.54	71.02
PAN[40]	88.02	79.61	80.00	71.32
LinkNet[41]	87.32	80.91	78.53	71.45
CMT[42]	87.85	80.49	79.65	71.78
HRNet[12]	87.65	81.39	79.07	71.85
PVT[19]	87.76	81.61	79.30	72.20
Hrvit[11]	88.47	80.89	80.51	72.45
DeepLabv3[8]	88.26	82.00	80.37	72.64
DBNet[24]	88.64	80.34	81.12	72.70
Unet[7]	88.87	81.00	81.48	73.38
DBPNet[43]	89.31	81.99	82.02	74.31
MMA(Our)	**89.68**	**84.38**	**82.12**	**75.60**

In the experiment, we compare our model with the current popular CNN and transformer models, where some networks have a network structure and information extraction method similar to our network. PVT is a pure transformer model based on self-attention mechanism, which enables the model to automatically extract and encodes key information in the input sequence, but lacks the extraction of detailed information. PVT uses a multiscale structure similar to CNN to make it more prominent in intensive prediction tasks. CVT, CMT, and other networks are used to integrate convolution into transformer, which improves the extraction of deep information by the network. ExtremeC3Net, BiseNetV2, DBNet, and DBPNet use a dual-branch network as an encoder to extract space and detail information, but they are composed in different ways. The first

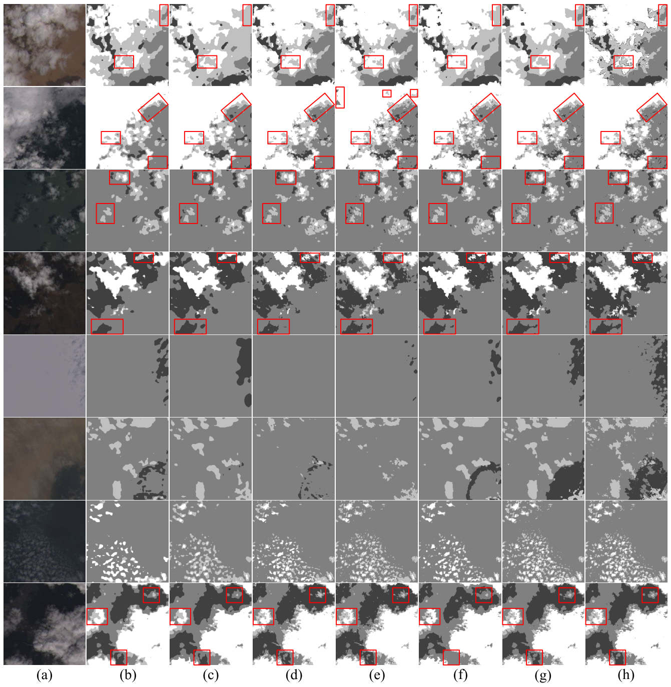


Fig. 11. Comparison of different methods under different scenarios in Biome 8 Dataset. (a) Test images. (b) Segmentation results of HRvit. (c) Segmentation results of DeepLabV3. (d) Segmentation results of DBNet. (e) Segmentation results of UNet. (f) Segmentation results of DBPNet. (g) Segmentation results of our method. (h) Label.

two are pure CNN based models, and the latter two are based on transformer and CNN. LinkNet innovates in the feature fusion upsampling part. CGNet fuses contextual semantic information. UNet proposes unique Encoder and Decoder structures. DeepLabV3 and PSPNet add multiscale modules at the end of encoder for feature extraction. HRNet and HRvit use multiscale at the same feature level. Fine features are combined with rough features. The comparison network has certain similarity with the network in this article. Table VIII shows the P (%), R (%), and Avg.BF (%) of each class. It can be seen from the table that in terms of cloud detection, the values of P (%) and Avg.BF (%) of our model have achieved the best results, and the value of R (%) is slightly lower than that of DBPNet. In terms of cloud shadow detection, our model's P (%), R (%), and comprehensive index Avg.BF (%) achieved the best results. In terms of thin cloud prediction, the P (%) of our model has achieved the best results, and Avg.BF (%) is slightly lower than DBPNet. Table IX shows the comparison of the comprehensive indicators of the model. Our model has achieved the best results in PA (%), MPA (%), Fwiou (%), and MeanMIOU (%). To more intuitively show the prediction effect of the model, this article visualizes several sets of prediction comparison charts.

