HyCloudX: A Multibranch Hybrid Segmentation Network With Band Fusion for Cloud/Shadow

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

***Abstract*—Semantic segmentation of cloud and shadow is an important task in remote sensing and atmospheric science. However, the complexity of cloud/shadow shapes, and noise disturbances (such as snow and ice, buildings, complex backgrounds, and atmospheric optics) make this task challenging. The traditional deep network has good details and generalization due to its local feature extraction ability and spatial invariance, but it is relatively weak in dealing with global context information, which leads to misjudgment and missed judgment in complex scenes. The transformer can effectively capture long-distance dependencies through the self-attention mechanism, but it may have challenges in extracting local image features and maintaining spatial consistency, resulting in loss of detail information and insufficient generalization. This article proposes a hybrid branch semantic segmentation network composed of convolutional network and transformer in parallel. A series of modules are designed to solve the problems of lack of multiscale feature extraction and insufficient fusion in some of the convolution–transformer hybrid networks. In particular, the network utilizes the rich information in auxiliary bands, such as near-infrared to improve the segmentation performance, so that the network can process a wider range of data and improve generalization. Experimental results on CloudSEN-12, 38-Cloud, and SPRCS-Val show that our network outperforms existing methods. After introducing the band fusion branch (HyCloudX), the network improves the segmentation performance and generalization, especially in the case of complex noise interference.**

***Index Terms*—Band fusion, cloud and cloud-shadow, deep learning, hybrid structure, image segmentation, multibranch.**

I. Introduction

CLOUD and cloud shadow detection is a crucial problem in remote sensing image processing. As an important part of meteorology, cloud changes are of key significance for analyzing climate change, forecasting, and studying disastrous weather. Many applications based on remote sensing technology, such as land cover classification [1], change detection [2], and water area segmentation [3], must also overcome the influence of cloud cover to ensure the accuracy and reliability of detection. Therefore, it is necessary to accurately identify clouds and cloud shadows to ensure the accuracy and reliability of remote sensing technology applications.

Before the introduction of deep learning, traditional methods based on machine learning are mainly used. Mainly includes: based on image processing methods [4], [5], using texture features, edge detection, and other computer vision technology to extract features, and then use the classifier to classify pixels; based on the image segmentation method [6], the cloud and cloud shadow are divided into different regions by using the pixel-based segmentation method, and then each region is classified. Based on the pixel classification method [7], semantic segmentation is achieved by using machine learning algorithms such as support vector machines to classify each pixel. Although machine learning-based methods may also achieve good results when dealing with simple cloud and cloud shadow images, due to the large changes in the shape and texture of clouds and cloud shadows, machine learning-based methods often perform poorly in more complex scenes [8].

In 2012, Krizhevsky et al. [9] proposed AlexNet, which proved the potential of deep convolution networks in the field of computer vision and promoted the development of deep learning in the fields of semantic segmentation and object detection. In 2014, Long et al. [10] proposed a fully convolutional network (FCN) for semantic segmentation. This network uses a convolution layer instead of a fully connected layer to achieve end-to-end pixel-level prediction of input of any size. However, this network has an obvious problem, due to the substitution of pooling layer and convolution layer, the output resolution is low, and the context information of the image is not considered. To solve the problem of FCN, in 2015, Ronneberger et al. [11] proposed U-Net, which adopts the structure of encoder and decoder. The encoder can transform the image into a feature representation, and the decoder can map the features back to the pixel-level prediction results. Since then, a series of codec-based networks have emerged, such as SegNet [12] and ENet [13]. In order to make better use of the context information of images, in 2016, Zhao et al. [14] proposed the spatial pyramid pooling network (PSPNet) to solve the problem of context information in semantic segmentation. Subsequently, DeepLab series models introduced dilated convolution and atrous spatial pyramid pooling (ASPP) modules [15], [16] to improve the accuracy of segmentation results. In 2018, Li et al. [17] proposed the pyramid attention network (PAN), which introduces an attention mechanism to increase the importance of regions of interest. With the introduction of transformer [18], it has been successfully tried in computer vision tasks [19], [20]. This model, which was

Manuscript received 27 January 2024; revised 19 February 2024; accepted 1 March 2024. Date of publication 7 March 2024; date of current version 21 March 2024. This work was supported by the National Natural Science Foundation of PR China under Grant 42075130. *(Corresponding author: Liguo Weng.)*

