Dual-Branch Network for Cloud and Cloud Shadow Segmentation

Chen Lu, Min Xia, *Member, IEEE*, Ming Qian, and Binyu Chen

***Abstract*— Cloud and cloud shadow segmentation is one of the most important issues in remote sensing image processing. Most of the remote sensing images are very complicated. In this work, a dual-branch model composed of transformer and convolution network is proposed to extract semantic and spatial detail information of the image, respectively, to solve the problems of false detection and missed detection. To improve the model's feature extraction, a mutual guidance module (MGM) is introduced, so that the transformer branch and the convolution branch can guide each other for feature mining. Finally, in view of the problem of rough segmentation boundary, this work uses different features extracted by the transformer branch and the convolution branch for decoding and repairs the rough segmentation boundary in the decoding part to make the segmentation boundary clearer. Experimental results on the Landsat-8, Sentinel-2 data, the public dataset high-resolution cloud cover validation dataset created by researchers at Wuhan University (HRC_WHU), and the public dataset Spatial Procedures for Automated Removal of Cloud and Shadow (SPARCS) demonstrate the effectiveness of our method and its superiority to the existing state-of-the-art cloud and cloud shadow segmentation approaches.**

***Index Terms*— Deep learning, dual branch, remote sensing image, segmentation.**

I. Introduction

CLOUD and cloud shadow detection is a crucial issue in remote sensing image processing. On the one hand, cloud is an important meteorological element, and climate change can be analyzed by observing cloud changes, which is of great significance to the prediction and research of disaster weather. On the other hand, many applications based on remote sensing technology, such as land cover classification, change detection, water area segmentation, and so on, are affected by cloud cover and often encounter problems, such as missed detection and false detection. Therefore, it is necessary to accurately identify clouds and cloud shadows.

The traditional cloud detection methods [1]–[3] used thresholds for cloud detection. Although the detection accuracy was improved to a certain extent, false detections and missed detections occurred from time to time. Moreover, the selection of spectrum and threshold value depended heavily on prior knowledge and was easily interfered by many factors. In recent years, deep convolutional neural networks (DCNNs) achieved great success in the field of computer vision. Some previous CNN-based methods [4]–[6] achieved excellent results in image classification, which laid the foundation for the pixel-level classification task, i.e., semantic segmentation. To achieve pixel-level classification, Long *et al.* [7] proposed the fully convolutional networks (FCNs). This method replaced the fully connected (FC) layer with a convolutional layer, which was effective for semantic segmentation tasks. Ronneberger *et al.* [8] proposed a U-shaped network structure (U-Net) to obtain contextual information and location information. The DeepLab proposed by Chen *et al.* [9] used Atrous convolution to expand the receptive field and used conditional random field (CRF) [10] to improve the ability of the model to capture details. The pyramid scene parsing network (PSPNet) proposed by Zhao *et al.* [11] could aggregate the contextual information of different regions, thereby improving the ability to obtain global information. For the cloud and cloud shadow segmentation task, Zhan *et al.* [12] designed a cloud detection network based on FCNs, which used multispectral satellite images to classify clouds and snow. Mohajerani and Saeedi [13] designed a network composed of FCNs, which could obtain better results under limited spectral band conditions. Although FCNs achieved end-to-end pixel-level prediction, a large number of misclassified regions appeared in this method in processing complex scenes of cloud and cloud shadow segmentation tasks. Li *et al.* [14] proposed a SegNet-based remote sensing multiangle cloud detection method to achieve effective cloud classification. SegNet [15] often had problems such as missed detection and false detection, which were caused by insufficient extraction of global context information. Yan *et al.* [16] used a pyramid pooling module (PPM) to extract contextual information. Also, they proposed a multilevel feature fused structure to combine semantic information with spatial information from different levels. Guo *et al.* [17] proposed a lightweight fully convolutional neural network (Cloudet), which was effective for large-scale cloud detection with larger receptive fields. Shao *et al.* [18] stacked many bands of an image to obtain the combined spectral information and used a method based on a multiscale features-convolutional neural network (MF-CNN) to detect thin clouds, thick clouds, and noncloud pixels. Compared with the traditional threshold method, these deep learning methods could process images of any size, and the learned features did not have to be manually adjusted,

Manuscript received May 8, 2022; accepted May 13, 2022. Date of publication May 16, 2022; date of current version May 26, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 42075130 and in part by the Postgraduate Research and Innovation Project of Jiangsu Province under Grant 1534052101072. *(Corresponding author: Min Xia.)*

Chen Lu, Min Xia, and Binyu Chen are with the Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, Jiangsu Key Laboratory of Big Data Analysis Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: xiamin@nuist.edu.cn).

Ming Qian is with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430072, China.

Digital Object Identifier 10.1109/TGRS.2022.3175613

which greatly increased the detection ability of the model. However, it was difficult for DCNNs to capture long-distance dependencies. Although this problem could be alleviated by expanding the receptive field, it could not capture global features after all. In particular, we hope that the model can see the global situation of the image, pay more attention to key information, and grasp the degree of correlation between pixels. Because each pixel in an image usually does not exist independently, it is more or less related to its surrounding or even further pixels. Global average pooling can help the model to know the global average of feature maps, but it cannot focus on key areas.

Some recent studies [19]–[21] extended the transformer to obtain global features, which was originally used in natural language processing tasks and demonstrated good performance in various visual tasks. Different from global average pooling, transformer could use the multihead attention mechanism to grasp the global information while paying attention to key areas. The core of the self-attention mechanism is to capture the correlation between pixels. In the calculation process of the self-attention mechanism, all pixels participate in the calculation, but the degree of participation of each pixel is not the same, which makes it possible to master the global information and focus on key areas. Dosovitskiy *et al.* [22] proposed a vision transformer (ViT) designed for image classification, and it directly applied the pure transformer module to the image block sequence to achieve image classification. This method obtained better results than convolution in multiple image classification tests. However, this method was aimed at image classification tasks and was not suitable for semantic segmentation tasks. To introduce transformer into dense prediction tasks, such as object detection and semantic segmentation, Wang *et al.* [23] proposed the pyramid vision transformer (PVT), which used pure transformer as the backbone such as the ViT, and introduced the pyramid structure to the transformer. The feature map was reduced, and the computational complexity and memory usage were reduced consequently, which was very effective for dense prediction. The convolutional vision transformer (CvT) proposed by Wu *et al.* [24] introduced convolutions into ViT to improve the performance of transformer and achieved the best result. However, these methods were still difficult to deal with some complex tasks, especially for cloud and cloud shadow segmentation, which used remote sensing image data. This was because satellite remote sensing images were larger, had diverse data types, and had many different spectral bands than images captured by traditional cameras with only three spectral bands of red green blue (RGB). Therefore, the methods mentioned earlier had limited capabilities for tasks, such as cloud and cloud shadow segmentation, that use remote sensing data: 1) the segmentation result was not accurate enough, and the segmentation boundary of cloud and cloud shadow was rough, and 2) under the interference of factors, such as surface objects and noise, false detections and missed detections were likely to occur, resulting in unsatisfactory segmentation results.

This work proposed a dual-branch model composed of transformer and convolutional neural network to solve the abovementioned problems. We used the transformer block in CvT and convolution as the backbone. On the one hand, transformer had dynamic attention, global context, and better generalization capabilities, which were not available in the convolution model [25], [26]. On the other hand, convolutional neural networks had shift, scaling, and distortion invariance, which the transformer module [27]–[29] lacked. Our proposed method effectively combined the two, so that the two branches could use each other's advantages, which helped the model extract features more efficiently. In terms of feature fusion, the use of mutual guidance module (MGM) enabled transformer branch and convolution branch to guide each other to perform feature mining and extract multiscale context information; as a result, the segmentation accuracy of clouds and cloud shadows of different scales was improved. In the decoding stage, the features of different levels extracted by the two branches were fully utilized for fusion and decoding, and hence, semantic information and spatial location information were effectively fused. As a result, the positioning of clouds and cloud shadows was more accurate, and segmentation boundaries show more details. Our main contributions were as follows.

1) We used the dual-branch network to classify clouds, cloud shadows, and backgrounds. This method was very useful when the spectral range was limited. It could complete end-to-end training without any manual parameter adjustment, making the process of cloud and cloud shadow detection very simple.
2) Considering the contextual relationship between clouds and cloud shadows, in the encoding stage, we designed a dual-branch model composed of transformer and convolution module to extract features of different scales and improved the ability to extract semantic information and spatial information at different scales. In the decoding stage, the model made full use of features extracted by dual branches for upsampling and gradually guided the restoration of the feature map, so that the positioning of clouds and cloud shadows was more accurate, and the segmentation boundary was clearer.
3) The proposed method used an MGM, so that the two branches could guide each other to extract features, and it improved the performance of cloud and cloud shadow segmentation and generated a clear segmentation boundary while accurately positioning the cloud and cloud shadow.