In Fig. 11, six methods with better performance on Biome 8 Dataset were selected for prediction and comparison. The first three rows of scenes are barren, forest, and grass, respectively. It can be seen from the graph that our model has a great advantage in dealing with the boundary segmentation of cloud and thin cloud, thin cloud, and ground. In Table VIII, our model has achieved 95.92% and 68.88% of Avg.BF (%) value in cloud and thin cloud detection, respectively, which verifies the excellent performance of our model in repairing cloud boundaries. The fourth row is the Shrubland scene. The lower frame in the scene shows the excellent positioning ability of the model for shadows. In addition, the upper frame in the Shrubland scene reflects the detailed processing of the boundary of the model for clouds, cloud shadows, and thin clouds. In Table VIII, our model achieved 70.90% of the Avg.BF (%) value in cloud shadow detection, ranking first. On the other hand, it shows the cloud shadow boundary repair ability of the model in this article. In the fifth and sixth rows of snow and urban scenes, due to the high similarity between snow and thin clouds and clouds, floods and cloud shadows, the probability of missed detection and false detection is increased. The model in this article is based on transformer and has excellent scene perception ability, so it has a good performance. In addition, MSP provides a richer global vision for the transformer model, which is conducive to strengthening the model's ability to locate the target. In the seventh row water scene, it can be seen from the graph that our model has excellent detection ability for small target clouds and thin clouds. This is due to the fact that MMA builds an interactive platform between CNN and transformer, which enhances the ability to extract local information. MGAM also enhances the small target detection ability of MMA to a certain extent. In Table VIII, the P (%) value of MMA in cloud detection is as high as 97.49%, and the P (%) value in thin cloud detection is as high as 78.88%. It also shows the powerful cloud and thin cloud detection ability of the model in this article. The last row of the Wetlands scene, the background is more complex, our model is more excellent to complete the interaction of thin clouds, shadows, clouds, cloud shadow positioning problem, and edge segmentation problem. In a word, compared with other methods, our method has obvious effects on boundary repair of cloud and cloud shadow, small target detection, thin cloud detection, target location, and so on.

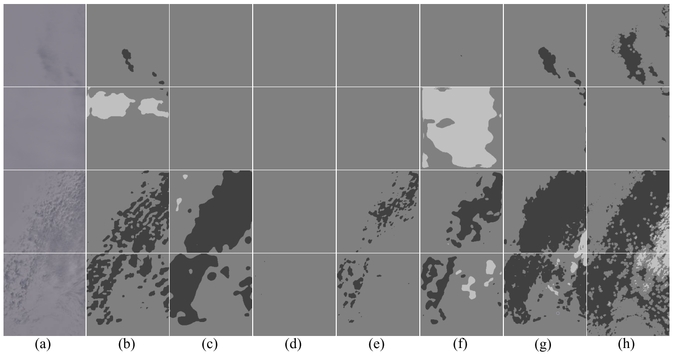


Fig. 12. Comparison of false detection of different models in Biome 8 Dataset. (a) Test images. (b) Segmentation results of HRvit. (c) Segmentation results of DeepLabV3. (d) Segmentation results of DBNet. (e) Segmentation results of UNet. (f) Segmentation results of DBPNet. (g) Segmentation results of our method. (h) Label.

In Fig. 12, we selected the snow scene with serious false detection and missed detection to compare the prediction effect. The color of the snow is very similar to that of clouds and thin clouds, and it is difficult to distinguish with the naked eye. Dealing with such problems has a huge test on the model.

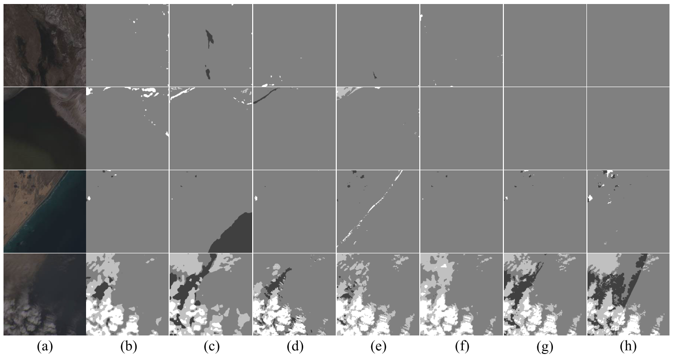


Fig. 13. Anti-noise comparison of different models in Biome 8 Dataset. (a) Test images. (b) Segmentation results of HRvit. (c) Segmentation results of DeepLabV3. (d) Segmentation results of DBNet. (e) Segmentation results of UNet. (f) Segmentation results of DBPNet. (g) Segmentation results of our method. (h) Label.