Ziwei Hu, 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/JSTARS.2024.3374233

originally used for natural language processing tasks, brings new possibilities to DCNN, which was originally difficult to break through in the field of computer vision, because of its advantages of long-range dependence and location independence. To introduce transformer into dense prediction tasks, such as object detection and semantic segmentation, Wang et al. [21] proposed pyramid vision transformer (PVT), which uses pure transformer as the backbone and introduces pyramid structure into transformer, reducing feature maps and thus reducing computational overhead. Swin transformer introduces windowed self-attention mechanism and cascade structure, which can greatly reduce the computational complexity, effectively alleviate the problem of gradient disappearance, and improve the performance of the model [22]. Some cutting-edge research has begun to explore strategies for combining these two network structures. These studies try to use the powerful local feature extraction ability of deep convolution networks and the excellent global context awareness of transformer to achieve better performance on various tasks. Convolutional vision transformer (CvT) proposed by Wu et al. [23] applied convolution to vision transformer to improve spatial information. Lee et al. [24] proposed multipath vision transformer (MPViT), explored multiscale path embedding and multipath structure, and introduced convolution branches to extract local feature information.

With the development of deep learning in the field of computer vision, networks used for semantic segmentation of remote sensing images of clouds and cloud shadows have also made significant progress. Guo et al. [25] proposed a lightweight fully convolutional neural network (ClouDet), which uses atrous separable convolution to improve the efficiency and accuracy of the network, and introduces multiscale feature fusion to deal with cloud shadows of different scales. Yan et al. [26] used the pyramid pooling module to extract context information. They proposed a multilevel feature fusion structure that combines semantic information with spatial information from different levels. There are also explorations to do more efficient convolutional features extraction of cloud and cloud shadow features by strip convolution [27], multisupervised feature fusion attention [28], multiscale feature extraction [29], [30], and other methods [31], [32], [33]. Through the introduction of adversarial learning, efforts are made to bridge the significant differences in the representation of remote sensing images between different cities to enhance generalization [34]. Hong et al. [35] applied the generative pretrained transformer structure to classification and segmentation tasks of remote sensing images, enabling progressive training to handle inputs of different sizes, resolutions, time series, and regions, thus making full use of extensive remote sensing big data. DBNet proposed by Lu et al. [36] decodes by combining different features extracted by transformer branch and convolution branch, and repairs the rough segmentation boundary in the decoding part. Wang et al. [37] proposed the multihead feature extraction module to strengthen the recognition ability of the target boundary and effectively avoid interclass ambiguity.

Nevertheless, existing networks still have limited ability in the semantic segmentation task of cloud and cloud shadow, especially since the cloud and cloud shadow segmentation boundaries are rough. Under the interference of surface objects, noise and other factors, false detection and missed detection are easy to occur. Based on deep convolutional neural network (DCNN) learning methods, which excel at extracting local image feature information and possess characteristics, such as parameter sharing and translational invariance [38], the network's parameter count and computational load can be significantly reduced, thereby enhancing its robustness and generalization performance. However, DCNNs struggle to capture long-range dependencies. Although this issue can be alleviated by enlarging the receptive field, it still fails to capture global features. Transformer, as a sequence modeling approach, although capable of handling long-range dependencies, cannot fully exploit spatial relationships between pixels, which traditional CNN structures can handle better. Existing convolutional–transformer hybrid network structures offer overly simplistic and rudimentary fusion of features extracted by these two structures, failing to fully utilize this feature information, resulting in no significant improvement in performance in cloud and cloud shadow semantic segmentation tasks. Furthermore, existing networks suffer a significant drop in performance in some complex scenarios, such as cloud and cloud shadow occlusions, interference from surface objects (e.g., ice and snow), variations in cloud, and cloud shadow reflectance due to lighting conditions. This is because they rely solely on visible light channels, making it difficult to differentiate boundaries and regions between clouds and cloud shadows with very similar reflectance within the visible spectrum. They are also highly sensitive to variations in cloud and cloud shadow reflectance. These common factors collectively affect the accuracy of segmentation. These networks exclusively use visible light channels for image segmentation, ignoring the potential value of other spectral bands. For instance, compared to the visible light spectrum, the infrared spectrum offers higher penetration capabilities and better discrimination, providing rich additional information, such as cloud thickness. Infrared bands are less affected by factors, such as lighting, shadows, complex atmospheric optical noise, and can provide more stable data. These capabilities can significantly aid in the semantic segmentation of remote sensing images of clouds and cloud shadows.