II. Methodology

This article proposed a dual-branch architecture composed of transformer and convolutional networks, which could effectively identify clouds and cloud shadows while generating clear and accurate segmentation boundaries. The overall architecture of the dual-branch network was shown in Fig. 1, which was mainly composed of encoder and decoder. Figs. 2 and 3 show the specific structure of each module in Fig. 1. For a given image of any size, we first used the dual-branch structure guided by transformer and convolutional network to extract features at different levels. Some previous studies [7], [8] simply fused high-level features and low-level features. Such fusion was insufficient, and there were still problems,

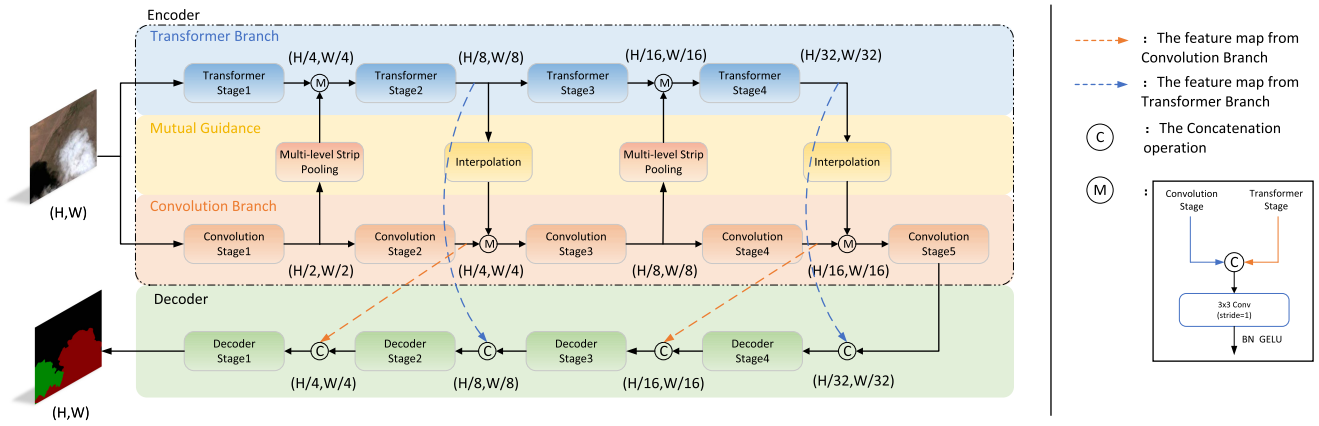

Fig. 1. Structure of dual-branch network. Conv represented convolution layer, BN represented batch normalization layer, and GELU represented the activation function GELU.

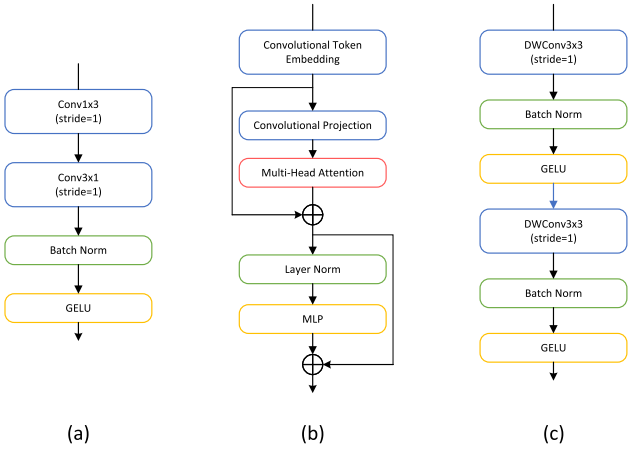

Fig. 2. Composition of different stages. (a) Composition of convolution stages. (b) Composition of transformer Stages. (c) Composition of decoder stages. Conv represented convolution layer, DWConv represented depthwise separable convolution layer, Batch Norm represented batch normalization layer, Layer Norm represented layer normalization layer, MLP represented multilayer perceptron, and GELU represented the activation function GELU.

such as missed detection, false detection, rough segmentation boundary, and so on, resulting in unsatisfactory segmentation results. The proposed method could combine respective advantages of transformer and convolutional networks to effectively fuse global features and local characteristics. In the decoding stage, aiming at the problem of inaccurate target positioning and rough segmentation boundaries caused by the loss of high-level semantic information and spatial detail information after upsampling, we used features from different levels of the two branches for fusion and achieved precise positioning and fine segmentation of clouds and cloud shadows.

A. Backbone

We used a dual-branch architecture of transformer and convolutional network as the backbone to extract features at different levels, and the specific structure was shown in Fig. 1. As we all knew, convolutional networks had excellent charac-

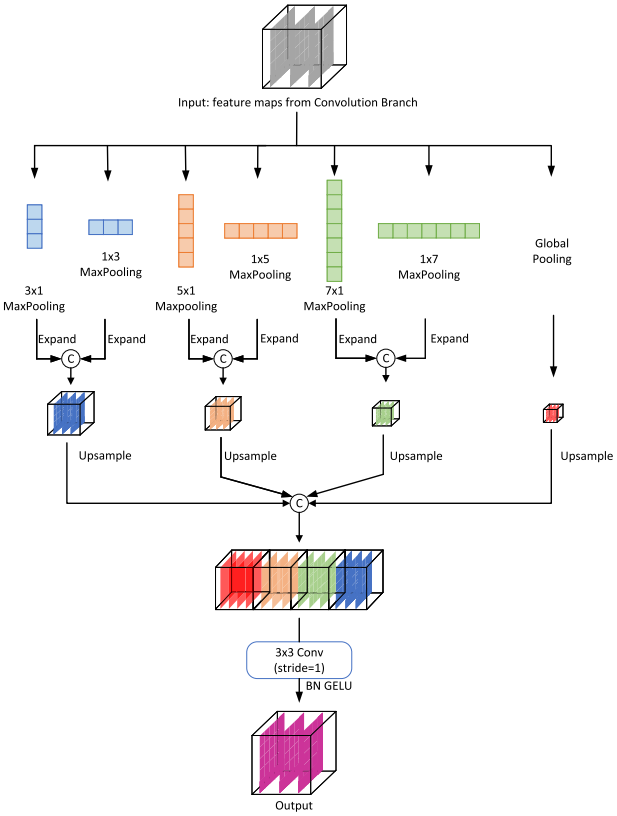

Fig. 3. Structure of multilevel strip pooling. Conv represented convolution layer, BN represented batch normalization layer, and GELU represented the activation function GELU.

teristics, such as shift, scaling, and distortion invariance, and transformer had dynamic attention, global receptive fields, and better generalization performance. When extracting features, the two could complement each other, thereby improving the feature extraction capability of the model, so that the model could extract high-level features more effectively. In particular, we used the transformer block in CvT as the module of transformer branch. This method shows better performance than pure convolution or pure transformer. Aiming at the

TABLE I
ARCHITECTURE OF THE PROPOSED METHOD

Level	Convolution Branch	Guidance Module	Transformer Branch	Decoder	Size
L1	Conv $\begin{bmatrix} 1\times3, 64, stride2 \\ 3\times1, 64, stride2 \\ 1\times3, 64, stride1 \\ 3\times1, 64, stride1 \end{bmatrix}$	$\longrightarrow$ Multi-level Strip Pooling	-	DWConv $\begin{bmatrix} 3\times3, 16, stride1, padding1 \\ 3\times3, 64, stride1, padding1 \end{bmatrix}$	1/2
L2	Conv $\begin{bmatrix} 1\times3, 128, stride2 \\ 3\times1, 128, stride2 \\ 1\times3, 128, stride1 \\ 3\times1, 128, stride1 \end{bmatrix}$	$\nwarrow$ $\quad$ $\searrow$	$Conv_{embed}(7\times7, 64, stride4, padding2)$ $Transformer(dim64, heads1)\times1$	DWConv $\begin{bmatrix} 3\times3, 192, stride1, padding1 \\ 3\times3, 256, stride1, padding1 \end{bmatrix}$	1/4
L3	Conv $\begin{bmatrix} 1\times3, 384, stride2 \\ 3\times1, 384, stride2 \\ 1\times3, 384, stride1 \\ 3\times1, 384, stride1 \end{bmatrix}$	$BilinearInterpolation$ $\longleftarrow$ $\longrightarrow$ Multi-level Strip Pooling	$Conv_{embed}(3\times3, 192, stride4, padding2)$ $Transformer(dim192, heads3)\times2$	DWConv $\begin{bmatrix} 3\times3, 384, stride1, padding1 \\ 3\times3, 512, stride1, padding1 \end{bmatrix}$	1/8
L4	Conv $\begin{bmatrix} 1\times3, 512, stride2 \\ 3\times1, 512, stride2 \\ 1\times3, 512, stride1 \\ 3\times1, 512, stride1 \end{bmatrix}$	$\nwarrow$ $\quad$ $\searrow$	$Conv_{embed}(3\times3, 384, stride4, padding2)$ $Transformer(dim384, heads6)\times10$	DWConv $\begin{bmatrix} 3\times3, 512, stride1, padding1 \\ 3\times3, 1024, stride1, padding1 \end{bmatrix}$	1/16
L5	Conv $\begin{bmatrix} 1\times3, 1024, stride2 \\ 3\times1, 1024, stride2 \\ 1\times3, 1024, stride1 \\ 3\times1, 1024, stride1 \end{bmatrix}$	$BilinearInterpolation$ $\longleftarrow$	$Conv_{embed}(3\times3, 384, stride4, padding2)$ $Transformer(dim384, heads6)\times10$	-	1/32