In the first sub-graph, only our model and HRvit detected the distribution of some cloud shadows, while other models did not detect cloud shadows. In the second sub-graph, HRvit and DBPNet mistakenly detected part of the snow area as a thin cloud. In the third subgraph, only our model in this article detects the distribution of some thin clouds. In the last image, only our model in this article roughly detected the distribution of cloud shadow and thin cloud. In the final analysis, so many missed and false detections are caused by insufficient global information extraction. Although the CVT branch is used in DBNet, the large number of CNN structures in DBNet may weaken the model's ability to extract global information. Therefore, the model performs poorly in this scenario and does not detect any cloud shadow and thin cloud information. HRvit and DBPNet have transformer module, which can detect the distribution of shadows more, but there are more false detections for thin clouds. DeepLabV3 enriches the semantic information of the model by virtue of its ASPP module, which has a good effect on shadow detection. In Table VIII, the *P* (%) of DeepLabV3 in cloud shadow detection reached 75.01%, which was the best among all models. Unet uses its excellent detail information extraction ability to detect and process the distribution of some cloud shadows. Its Avg.BF (%) in cloud shadow detection is 70.63%, which is excellent and ranks second among all models. In general, our model effectively utilizes the advantages of the multiscale transformer model. The fusion of fine features at the same feature level can effectively enrich the ability to obtain global information, so it performs well in snow scenes.

In Fig. 13, the anti-interference performance of the model is compared. The white snow and black waters are very similar to clouds and cloud shadows, respectively. The problem of noise interference in cloud and cloud shadow segmentation has always been one of the more important issues. In these scenes, some of them cannot be accurately judged by the naked eye. The actual segmentation diagram of each model in the figure shows that the white snow and black water do cause some interference to the accuracy of the model. Some models have a large area of missed detection and false detection. Our model has AGFM module and self-attention mechanism, which can effectively remove the interference through the scene where the cloud and cloud shadow are located. In the actual segmentation, it also reflects the superiority of our model over other models.

TABLE X

Evaluation Results of Different Models on the HRC-WHU Dataset

Method	PA(%)	MPA(%)	R(%)	Avg.BF(%)	MeanMIoU(%)
ExtremeC3Net	96.65	94.83	95.60	93.31	90.93
BiseNetV2	96.84	95.23	95.78	93.57	91.52
DeepLabV3	97.30	95.99	94.81	94.47	92.78
CMT	97.55	96.25	97.78	95.00	93.31
CGNet	97.56	96.47	96.62	94.96	93.26
PAN	97.62	96.25	96.96	95.18	93.89
Unet	97.74	96.31	97.24	95.47	93.77
PVT	97.76	96.70	96.93	95.38	93.90
HRNet	97.82	96.64	97.15	95.55	94.02
LinkNet	97.84	96.51	97.34	95.66	94.03
PSPNet	97.86	96.85	97.09	95.59	94.18
CVT	97.91	96.96	97.12	95.68	94.29
HRvit	97.96	97.06	97.16	95.77	94.45
DBNet	98.01	96.82	97.51	95.98	94.47
DBPNet	98.07	97.08	97.44	96.05	94.70
MMA(Our)	**98.64**	**98.02**	**98.14**	**97.18**	**96.23**