The multibranch network proposed in this article is composed of convolution and transformer and adopts the encoder–decoder structure. In the encoder, we use ResNet-50 and lightweight swin transformer as the backbone. On the one hand, convolution structure has transform scaling and distortion invariance, and can effectively extract local and low-level information, which is lacking in transformer structure. On the other hand, transformer has dynamic attention, global context and better generalization ability, which can effectively extract global and high-level semantic information [18], which is weak in convolution structure. The proposed method effectively combines the advantages of the two branches, so that the two branches can exert their respective advantages and help the model extract features more effectively. Studies have shown that for semantic segmentation tasks, it is significant to have a large receptive field [10] and extract multiscale feature information [39]. Therefore, in terms of feature fusion, we design cross feature fusion (CFF) module to fuse the global and local features extracted by the transformer branch and ResNet branch of each stage to guide each other

for feature mining. In order to expand the receptive field and better extract multiscale context information, we design cross layer fusion (CLF) module. In the decoder, we design cross hierarchy fusion (CHF) module to make full use of the different levels of features extracted by the two branches to guide the decoding, so as to effectively fuse the semantic information and spatial location information, so that the position of cloud and cloud shadow is more accurate, and the segmentation boundary contains more details. In particular, we design cross band fusion (CBF) in a branching form, which integrates multiple auxiliary spectral information into the visible light channel through the attention mechanisms, fully leveraging the abundant information provided by the auxiliary spectrum. We name the network based on the dual-branch structure with CFF, CLF, and CHF modules as HyCloud. On this basis, the multibranch network with auxiliary band branch and band fusion module CBF is named HyCloudX.

Our work has made the following contributions.

1) By adopting a convolution–transformer hybrid structure network, we successfully integrated the advantages of both convolution and transformer architectures. Through the synergistic utilization of convolution's local detailed features and transformer's global semantic information, we enhanced the semantic segmentation performance for clouds, cloud shadows, and backgrounds. This approach resulted in more precise and detailed segmentation outcomes.
2) It is different from the simple processing of the dual-branch feature in the same type of network with mixed structure, in the coding stage, we design CFF module and CLF module to fuse the multiscale feature information of each branch, which improves the performance of extracting semantic information and spatial information at different scales. In the decoding phase, the CHF module is designed to make full use of features extracted from multiple branches for upsampling, gradually guiding the restoration of the feature maps. By extensively leveraging features extracted from both branches, we achieve a more effective extraction of semantic and spatial information at different scales, resulting in clearer and more accurate segmentation boundaries.
3) We exploratively attempted to introduce auxiliary bands to assist in guiding semantic segmentation within the visible light channel. Through experiments, we demonstrated the feasibility of this approach. Furthermore, we observed that employing the CBF module, which separately processes and fuses bands for multimodal band extraction, leads to a more significant improvement in segmentation performance compared to directly concatenating these bands as a unified input into the network. This improvement is particularly notable in the precision of boundary delineation, thereby enhancing the overall segmentation performance and generalization of the network. This exploration suggests that incorporating additional auxiliary bands to complement the information in the visible light channel and effectively fusing features from these multimodal bands can greatly enhance the final segmentation performance. This has profound implications for cloud detection-related work and various downstream applications in atmospheric science, opening up new possibilities for segmentation models in atmospheric science applications.

II. Methodology

This article proposes a multibranch architecture network composed of convolution and transformer, which can effectively identify clouds and cloud shadows and generate clear and accurate segmentation boundaries. The overall architecture of the network is shown in Fig. 1, which is divided into two parts: encoder and decoder. We first use a dual-branch structure composed of convolution and transformer to extract features at different levels. Some previous studies [10], [11] simply combined high-level features and low-level features. This fusion is too simple, and there are still problems, such as false detection and rough segmentation edges, resulting in unsatisfactory segmentation results. The proposed network can combine the advantages of convolution network and transformer, effectively fuse local features and global features, and perform multiscale feature extraction and fusion. The proposed method also introduces and integrates infrared band information to assist the network in identifying clouds and cloud shadows. In the decoding stage, for the coarse segmentation boundary caused by the loss of high-level semantic information and spatial detail information after upsampling, we use the multiscale features in the encoder to fuse, and realize the precise positioning and fine segmentation of cloud and cloud shadow.