problem of rough segmentation boundaries in the existing methods, strip convolution was used as the module of convolution branch to extract the edge and position information of segmentation targets to achieve precise segmentation of cloud and cloud shadow boundaries. Table I shows the specific parameters of the model. We divided the feature layers with the same output size into one level and got a total of five levels. The expression of the transformer branch was as follows:

$$d_i = \begin{cases} x_0, & i = 0 \\ {x_i}^t, & i = 2, 4 \\ \text{Concat}\left({x_i}^t, {x_i}^{mc}\right), & i = 1, 3 \end{cases} \tag{1}$$

$$T_1 = \text{Conv}_{\text{embed}}(d_i) \tag{2}$$

$$T_2 = \text{MHA}\{\text{Flatten}[\text{Conv}_{\text{proj}}(T_1)]\} + T_1 \tag{3}$$

$${x_{i+1}}^t = \text{Reshape}\{\text{MLP}[\text{Norm}(T_2) + T_2]\}. \tag{4}$$

Among them, d_i represented the input matrix of the ith layer of the transformer branch ($i = 0, 1, 2, 3, 4$), x_0 represented the matrix input to the model, x_i^t and ${x_{i+1}}^t$, respectively, represented the output matrix of the ith and $i+1$th layers of the transformer branch, x_i^{mc} represented the feature map of the ith layer outputs of the convolution branch after multilevel strip pooling, Concat($\cdot$) meant concatenation, $\text{Conv}_{\text{embed}}(\cdot)$ indicated the convolutional embedding layer, $\text{Conv}_{\text{proj}}(\cdot)$ indicated the convolutional projection layer, Flatten($\cdot$) indicated the expansion of 2-D data into 1-D data, MHA($\cdot$) indicated the multihead attention layer, Norm($\cdot$) meant layer normalization, MLP($\cdot$) meant multilayer perceptron, and Reshape($\cdot$) meant changing 1-D data into 2-D data.

The expression of the convolution branch was as follows:

$$e_i = \begin{cases} x_0, & i = 0 \\ {x_i}^c, & i = 1, 3, 5 \\ \text{Concat}\left({x_i}^c, {x_i}^{ut}\right), & i = 2, 4 \end{cases} \tag{5}$$

$$C_1 = \sigma\{\text{BN}[\text{Conv}_{1\times3}(e_i)]\} \tag{6}$$

$${x_{i+1}}^c = \sigma\{\text{BN}[\text{Conv}_{3\times1}(C_1)]\} \tag{7}$$

where e_i represented the inputs of the ith layer of the convolution branch ($i = 0, 1, 2, 3, 4$), x_0 represented the original image of the inputs, ${x_i}^c$ and ${x_{i+1}}^c$, respectively, represented the outputs of the ith layer and the $i+1$th layer of the convolution branch, ${x_i}^{ut}$ indicated that the ith layer of the transformer branch was upsampled by bilinear interpolation to be a feature map of the same size as ${x_i}^c$, $\sigma(\cdot)$ denoted the activation function GELU, BN($\cdot$) denoted the batch normalization layer, and $\text{Conv}_{1\times3}(\cdot)$ and $\text{Conv}_{3\times1}(\cdot)$, respectively, denoted convolutions with a kernel size of 1×3 and 3×1.

The receptive field of convolution was limited [30]–[32]; to extract high-level semantic information, some of the previous methods were implemented using Atrous convolution or self-attention. These methods were effective in detecting large-scale clouds, but the detection of small-scale clouds and cloud shadows was not good, because the large square window would extract too much information from irrelevant areas, which interfered model prediction. In response to the above-mentioned problems, Hou *et al.* [33] proposed strip pooling, which could reduce interference in irrelevant areas. Inspired by this method, we designed the convolution branch consisting of strip convolutions to extract multiscale spatial information, which enhanced the detection ability of small targets and increased the detailed information of the segmentation boundary. The basic strip convolution block consisted of 1×3 and 3×1 strip convolutions, batch normalization layer, and activation function GELU. On the one hand, clouds and cloud shadows had similar shapes, and they might be classified by contextual information. On the other hand, the strip-shaped kernel could reduce the interference of irrelevant areas, so that the model could more effectively identify smaller-scale clouds and cloud shadows, and achieved accurate classification.

B. Mutual Guidance Module

The shapes of clouds and cloud shadows were similar, which often leaded to misjudgments in the detection process. In addition, interference from surface objects (such as houses, vegetation, and other objects with similar attributes to clouds and cloud shadows) and noise could also cause misclassification. We believed this was caused by insufficient fusion of location information and category information. To obtain more accurate results, the high-level features with rich category information extracted by transformer branch could be weighted to low-level features extracted by the convolution branch. Compared with

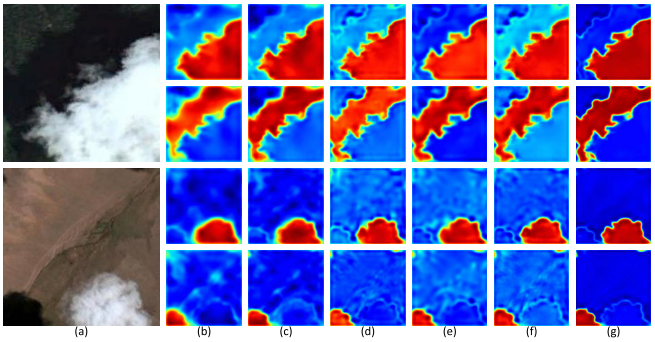

Fig. 4. Heat map representation. (a) Test image. (b) Transformer branch. (c) Transformer branch + convolution branch [T]. (d) Transformer branch + convolution branch [C]. (e) Transformer branch + convolution branch + MGM [T]. (f) Transformer branch + convolution branch + MGM [C]. (g) Proposed method (transformer branch + convolution branch + MGM + IFD). Among them, [T] represented decoding only with the features of transformer branch. [C] represented decoding only with the feature of convolution branch. The first row of each sample was the attention to clouds, and the second row was sattention to cloud shadows.

high-level features, low-level features extracted by the convolution branch retained more spatial information. Therefore, the convolution branch could provide location information guidance for the transformer branch for deep semantic feature mining.

Inspired by BiseNet [34], [35], we designed an MGM for feature fusion. This module consisted of two parallel lines; one line sent the features extracted by the convolution branch into a multilevel strip pooling to capture low-level information of different scales. It included global pooling, $1 \times n$ max pooling, and $n \times 1$ max pooling ($n = 3, 5, 7$). This module further extracted multiscale features based on the features extracted by the convolution branch, then resized these features to the same size of the transformer branch at the same level, and finally fused them through concatenation. The other line used bilinear interpolation to upscale high-level features to the same size of convolution branch at the same level, then concatenated the two, and fed them to the subsequent network. On the one hand, high-level features had rich category information, which could guide low-level features to classify. On the other hand, low-level features retained relatively more location information, which could complement high-level features in spatial location information.

Different from DeepLab's scheme of adding an atrous spatial pyramid pooling (ASPP) module at the end of the backbone network to extract multiscale features, we used multilevel strip pooling (Fig. 3) in the first and third levels of the convolution branch to extract the multiscale features of this layer and concatenate them with the features extracted by transformer. These features were then fed into subsequent level of the transformer branch. It is known that max pooling can extract the spatial features of images very well, so we used the strip max pooling modules of different scales to supplement the transformer branch with rich semantic features. The third and fifth levels of the convolution branch received the features from the upper level and the features with rich semantic information extracted by the second and fourth levels of the transformer branch, which were obtained after interpolation. The multilevel strip pooling module and the interpolation method enabled the features extracted by the two branches to be fused with each other to achieve mutual guidance between the two branches. Therefore, we call the module composed of multilevel strip pooling and interpolation as the MGM. It is known that both ASPP and PPM were used to extract multiscale information, so we compared the proposed multilevel strip pooling with them. As shown in Table IV, the model with MGM yielded better results than with ASPP or PPM.