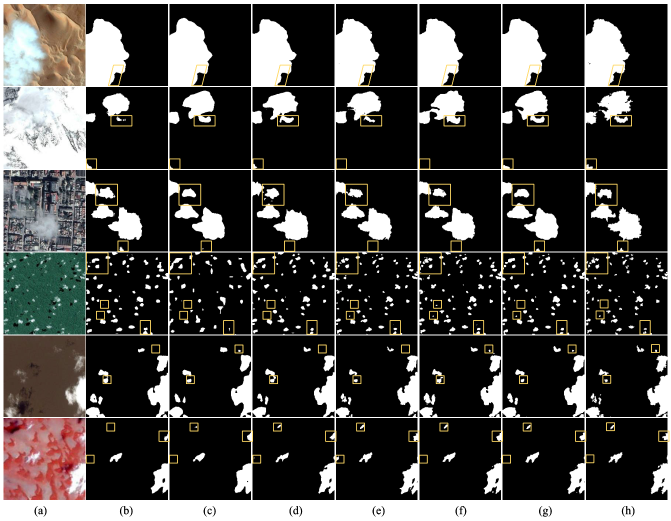


Fig. 14. Comparison of different methods under different scenarios in HRC-WHU Dataset. (a) Test images. (b) Segmentation results of PSPNet. (c) Segmentation results of CVT. (d) Segmentation results of HRvit. (e) Segmentation results of DBNet. (f) Segmentation results of DBPNet. (g) Segmentation results of our method. (h) Label.

2) Generalization Experiment of HRC-WHU Dataset: To further verify the validity of the model in this article, we conducted a generalization comparison test on the HRC-WHU dataset. The parameters in the experiment were set as default parameters, and both Flops and Paras were consistent with the data in the Biome 8 Dataset. The experimental results are shown in Table X. Since it is a two-class classification, most of the models perform well. Among them, most of the models combined with transformer model and transformer and CNN show excellent performance on this dataset. Compared with other networks, our model has the highest values in PA (%), MPA (%), *R* (%), Avg.BF (%), and MeanMIOU (%), which verifies the generalization ability of MMA (Our).

In Fig. 14, we compare the prediction maps of our network with the other five highest MeanMIOU models. The images of desert, woodland, city, snow, and other scenes are selected,

TABLE XI
COMPARISON OF EVALUATION METRICS OF DIFFERENT MODELS ON SPARCS DATASET

	Class Pixel Accuracy					Overall Results					
Method	Cloud(%)	Cloud Shadow(%)	Snow/Ice(%)	Water(%)	Land(%)	PA(%)	MPA(%)	R(%)	Avg.BF(%)	FWIOU(%)	MeanMIOU(%)
CMT	86.80	70.28	93.48	85.71	95.61	91.20	86.37	88.64	83.20	84.45	78.54
BiseNetV2	91.11	73.36	96.40	84.68	96.87	93.16	88.48	91.51	86.48	87.71	82.26
ExtremeC3Net	91.63	76.98	96.72	89.15	96.63	93.70	90.22	91.49	87.37	88.53	83.73
CGNet	91.79	79.59	97.00	86.07	96.95	93.95	90.28	92.04	87.81	88.94	84.09
PVT	91.11	79.09	94.57	90.10	96.80	93.77	90.33	91.83	87.69	88.60	84.04
LinkNet	91.53	79.30	96.81	88.42	97.31	94.07	90.28	92.47	88.21	89.20	84.54
CVT	91.57	78.59	96.78	90.39	96.93	94.07	90.85	92.05	88.13	89.14	84.78
PSPNet	92.49	79.69	96.62	88.42	97.02	94.26	90.85	92.41	88.44	89.45	84.94
PAN	92.27	79.02	96.98	89.08	97.14	94.30	90.90	92.45	88.50	89.55	85.02
HRNet	93.03	76.93	97.45	89.11	97.25	94.38	90.75	92.87	88.75	89.74	85.18
Hrvit	91.98	79.10	96.86	89.81	97.29	94.31	91.01	92.64	88.75	89.67	85.22
DeepLabV3	92.55	78.79	**97.70**	**91.97**	97.14	94.56	91.63	92.55	88.90	90.00	85.68
Unet	92.39	78.11	97.13	90.66	97.60	94.63	91.18	93.24	89.32	90.17	85.87
DBNet	92.11	80.13	96.92	91.69	97.60	94.79	91.69	93.23	89.55	90.40	86.31
DBPNet	91.73	81.22	96.71	91.54	97.60	94.77	91.76	93.18	89.57	90.35	86.36
MMA(Our)	**94.47**	**84.42**	97.48	89.83	**97.63**	**95.56**	**92.80**	**93.91**	**90.70**	**91.80**	**87.76**