It could be seen that MGM can improve the performance of the model no matter which way the model was used for decoding in Fig. 4. In particular, by comparing Fig. 4(c) and (e), it could be seen that the model that only used the features from transformer for decoding paid more attention to the interior of clouds and cloud shadows, but not enough attention to the boundary. After adding MGM, the boundary became more clearer. The results show that MGM could further extract multiscale features from the convolution branch and supplemented the transformer branch, so as to realize the guidance of the convolution branch for the transformer branch. The comparison of Fig. 4(d) and (f) shows that the model that only used the features from convolution branch for decoding had insufficient attention to the interior of the cloud and cloud shadow. After adding MGM, the amount of attention to the interior of cloud and cloud shadow was supplemented. This shows that MGM could further supplement the convolution branch and realize the guidance of the transformer branch for the convolution branch.

C. Interleaved Fusion Decoding

The size and shape of clouds and cloud shadows were variable, which made it very difficult to detect their boundaries. In the existing methods, the segmentation boundary of the feature map after multiple downsampling and upsampling was very rough, and the details were insufficient. The method proposed in this article used the dual-branch architecture in the encoding part to extract high-level semantic features and spatial position information, respectively. In the decoding stage, our method used the interleaved upsampling of the two branch features to fully fuse the features extracted from these two branches, making the segmentation boundary clearer and more detailed. Its specific operations are as follows:

$$D_i' = \mathrm{Up}(\sigma\{\mathrm{BN}[\mathrm{DWConv}(D_i)]\}), \quad i = 1, 2, 3, 4 \tag{8}$$

$$D_i = \sigma\{\mathrm{BN}[\mathrm{DWConv}(M_i)]\}, \quad i = 1, 2, 3, 4 \tag{9}$$

$$M_i = \begin{cases} \mathrm{Concat}\left(D_{i+1}, {x_i}^c\right), & i = 1, 3 \\ \mathrm{Concat}\left(D_{i+1}, {x_i}^t\right), & i = 2 \\ \mathrm{Concat}\left({x_i}^t, {x_{i+1}}^c\right), & i = 4 \end{cases} \tag{10}$$

where ${x_i}^t$ and ${x_i}^c$ represented the outputs of the ith layer of the transformer branch and the convolution branch, respectively, D_i' represented the outputs of the ith layer of the decoder, $\mathrm{Up}(\cdot)$ represented the bilinear interpolation upsampling, $\sigma(\cdot)$ represented the function GELU, $\mathrm{BN}(\cdot)$ represented the batch normalization, $\mathrm{DWConv}(\cdot)$ represented the depthwise separable convolution, and $\mathrm{Concat}(\cdot)$ represented the concatenation.

By comparing Fig. 4(e)–(g), it could be seen that the result of decoding using interleaved fusion decoding (IFD) was better

than only using features from transformer branch or features from convolution branch. The proposed method using IFD paid more attention to clouds and cloud shadows and had clearer boundaries. This shows that IFD could effectively use the features of the two branches to repair the lost information during the decoding process.

III. Experimental Analysis

A. Datasets

1) Cloud and Cloud Shadow Dataset: This dataset was collected from the Landsat-8 and Sentinel-2. The Landsat-8 carried two sensors, operational land imager (OLI) and thermal infrared sensor (TIRS), with a total of 11 bands. The data collected by OIL had nine bands with a spatial resolution of 30 m. The Sentinel-2 carried a multispectral image that could cover 13 spectral bands. We used bands 2–4 of the two satellites to evaluate the model's performance under limited spectral information. Table II shows the detailed information of each band of the two satellites.

The resolution of the original remote sensing image was usually very large. Due to the limitation of GPU memory, it was difficult to achieve training of the original size image. Therefore, we cropped each remote sensing image into a small image of 224 × 224 pixels for model training. A total of 25 314 small images were obtained. To train and test the model separately, according to the experimental requirements, the dataset was divided into a training set (20 000 pictures) and a validation set (5314 pictures). The ratio of the training set to the test set was 8:2. Our dataset used three semantically classified pixel tags of cloud, cloud shadow, and background to test the performance of the model in different types of scenes. The dataset had a wide range of collection, including complex areas, such as plateaus, plains, hills, cities, and farmland.

2) HRC_WHU Dataset: To further verify the generalization performance of the proposed algorithm, we used the high-resolution cloud cover validation dataset created by researchers at Wuhan University (HRC_WHU) dataset created by Li *et al.* [36] to test the generalization performance of the model. The images of this dataset were collected from Google Earth. Experts in the field of remote sensing image interpretation of Wuhan University digitally processed the relevant reference cloud masks. There were 150 pictures in total, and each picture was composed of three RGB channels. It had the characteristics of high resolution and diverse scenes. The resolution of each picture was 1280 × 720. We cropped the picture into a 256 × 256 small image for training. When predicting, the original image was also cropped into 256 × 256 small images for prediction, and these small images were finally spliced together to synthesize a complete prediction image with a size of 1280 × 720. Fig. 5 shows the images of Landsat-8, Sentinel-2, and HRC_WHU. In addition, to prevent overfitting and enhance the generalization ability of the model, we randomly flipped and rotated images and added noise to the dataset to achieve data enhancement.

3) SPARCS Dataset: We used the public dataset Spatial Procedures for Automated Removal of Cloud and Shadow (SPARCS) [37] to verify the performance of the proposed method in multispectral scenarios, as shown in Table III. This dataset was originally created by M. Joseph Hughes, Oregon State University, and was derived manually from pre-collection Landsat-8 scenes. This collection of cloud validation data contains 80 1000 × 1000 pixels, subsets of pre-collection Landsat-8 scenarios, including the classes of cloud, cloud shadow, snow/ice, water, and background. Limited by the video memory capacity of the GPU, we cut the original dataset into 256 × 256 pixels and got a total of 1280 images. We split the dataset into training and validation sets in a ratio of 8:2. To enhance the generalization ability of the model, we expanded the dataset by horizontal flipping, vertical flipping, and random rotation, and finally got a total of 5068 training sets and 1268 validation sets.

TABLE II
Cloud and Cloud Shadow Dataset

Satellite	Band	Wavelength(μm)	Resolution(m)
Landsat-8	Band 2 (Blue)	0.45 - 0.51	30
	Band 3 (Green)	0.53 - 0.59	30
	Band 4 (Red)	0.64 - 0.67	30
Sentinel-2	Band 2 (Blue)	0.45 - 0.52	10
	Band 3 (Green)	0.54 - 0.57	10
	Band 4 (Red)	0.65 - 0.68	10

TABLE III
SPARCS Dataset

Band	Wavelength(μm)	Resolution(m)
Band 1 (Coastal)	0.43 - 0.45	30
Band 2 (Blue)	0.45 - 0.51	30
Band 3 (Green)	0.53 - 0.59	30
Band 4 (Red)	0.64 - 0.67	30
Band 5 (NIR)	0.85 - 0.88	30
Band 6 (SWIR-1)	1.57 - 1.65	30
Band 7 (SWIR-2)	2.11 - 2.29	30
Band 8 (Pan)	0.50 - 0.68	15
Band 9 (Cirrus)	1.36 - 1.38	30
Band 10 (TIRS-1)	10.6 - 11.19	100

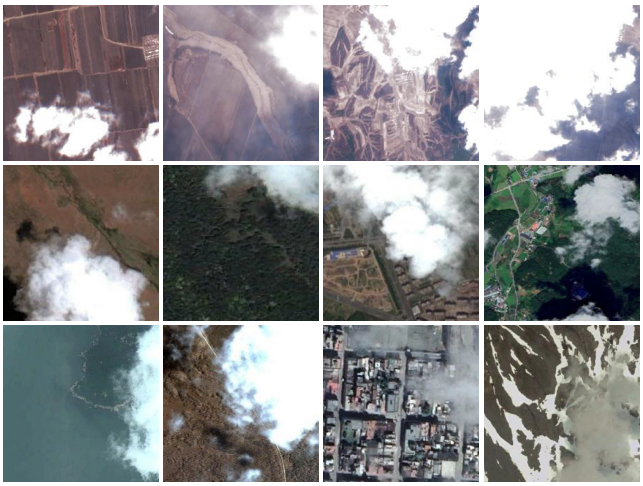

Fig. 5. Some of the training data in Landsat-8 (first row), Sentinel-2 (second row), and HRC_WHU (third row).

B. Experiment Details

1) Optimization: The experiment was carried out on an RTX 3080 GPU using the Pytorch. In this article, stochastic gradient descent (SGD) was used to optimize the momentum with a coefficient of 0.9, and the poly learning rate (LR) policy

was used in training, the initial LR was set to 0.001, and the poly power was 2. The model was trained 300 epochs in total. To prevent overfitting, we used $L2$ regularization and set the weight attenuation to 0.01. The LR of each round of training was described as follows:

$$\mathrm{LR} = 0.001 \times \left(1 - \frac{E}{300}\right)^{\gamma} \tag{11}$$

where E was the number of training epochs, and γ was 2.

2) Metrics: We chose precision (P), recall (R), F_1 score, overall accuracy (OA), pixel accuracy (PA), mean pixel accuracy (MPA), and mean intersection over union (MIoU) to evaluate the performance of this method in cloud and cloud shadow segmentation tasks. The calculation formula of each evaluation index was as follows:

$$P = \frac{(\mathrm{TP})}{(\mathrm{TP}) + (\mathrm{FP})} \tag{12}$$

$$R = \frac{(\mathrm{TP})}{(\mathrm{TP}) + (\mathrm{FN})} \tag{13}$$

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{14}$$

$$\mathrm{PA} = \frac{\sum_{i=0}^{k} p_{i,i}}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{i,j}} \tag{15}$$

$$\mathrm{MPA} = \frac{1}{k} \sum_{i=0}^{k} \frac{p_{i,i}}{\sum_{j=0}^{k} p_{i,j}} \tag{16}$$

$$\mathrm{MIoU} = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{i,i}}{\sum_{j=0}^{k} p_{i,j} + \sum_{j=0}^{k} p_{j,i} - p_{i,i}} \tag{17}$$

where true positive (TP) represented the number of cloud (cloud shadow) pixels that were correctly predicted, false positive (FP) represented the number of cloud (cloud shadow) pixels that are predicted incorrectly, and true negative (TN) represented correctly classified noncloud (noncloud) pixels. False negative (FN) represented misclassified cloud (cloud shadow) pixels, k represented the number of categories (excluding background), $p_{i,i}$ represented the number of TPs, and $p_{i,j}$ represented the number that belongs to category i but was predicted to be category j.

C. Ablation Study

In this section, we first used the original CvT as the benchmark model, and we upsampled each of its layers and then concatenated them together for output. Next, we gradually added the proposed modules to the model to verify the feasibility of each module and the entire model. The research in this section mainly used MioU to evaluate. Table IV shows the result, and it could be seen that the proposed model including all modules achieved the best performance.

1) Ablation for Dual Branch: To extract multiscale spatial position information, the proposed convolution branch had shift, scaling, and distortion invariance and shows excellent performance in extracting local features. In addition, the use of a long strip of convolution window avoided the interference of too many irrelevant areas. The model with the dual branch increased MIoU to 92.14%, which shows that the dual branch was effective in extracting spatial and semantic information.

TABLE IV
ABLATION FOR DIFFERENT MODULES IN THE MODEL

Method	MIoU(%)
Transformer Branch	91.57
Transformer Branch + Convolution Branch	**92.14(0.57↑)**
Transformer Branch + Convolution Branch + PPM	92.51
Transformer Branch + Convolution Branch + ASPP	92.73
Transformer Branch + Convolution Branch + MGM	**93.29(1.15↑)**
Transformer Branch + Convolution Branch + MGM + IFD	**93.42(0.13↑)**

2) Ablation for MGM: This module was used to fuse the global context information extracted by the transformer branch and the spatial position information extracted by the convolution branch, so that the high-level features with accurate category information and the rich low-level features guided each other, which helped to improve the recognition ability of similar objects. It is known that the ASPP module in DeepLab and the PPM module in PSPNet are also used to extract multiscale information. We compared the proposed multilevel strip pooling with the ASPP module and the PPM module, as shown in Table IV, and our proposed method was the best performer among these modules. The results in Table IV show that the MGM was an effective module, which could increase the MIoU of the model by 1.15%.

3) Ablation for IFD: Due to inconstant shapes and sizes of clouds and cloud shadows, the existing methods [7], [11], [38] generated very rough boundaries and insufficient details. To repair edge details, we added the features extracted by the transformer branch and the convolution branch for decoding to improve the segmentation accuracy of the boundary. Compared with the single branch decoding method, the IFD performed better. This method could further improve the MIoU of the model from 93.29% to 93.42%.

D. Comparison Test of the Cloud and Cloud Shadow Dataset

In this section, the proposed method was compared with the current excellent models, such as FCN, pyramid attention network (PAN), PSPNet, DeepLab V3plus, and so on, for the purpose of proving the feasibility of the algorithm. Each of these methods had its own characteristics. FCN adopted a fully convolutional structure to achieve pixel-level classification. To extract multiscale semantic information, PSPNet used pooling layers with different sizes, and the DeepLab series adopted ASPP modules composed of Atrous convolutions with different Atrous rates. However, Li *et al.* [38] believed that Atrous convolution may lead to grid artifact, and PPM in PSPNet may lead to loss of pixel-level localization information. The feature pyramid attention (FPA) module was used to extract multiscale information. In terms of real-time semantic segmentation, BiSeNetV2, ExtremeC3Net, and DANet adopted a dual-branch architecture to extract spatial and semantic information, but their composition is different. LinkNet innovated in the way of connecting encoder and decoder. CGNet used a context guided (CG) module to construct a model, which was characterized by the ability to combine local features and surrounding features, and used global features to improve the joint features. In the model constructed based on the transformer, PVT introduced the pyramid structure into the transformer, so that the feature

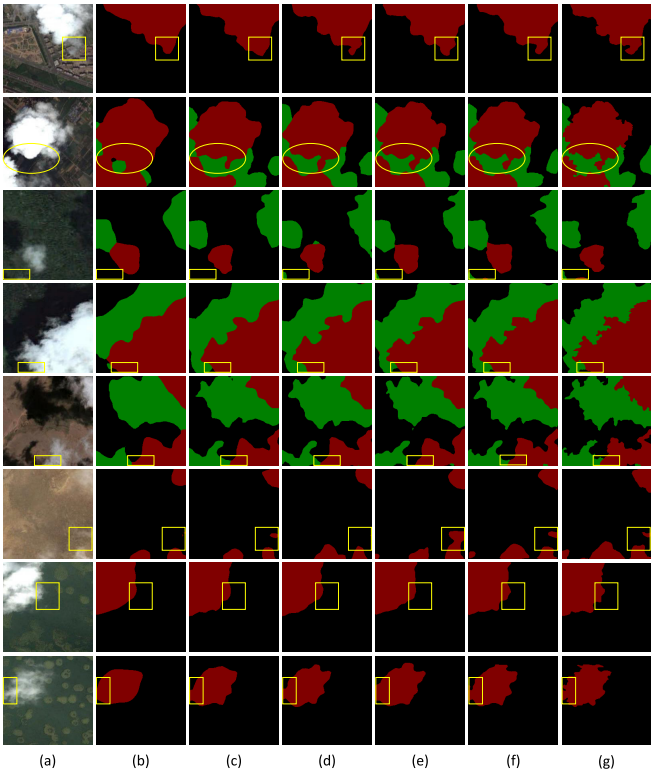

Fig. 6. Comparison of different methods in various scenarios. (a) Test image. (b) Segmentation results of PAN. (c) Segmentation results of BiseNet V2. (d) Segmentation results of PSPNet. (e) Segmentation results of DeepLab V3plus. (f) Segmentation results of our method. (g) Label.

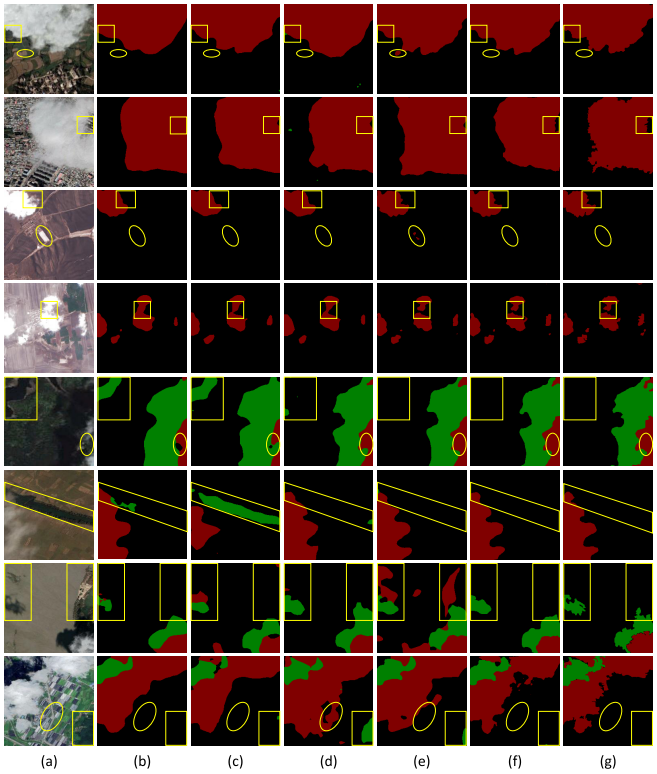

Fig. 7. Comparison of misdetection of different models. (a) Test image. (b) Segmentation results of PAN. (c) Segmentation results of BiseNet V2. (d) Segmentation results of PSPNet. (e) Segmentation results of DeepLab V3plus. (f) Segmentation results of our method. (g) Label.

map can be gradually reduced, which is very suitable for dense prediction. CvT introduced convolution into transformer to improve performance.