respectively. In the first line of desert background, our model is more precise in boundary segmentation, the highest value of Avg.BF (%) in Table X is 97.18%, which also shows the boundary segmentation ability of the model in this article. In the background of the second row of snow and cloud, the interference of snow caused some interference to the correct segmentation. Our model is superior to other models in the boundary segmentation of snow and cloud and the detection of small target cloud, In Table X, MPA (%) achieved the highest value of 98.02%, which also shows the excellent positioning ability of the model for cloud and small targets. In the third row of urban background, the accurate positioning of thin clouds has always been a major difficulty in cloud and cloud shadow segmentation. The first few models are significantly weaker than our model in the positioning of thin cloud boundaries and the detection of small target thin clouds. In the background of the fourth row of woodland, a large number of clouds and cloud shadows of small targets are collected, which is very easy to miss detection and false detection. From the experimental results, our model has certain advantages in the recognition ability of small targets. In the fifth line of desert background, there are some residual clouds that are difficult to segment, and our model can effectively locate them. In the last line of color background, our model has better recognition ability for small target positioning than other networks. In summary, our model performs better than other models on the dataset of HRC-WHU in terms of small target recognition and positioning, thin cloud detection, anti-noise, and boundary refinement.

3) Generalization Experiment of SPARCS Dataset: To further verify the generalization ability of the model, we selected the SPARCS five classification dataset to verify the validity of the model after selecting the three classification and two classification datasets. The five classification dataset further tests the anti-interference ability of the model. The experimental results are shown in Table XI. In terms of MeanMIOU, our model has achieved an excellent result of 87.76%. In other comprehensive indicators PA (%), MPA (%), R (%), Avg.BF (%), and FWIOU (%) also achieved the highest value. Class PA refers to the P (%) of each class, which shows the accuracy of the model for different categories. The results in the table show that our model has achieved the highest P (%) in cloud, cloud shadow, and land detection, but it is slightly lower than DeepLabV3 in snow/ice and water detection accuracy. In the detection of clouds and cloud shadows, our prediction accuracy is as high as 94.47% and 84.42%. The prediction level of our model is far ahead of other models, which shows the advantages of our model for cloud and cloud shadow detection.

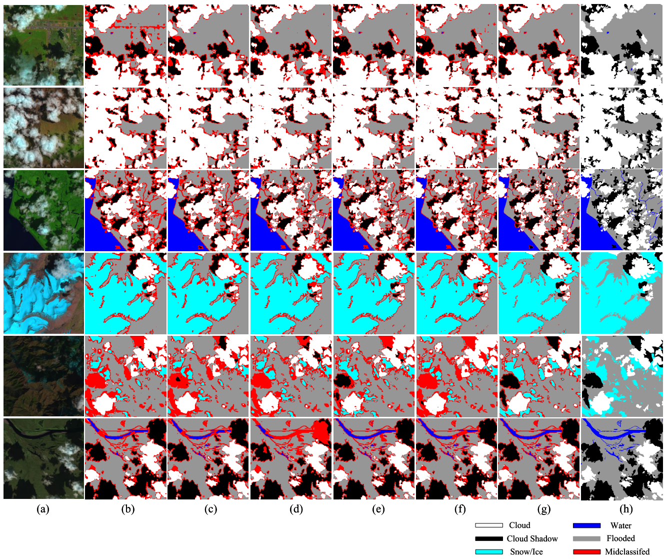


Fig. 15. Comparison of different methods under different scenarios in SPARCS dataset. (a) Test images. (b) Segmentation results of HRvit. (c) Segmentation results of DeepLabV3. (d) Segmentation results of Unet. (e) Segmentation results of DBNet. (f) Segmentation results of DBPNet. (g) Segmentation results of our method. (h) Label.

In Fig. 15, we show the segmentation map comparison map of our model with the other five models with higher MeanMiou values. The first and second rows mainly show the segmentation effect of clouds, cloud shadows, and land, the comprehensive performance index Avg.BF (%) in Table XI is 90.70%, which also shows the excellent ability of the model in boundary repair. Our segmentation edges are fine and less affected by noise such as roads. The third and sixth lines show the segmentation effect of water area and cloud shadow. In dealing with the similar segmentation of river and cloud shadow, there is no large-scale missed detection and false detection, and the segmentation edge is fine. The fourth and fifth lines show the segmentation effect of large target snow and small target snow, cloud, and cloud shadow. In the complex scene with hilly shadow, it is very easy to misdetect into cloud shadow. Our model does not have a wide range of false detection, and the segmentation effect is better. In summary, the red area is the area of missed detection and false detection. Our model has obvious advantages over other networks in the positioning of clouds and cloud shadows. There are fewer missed detection and false detection problems, and the boundary segmentation is clearer. This is determined by the characteristics of our network. The network combined with transformer and CNN can better combine global information and detailed information. MGAM and AGFM enhance the positioning accuracy of the network for small targets. The P (%) values of cloud, cloud shadow and land are 94.47%, 84.42%, and 97.63%, respectively, and the detection accuracy is greatly ahead of other networks. When detecting small target objects, the figure shows obvious advantages. The unique CIF Decoder fully integrates contextual semantic information. Enhance the decoding ability of the network while ensuring that the information is not seriously lost.