Table V shows the comparison results of different methods. For cloud detection, our method was better than the other methods in OA and F_1 scores, reaching 98.76% and 97.08%, respectively. For cloud shadow detection, our proposed method was the best in OA, R, and F_1 scores, reaching 98.73%, 94.39%, and 94.39%, respectively. Although P and R on cloud detection and P on cloud shadow detection were not the highest, there were only a small gap behind the best-performing method. The right-hand side of Table V shows the comparison of different models on the three metrics of PA, MPA, and MIoU. From the performance point of view, FCN-8S and DANet had the worst performance, followed by PAN, BiseNet V2, PSPNet, CGNet, and CvT. DeepLab V3plus, LinkNet, ExtremeC3, PVT, modified VGG, CloudNet, and GAFFNet performed better. However, they were not as good as the model proposed in this article. The PA, MPA, and MioU of our proposed model reached 97.56%, 96.77%, and 93.42%, respectively.

Fig. 6 shows the comparison of the segmentation results of cloud and cloud shadow images by different methods, where the red represented the cloud, the green represented the cloud shadow, and the black was the background. It could be seen that the segmentation results of PAN and BiseNet V2 were rough, small clouds and cloud shadows were not detected, and there were many missed detection and false detection pixels along the boundary of clouds and cloud shadows. The segmentation of PSPNet and DeepLab V3plus was much better than PAN and BiseNet V2. They could accurately locate clouds and cloud shadows. The detection ability of small-scale clouds and cloud shadows was also improved, but the details were insufficient, especially the boundary details were seriously lost. Our proposed method used transformer branch and convolution branch to fully extract multiscale context information and used MGM to fuse the features of the two branches, which could effectively fuse high-level semantic information and spatial position information, so that the model could accurately locate clouds and cloud shadows. In addition, we used the convolution branch to guide the model to repair the detailed information of the segmentation boundary in the decoding stage, which reduced the problem of severe loss of boundary details after deep downsampling and was crucial for cloud and cloud shadow segmentation. Experiments show that our proposed method not only improved the accuracy of cloud and cloud shadow segmentation, but also kept rich boundary details. Among these methods, our proposed method shows the best performance.

Fig. 7 selected the test results of different models in complex scenarios, such as urban areas, towns, farmland, and water areas, that are prone to false detections and missed detections, which further verified the effectiveness of the method under the interference of surface objects. The red represented clouds, the green represented cloud shadows, and the black represented background. The first and second sets of images

TABLE V
COMPARISON OF EVALUATION METRICS OF DIFFERENT MODELS ON CLOUD AND CLOUD SHADOW DATASET

	Cloud				Cloud Shadow						
Method	OA(%)	P(%)	R(%)	F_1(%)	OA(%)	P(%)	R(%)	F_1(%)	PA(%)	MPA(%)	MIoU(%)
FCN-8S [7]	95.87	88.47	92.8	90.63	97.19	86.87	88.72	87.79	93.40	90.52	84.01
PAN [38]	98.25	96.29	95.49	95.89	98.31	92.71	92.46	92.59	96.73	95.52	91.26
BiseNet V2 [35]	98.27	96.57	95.28	95.92	98.34	94.18	90.99	92.59	96.68	95.96	91.20
PSPNet [11]	98.35	96.13	96.13	96.13	98.40	92.70	93.27	92.99	96.87	95.55	91.69
DeepLab V3Plus [9]	98.65	97.70	95.94	96.82	98.66	93.88	94.32	94.10	97.37	96.48	92.99
LinkNet [39]	98.61	96.59	96.91	96.75	98.54	94.19	92.91	93.55	97.23	96.24	92.55
ExtremeC3Net [40]	98.64	97.32	96.28	96.80	98.60	**94.68**	92.95	93.82	97.30	96.57	92.76
DANet [41]	96.45	91.68	91.72	91.71	97.29	88.40	87.68	88.04	94.03	91.93	85.07
CGNet [42]	98.37	95.93	96.48	96.20	98.27	93.33	91.34	92.34	96.73	95.60	91.27
PVT [23]	98.57	97.45	95.84	96.65	98.55	93.28	94.08	93.68	97.21	96.18	92.55
CvT [24]	98.44	95.89	**96.88**	96.38	98.32	92.90	92.24	92.57	96.85	95.54	91.57
modified VGG [12]	98.40	**98.13**	94.30	96.22	98.57	94.41	92.88	93.64	97.04	96.56	92.17
CloudNet [13]	98.70	97.22	96.68	96.95	98.40	92.05	94.05	93.05	97.17	95.77	92.36
GAFFNet [43]	98.53	96.49	96.63	96.56	98.41	92.71	93.40	93.05	97.06	95.73	92.08
Our	**98.76**	97.95	96.22	**97.08**	**98.73**	94.39	**94.39**	**94.39**	**97.56**	**96.77**	**93.42**

were taken over the urban area. The third and fourth sets of images were taken with the town as the background. In these groups of pictures, there were white houses on the ground. The white houses and clouds had similar attributes. Under the interference of houses, although PAN and BiseNet V2 did not have the problem of misdetecting surface objects as cloud clusters or cloud shadows, such as PSPNet and DeepLab V3plus, the segmentation boundary was very rough. This was caused by inadequate extraction of semantic information. On the other hand, PSPNet and DeepLab V3plus excessively extracted high-level semantic information while ignoring spatial location information, which was the cause of their misclassification. Also, the method we proposed shows better results. The fifth and sixth groups of pictures were segmentation results with river channels as the background. Because of weather, satellite shooting angles, and other reasons, some of the waters captured were dark and black, which was very similar to cloud shadows and was very confusing. Due to the insufficient fusion of semantic information and spatial location information, these models had different degrees of misdetection, except for our proposed method. PAN and BiseNet V2 classified most of the rivers into cloud shadows. The segmentation of PSPNet and DeepLab V3plus was slightly better than that of PAN and BiseNet V2, but their segmentation results were not as good as the method proposed in this article. The seventh and eighth sets of pictures show the segmentation results of different methods under the co-interference of water, vegetations, farmland, and houses. It could be seen that the segmentation results of PAN and BiseNet V2 not only had rough segmentation boundaries, but missed small clouds and cloud shadows; in the result, a large area of rivers and farmland was tagged into clouds. The segmentation accuracy of PSPNet and DeepLab V3plus was improved, but there were still cases where houses are classified as clouds and green plants as cloud shadows. Our model had detailed segmentation boundaries, and there were no large-scale false detections and missed detections. The method proposed in this article used transformer branch and convolution branch to extract high-level semantic information and spatial position information, respectively, which reduced the problem of insufficient extraction of semantic information or spatial position information in the existing methods. Our method used MGM to effectively and fully integrated high-level features and low-level features. Compared with the ordinary square convolution kernel, the strip convolution kernel could avoid the interference of excessive background on the extraction of cloud and cloud shadow segmentation boundary information. For the problem of rough segmentation boundary, different features extracted by transformer branch and convolution branch were fully combined in the decoding stage to strengthen semantic information and edge detail information. These modules work together to greatly mitigate the problems of false detection and missed detection caused by the interference of surface objects, thereby obtaining excellent segmentation results. Experiments proved that our proposed method could achieve precise segmentation in response to ground interference, greatly reducing the occurrence of misclassification, and was the best of these methods.

Fig. 8 shows the segmentation results of various methods under noise interference. It could be seen that under the interference of noise, the segmentation boundaries of PAN, BiseNet V2, and PSPNet were blurred, and there were different degrees of missed detection and false detection. This was because there was a lot of noise along the cloud boundary, which could mislead the judgment of the model. In addition, the semantic information and spatial information extracted by these methods were insufficient. In the case of noise interference, accurate segmentation could not be achieved, especially the detection capabilities of segmentation boundaries and small-scale clouds and cloud shadows. In contrast, DeepLab V3plus had strong anti-interference ability and fine segmentation. It could be found that the number of misclassified points was reduced. With the help of transformer branch and convolution branch, the proposed method could effectively extract and fuse multiscale semantic information and spatial information and used the features extracted from different

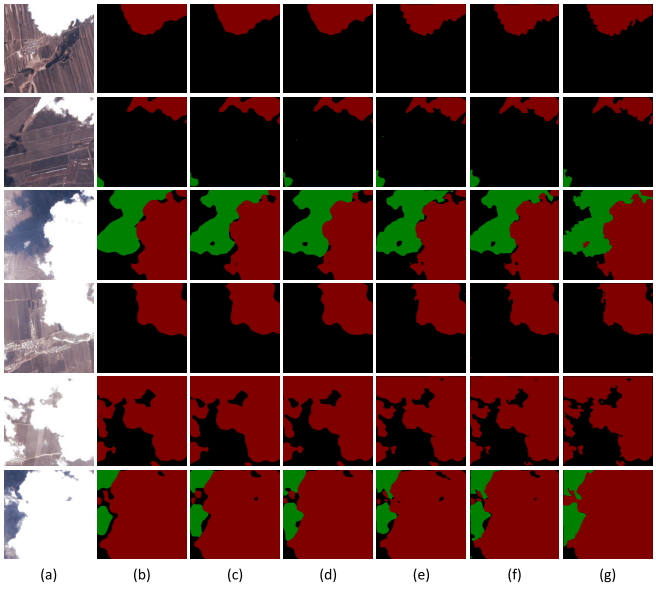

Fig. 8. Comparison of different models under noise. (a) Test image. (b) Segmentation results of PAN. (c) Segmentation results of BiseNet V2. (d) Segmentation results of PSPNet. (e) Segmentation results of DeepLab V3plus. (f) Segmentation results of our method. (g) Label.

branches to perform upsampling in the decoding stage to form accurate result. Therefore, the segmentation performance of our proposed method under noise interference was the best among the abovementioned methods.

Fig. 9 shows the heatmaps from the attention of different models, where the first row of each sample was the model's attention to clouds, and the second row was the attention to cloud shadows. The red region was the area that the model paid the most attention to, followed by orange and green, and blue is the region where no attention was needed. From the results of the heat map, PAN did not pay enough attention to clouds and cloud shadows, which also leaded to poor final segmentation results. The attention of BiseNet V2, PSPNet, and DeepLab V3plus was more concentrated, but the attention to the boundary was still insufficient, and some attention was put to the place without any target. The method we propose could reduce the distraction, so that the attention could be focused on the target objects, that was, clouds and cloud shadows, especially on the target boundary, so it could obtain more accurate target location and finer boundary segmentation.

E. Comparison Test of the HRC_WHU Dataset

Fig. 10 shows the segmentation of different models in five different scenarios on the HRC_WHU dataset. In Fig. 10, the white was classified as the cloud, and the black was the background. Through comparison, it could be seen that the segmentation results of PAN and BiseNet V2 were not good in all scenarios, especially the edge information of the cloud was severely lost. Also, there were a large number of missed and misdetected points under complex background, such as snow and house interference. In contrast, PSPNet and DeepLab V3plus demonstrated much stronger anti-interference ability, and the numbers of false detection and missed detection points were reduced, but the boundary information of the cloud was

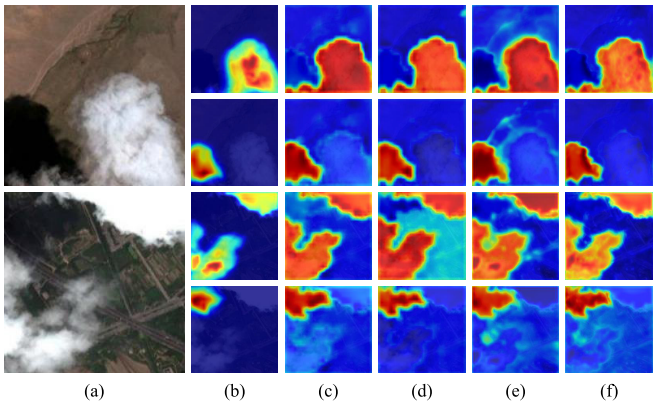

Fig. 9. Visual contract of different modules. (a) Test image. (b) PAN. (c) BiseNet V2. (d) PSPNet. (e) DeepLab V3plus. (f) Our method.

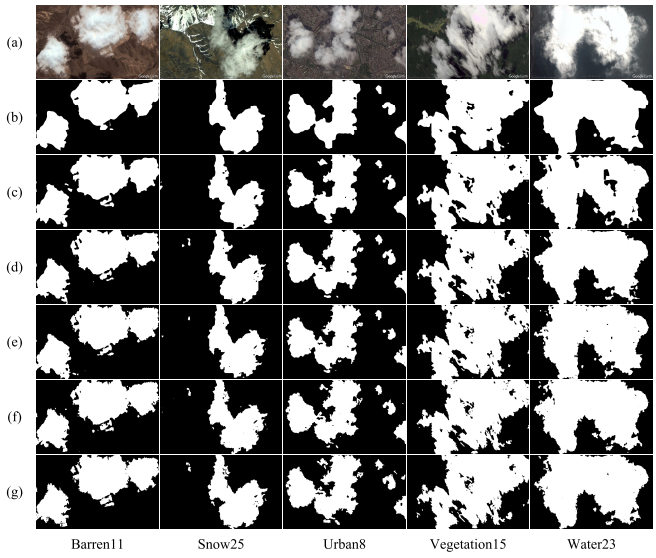

Fig. 10. Comparison of different models on the HRC_WHU Dataset. (a) Test image. (b) Segmentation results of PAN. (c) Segmentation results of BiseNet V2. (d) Segmentation results of PSPNet. (e) Segmentation results of DeepLab V3plus. (f) Segmentation results of our method. (g) Label.

still rough, and the details were insufficient. The proposed mothed achieved the best results, due to the fact that the MGM effectively allowed the two branches to form mutual guidance, thereby achieving better detection and location of clouds. In addition, because of the full use of features from the two branches in the decoding stage, the segmentation boundary was clearer, and the details were richer.

F. Comparison Test of the SPARCS Dataset

To further demonstrate the segmentation performance of the proposed method for multispectral remote sensing images, we also conducted comparative experiments on the SPARCS dataset, and the experimental results are shown in Table VI. The left-hand side of Table VI shows the PA of different models for each category. It can be seen that our model had the highest PA in the four categories of cloud, cloud shadow, snow/ice, and water, reaching 91.12%, 78.38%, 96.59%, and 89.99%, respectively. Although the PA of land was not the highest, it was very close to the highest model. On the right-hand side of Table VI, we can see that our proposed

TABLE VI
COMPARISON OF EVALUATION METRICS OF DIFFERENT MODELS ON THE SPARCS DATASET

	Class Pixel Accuracy					Overall Results				
Method	Cloud(%)	Cloud Shadow(%)	Snow/Ice(%)	Water(%)	Land(%)	PA(%)	Recall(%)	Precision(%)	F_1(%)	MIoU(%)
PAN [38]	89.10	75.27	86.60	79.96	95.64	91.20	87.34	85.32	81.53	76.57
BiSeNet V2 [35]	85.87	64.75	93.84	81.44	97.17	91.31	89.77	84.61	83.09	77.79
PSPNet [11]	90.79	63.75	94.22	77.73	96.84	91.78	90.29	84.67	83.48	78.20
DeepLab V3Plus [9]	87.81	72.12	85.17	81.27	**97.84**	91.99	90.75	84.85	84.01	78.44
LinkNet [39]	85.35	74.38	91.92	80.30	96.44	91.31	88.66	85.68	82.81	77.87
ExtremeC3Net [40]	91.09	75.47	95.43	83.62	96.13	92.77	90.32	88.35	85.46	81.29
DANet [41]	82.06	42.25	91.28	73.65	95.03	86.92	83.86	76.85	74.64	68.33
CGNet [42]	90.63	72.78	95.37	83.30	96.51	93.22	91.00	88.95	86.30	82.28
PVT [23]	88.22	75.77	92.00	86.27	95.92	92.02	89.76	87.64	84.66	80.24
CvT [24]	88.24	71.63	95.41	87.71	96.14	92.17	89.83	87.83	84.80	80.55
modified VGG [12]	85.55	58.53	94.87	79.35	95.98	89.99	86.38	82.85	79.36	74.00
CloudNet [13]	85.99	74.58	91.78	80.34	96.52	91.50	88.49	85.84	82.79	77.95
GAFFNet [43]	86.97	59.00	85.21	78.06	94.47	88.62	86.70	80.74	78.56	72.32
our	**91.12**	**78.38**	**96.59**	**89.99**	97.52	**94.31**	**92.90**	**90.72**	**88.83**	**85.26**