IV. Conclusion

This article proposes a multipath transformer and CNN combined network to achieve cloud and cloud shadow semantic segmentation of high-resolution remote sensing images of visible spectrum. This method uses multiscale transformer network to integrate multiscale global information at the same feature level, and uses strip convolution to extract detailed information to supplement the information extracted by transformer module. At the same time, MGAM with multiscale receptive field aggregates global information for the model and increases the detection ability of the model for small targets. In the decoding stage, the AGFM is used to guide the decoding and focus on the key information, and then the context fusion decoding method is performed through the CIF module to ensure the segmentation accuracy and repair the edge information. The experimental results show that the MMA model has a good performance compared with other networks on the Biome 8 Dataset, HRC-WHU Dataset, and SPARCS Dataset.

References

[1] Z. Zhu and C. E. Woodcock, “Object-based cloud and cloud shadow detection in Landsat imagery,” *Remote Sens. Environ.*, vol. 118, pp. 83–94, Mar. 2012.

[2] D. G. Manolakis, G. A. Shaw, and N. Keshava, “Comparative analysis of hyperspectral adaptive matched filter detectors,” in *Algorithms for Multispectral, Hyperspectral, and Ultraspectral Imagery*, vol. 4049. Bellingham, WA, USA: SPIE, 2000, pp. 2–17.

[3] H. Zhai, H. Zhang, L. Zhang, and P. Li, “Cloud/shadow detection based on spectral indices for multi/hyperspectral optical remote sensing imagery,” *ISPRS J. Photogramm. Remote Sens.*, vol. 144, pp. 235–253, Oct. 2018.

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

[5] X. Dai, M. Xia, L. Weng, K. Hu, H. Lin, and M. Qian, “Multi-scale location attention network for building and water segmentation of remote sensing image,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3276703.

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

[7] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* (Lecture Notes in Computer Science), vol. 9351, 2015, pp. 234–241.

[8] 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. (ECCV)*, 2018, pp. 801–818.

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

[10] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, “BiSeNet: Bilateral segmentation network for real-time semantic segmentation,” in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 325–341.

[11] J. Gu et al., “Multi-scale high-resolution vision transformer for semantic segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 12094–12103.

[12] C. Yu et al., “Lite-HRNet: A lightweight high-resolution network,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10440–10450.

[13] S. Mohajerani and P. Saeedi, “Cloud-Net: An end-to-end cloud detection algorithm for Landsat 8 imagery,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 1029–1032.

[14] D. Chai, S. Newsam, H. K. Zhang, Y. Qiu, and J. Huang, “Cloud and cloud shadow detection in Landsat imagery based on deep convolutional neural networks,” *Remote Sens. Environ.*, vol. 225, pp. 307–316, May 2019.

[15] L. Jiao, L. Huo, C. Hu, and P. Tang, “Refined UNet: UNet-based refinement network for cloud and shadow precise segmentation,” *Remote Sens.*, vol. 12, no. 12, p. 2001, Jun. 2020.

[16] N. Kitaev, Ł. Kaiser, and A. Levskaya, “Reformer: The efficient transformer,” 2001, *arXiv:2001.04451*.

[17] A. Dosovitskiy et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” 2010, *arXiv:2010.11929*.

[18] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

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

[20] H. Wu et al., “CvT: Introducing convolutions to vision transformers,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 22–31.

[21] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, “Transformer in transformer,” in *Proc. NIPS*, 2021, pp. 15908–15919.

[22] S. Zheng et al., “Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6881–6890.

[23] C. Yang, Z. Huang, and N. Wang, “QueryDet: Cascaded sparse query for accelerating high-resolution small object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13668–13677.

[24] C. Lu, M. Xia, M. Qian, and B. Chen, “Dual-branch network for cloud and cloud shadow segmentation,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3175613.

[25] 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 Obser. Remote Sens.*, vol. 16, pp. 32–43, 2023.