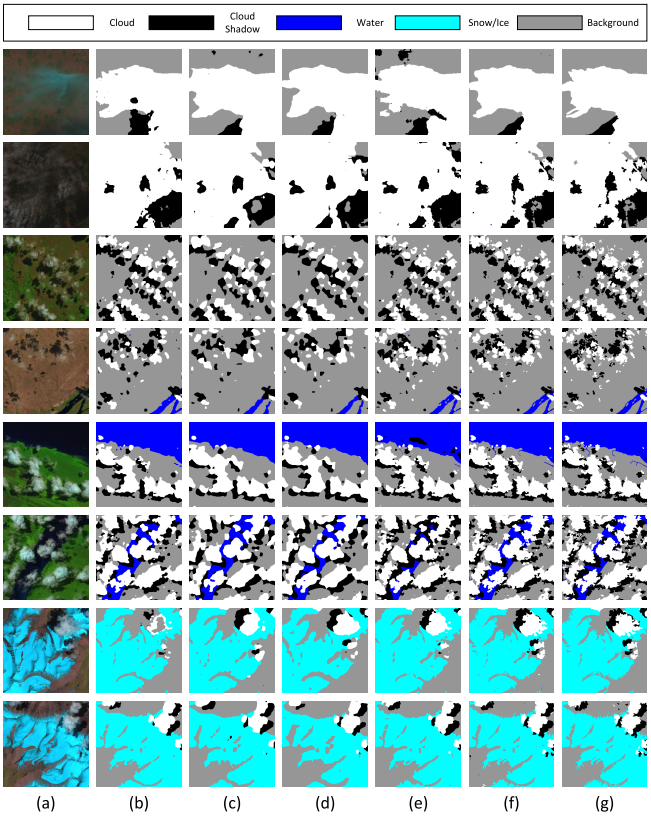

Fig. 11. Comparison of different models on the SPARCS Dataset. (a) Test image. (b) Segmentation results of PAN. (c) Segmentation results of BiSeNet V2. (d) Segmentation results of PSPNet. (e) Segmentation results of DeepLab V3Plus. (f) Segmentation results of our method. (g) Label.

method outperformed other models on PA, recall, precision, F_1, and MIoU. This shows that the dual-branch structure composed of transformer and convolution was effective in feature extraction. At the same time, because of the MGM, the two branches could guide each other to achieve efficient extraction of semantic and spatial information.

Fig. 11 shows the segmentation results of different methods in different scenarios of this dataset. Among them, the first and second sets were the segmentation results of thin clouds, and the third and fourth sets were the segmentation results of small-scale clouds and cloud shadows. Thin clouds and small clouds/cloud shadows were the difficult parts to detect in the cloud and cloud shadow segmentation. It could be seen from Fig. 11 that the segmentation results of PAN, BiSeNetV2, and PSPNet were not good, there were large-scale false detections, and the ability to detect small targets was insufficient. The segmentation results of DeepLab V3Plus were relatively good, but there was still a certain range of false detections. Furthermore, its segmentation boundaries were rough and lacked details. The proposed method greatly reduced the occurrence of false detections, because the dual-branch structure could efficiently and accurately extract multiscale information of images. Also, according to the different characteristics of transformer and convolution, the different features extracted by them were used for decoding, so that the segmentation boundary was clearer. The fifth and sixth sets mainly show the segmentation results of clouds, cloud shadows, and waters by different models. The seventh and eighth sets mainly show the segmentation results of different models for clouds, cloud shadows, and snow/ice. Detecting water and snow areas was prone to false detections, which were challenging tasks in cloud and cloud shadow segmentation. Due to the existence of the MGM, our proposed method can not only extract the spatial information and semantic information of the image through the dual-branch structure, but also realize the mutual guidance between the two branches, so that the model could more effectively extract the multiscale information. Therefore, the accurate extraction of water and snow information was achieved, and the occurrence of missed detection and false detection was reduced. In addition, the IFD could repair the lost boundary information during the decoding process, making the segmentation boundary clearer. As shown in Fig. 11, in these challenging tasks, our proposed method greatly reduced the probability of false detection and missed detection compared with other methods and, at the same time, produced a clear segmentation boundary.

IV. CONCLUSION

This article proposed a dual-branch network to realize end-to-end cloud and cloud shadow segmentation in visible and multispectral high-resolution remote sensing images. This method used transformer branch and convolution branch to extract high-level semantic information and spatial position information of an image, respectively. The MGM was used to enable branches with different characteristics to guide each other for feature extraction. In the decoding stage, through

the interleaved upsampling of the two branch features, accurate classification information was retained, and the rough segmentation boundary was repaired. Experiments show that compared with other methods, the proposed method significantly improved the accuracy of the model and could deal with various complex scenarios. However, this method still has room for improvement. We will reduce the number of parameters to improve the inference speed of the model while ensuring the segmentation accuracy.

REFERENCES

[1] S. S. Shen and M. R. Descour, "Comparative analysis of hyperspectral adaptive matched filter detectors," *Proc. SPIE*, vol. 4049, Aug. 2000, Art. no. 410332.

[2] X. Liu, J. Xu, and B. Du, "A BI-channel dynamic thershold algorithm used in automatically identifying clouds on GMS-5 imagery," *J. Appl. Metrol. Sci.*, vol. 16, no. 4, pp. 134–444, 2005.

[3] F. Ma, Q. Zhang, N. Guo, and J. Zhang, "The study of cloud detection with multi-channel data of satellite," *Chin. J. Atmos. Sci.-Chin. Ed.*, vol. 31, no. 1, p. 119, 2007.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[6] 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.

[7] 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.

[8] 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.* Cham, Switzerland: Springer, 2015.

[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2016.

[10] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn. (ICML)*, Jun. 2001, pp. 282–289.

[11] 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.

[12] Y. Zhan, J. Wang, J. Shi, G. Cheng, L. Yao, and W. Sun, "Distinguishing cloud and snow in satellite images via deep convolutional network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1785–1789, Oct. 2017.

[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. (IGARSS)*, Jul. 2019, pp. 1029–1032.

[14] J. Li, P. Zhao, W. Fang, and S. Song, "Cloud detection of multi-angle remote sensing image based on deep learning," *J. Atmos. Environ. Opt.*, vol. 15, no. 5, p. 380, 2020.

[15] 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, Oct. 2016.

[16] Z. Yan *et al.*, "Cloud and cloud shadow detection using multilevel feature fused segmentation network," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 10, pp. 1600–1604, Oct. 2018.

[17] H. Guo, H. Bai, and W. Qin, "ClouDet: A dilated separable CNN-based cloud detection framework for remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 9743–9755, 2021.

[18] S. Zhenfeng *et al.*, "Cloud detection in remote sensing images based on multiscale features-convolutional neural network," *IEEE Trans. Geosci. Remote Sens.* vol. 57, no. 6, pp. 4062–4076, Jun. 2019.

[19] N. Carion *et al.*, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020.

[20] N. Engel, V. Belagiannis, and K. Dietmayer, "Point transformer," *IEEE Access*, vol. 9, pp. 134826–134840, 2021.

[21] H. Chen *et al.*, "Pre-trained image processing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12294–12305.

[22] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[23] W. Wang *et al.*, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," 2021, *arXiv:2102.12122*.

[24] H. Wu *et al.*, "CvT: Introducing convolutions to vision transformers," 2021, *arXiv:2103.15808*.

[25] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*.

[26] 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.

[27] 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.

[28] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "DETR3D: 3D object detection from multi-view images via 3D-to-2D queries," in *Proc. 5th Conf. Robot Learn.*, 2022, pp. 180–191.

[29] R. Liu, Z. Yuan, T. Liu, and Z. Xiong, "End-to-end lane shape prediction with transformers," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3694–3702.

[30] 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, Dec. 2021, Art. no. 104940.

[31] M. Xia, Y. Qu, and H. Lin, "PADANet: Parallel asymmetric double attention network for clouds and its shadow detection," *J. Appl. Remote Sens.*, vol. 15, no. 4, 2021, Art. no. 046512.

[32] 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. (CVPR)*, Jun. 2019, pp. 5693–5703.

[33] Q. Hou, L. Zhang, M.-M. Cheng, and J. Feng, "Strip pooling: Rethinking spatial pooling for scene parsing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4003–4012.

[34] 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. (ECCV)*, 2018, pp. 325–341.

[35] 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," 2020, *arXiv:2004.02147*.

[36] 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.

[37] M. J. Hughes and D. J. 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.

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

[39] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2017, pp. 1–4.

[40] H. Park, L. Lowe Sjösund, Y. Yoo, J. Bang, and N. Kwak, "ExtremeC3Net: Extreme lightweight portrait segmentation networks using advanced C3-modules," 2019, *arXiv:1908.03093*.

[41] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[42] 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.

[43] M. Xia, T. Wang, Y. Zhang, J. Liu, and Y. Xu, "Cloud/shadow segmentation based on global attention feature fusion residual network for remote sensing imagery," *Int. J. Remote Sens.*, vol. 42, no. 6, pp. 2022–2045, Mar. 2021.