[26] T. Y. Lin, P. Dollàr, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2117–2125.

[27] W. Wang et al., “PVT V2: Improved baselines with pyramid vision transformer,” *Comput. Vis. Media*, vol. 8, no. 3, pp. 415–424, Sep. 2022.

[28] M. Xia, X. Zhang, W. Liu, L. Weng, and Y. Xu, "Multi-stage feature constraints learning for age estimation," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2417–2428, 2020, doi: 10.1109/TIFS.2020.2969552.

[29] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 1–15, 2017.

[30] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, "Segnext: Rethinking convolutional attention design for semantic segmentation," 2022, *arXiv:2209.08575*.

[31] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, "ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 552–568.

[32] L. Song, M. Xia, J. Jin, M. Qian, and Y. Zhang, "SUACDNet: Attentional change detection network based on Siamese U-shaped structure," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 105, Dec. 2021, Art. no. 102597.

[33] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.

[34] E. Vermote, C. Justice, M. Claverie, and B. Franch, "Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product," *Remote Sens. Environ.*, vol. 185, pp. 46–56, Nov. 2016.

[35] Z. Li, H. Shen, Q. Cheng, Y. Liu, S. You, and Z. He, "Deep learning based cloud detection for medium and high resolution remote sensing images of different sensors," *ISPRS J. Photogramm. Remote Sens.*, vol. 150, pp. 197–212, Apr. 2019.

[36] M. Hughes and D. Hayes, "Automated detection of cloud and cloud shadow in single-date Landsat imagery using neural networks and spatial post-processing," *Remote Sens.*, vol. 6, no. 6, pp. 4907–4926, May 2014.

[37] G. Csurka, D. Larlus, F. Perronnin, and F. Meylan, "What is a good evaluation measure for semantic segmentation?" *Bmvc*, vol. 27, no. 2013, pp. 5210–5244, 2013.

[38] H. Park, L. L. Sjösund, Y. Yoo, J. Bang, and N. Kwak, "Extremec3Net: Extreme lightweight portrait segmentation networks using advanced C3-modules," 2019, *arXiv:1908.03093*.

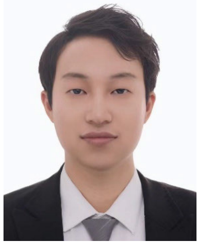
[39] T. Wu, S. Tang, R. Zhang, J. Cao, and Y. Zhang, "CGNet: A light-weight context guided network for semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 1169–1179, 2021.

[40] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," 2018, *arXiv:1805.10180*.

[41] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," *IEEE Vis. Commun. Image Process. (VCIP)*, 2017, pp. 1–4, doi: 10.1109/VCIP.2017.8305148.

[42] J. Guo et al., "CMT: Convolutional neural networks meet vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12175–12185.

[43] 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, p. 1536, Mar. 2023.



Guowei Gu received the B.S. degree from Huaiyin Institute of Technology, Huaian, China, in 2022.

He was a Graduate Student of electronic information at Nanjing University of Information Science and Technology, Nanjing, China. His research interests include deep learning and its applications.

Liguo Weng received the Ph.D. degree from North Carolina A&T State University, Greensboro, NC, USA, in 2010.

He is currently an Associate Professor with the College of Automation, Nanjing University of Information Science and Technology, Nanjing, China. His main research interests include deep learning and its application in remote sensing image analysis.

Min Xia (Member, IEEE) received the Ph.D. degree in cybernetics control engineering from Donghua University, Shanghai, China, in 2009.

He is currently a Professor with Nanjing University of Information Science and Technology, Nanjing, China, and the Deputy Director of Jiangsu Key Laboratory of Big Data Analysis Technology. His principal research interests include machine learning theory and its application.

Kai Hu received the bachelor's degree from China University of Metrology, Hangzhou, China, in 2003, the master's degree from Nanjing University of Information Science and Technology, Nanjing, China, in 2008, and the Ph.D. degree in instrument science and engineering from Southeast University, Nanjing, in 2015.

He is currently an Associate Professor with Nanjing University of Information Science and Technology. His main research focus on deep learning and its applications in remote sensing images.

Haifeng Lin is currently a Professor with the College of Information Science and Technology, Nanjing Forestry University, Nanjing, China. His main research interests include networking, wireless communication, deep learning, pattern recognition, and the Internet of Things.