A. Backbone

After experimental comparisons and considering tradeoffs between segmentation performance and computation complexity (refer to Section III-C), we use ResNet-50 as the convolution branch. Convolution structure inherently possess advantages, such as local perception and translation invariance. The introduction of residual connections by ResNet addresses the problem of gradient vanishing, enabling the construction of deeper networks and thus improving performance [40]. For the transformer branch, we employed modules from the swin transformer. Swin transformer introduces window-based self-attention mechanisms and a cascading structure, which can significantly reduce computational complexity and improve model performance. Transformer structure have the advantage of capturing long-range dependencies through dynamic self-attention mechanisms. The modules of the two structures of convolution and transformer form a dual-branch backbone as the basis of the network, which can form complementary advantages to obtain different levels of features and enhance the feature extraction ability of the network in the process of feature extraction. Table I gives the specific parameters of the backbone, where win.size denotes the size of shifted window in swin transformer.

Transformer branch first partitions the input image into multiple small patches, each considered as a token. These tokens are then fed into the swin transformer block for processing. Each swin transformer block consists of two consecutive blocks. The local self-attention mechanism is applied to these tokens through the shifted window mechanism within each stage, and then these local features are integrated through cross-window connections

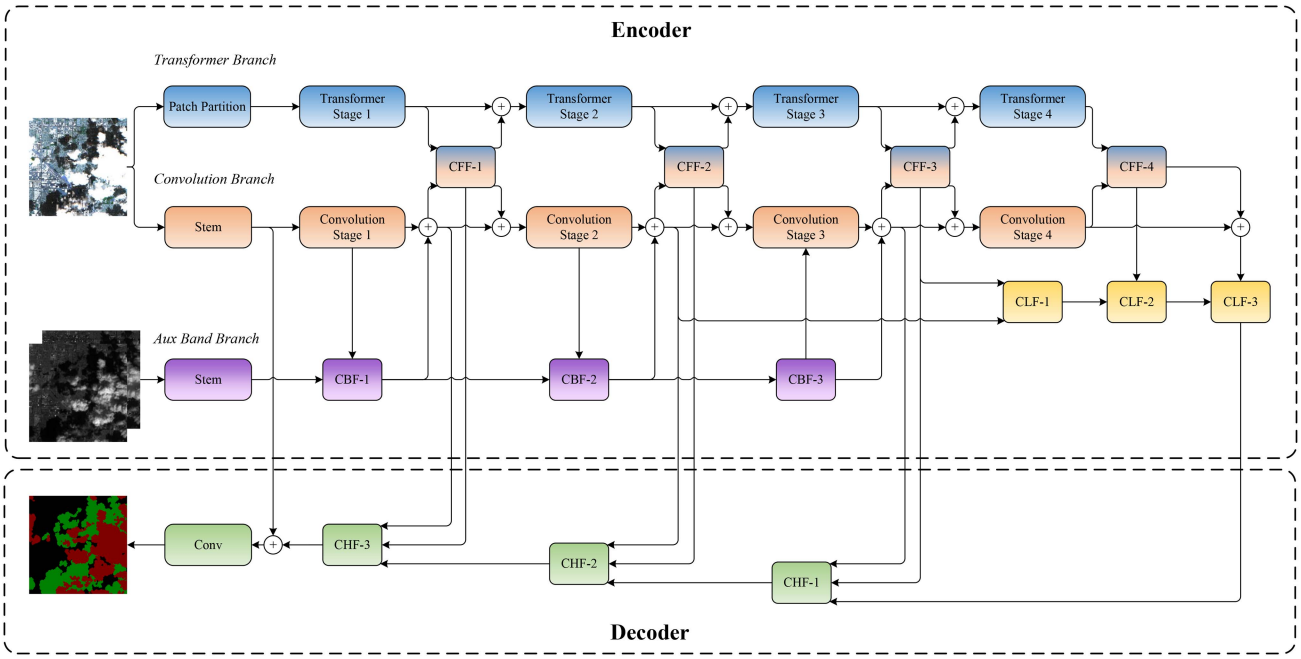

Fig. 1. Network structure of HyCloudX. CFF represents cross feature fusion module, CLF represents cross layer fusion module, CHF represents cross hierarchy fusion module, CBF represents cross band fusion module. HyCloud excludes the auxiliary band branch. $\oplus$ is the addition operation and Conv is the convolution operation.

TABLE I
BACKBONE STRUCTURE OF PROPOSED NETWORK

Stage	Conv branch	Output size	Transformer branch	Output size
0	$\begin{bmatrix} 7\times7, 64 \\ 3\times3 \text{ maxpool, stride2} \end{bmatrix}$	[64,112,112]	4×4, 64, stride 4	[64,56,56]
1	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}$	[256,112,112]	concat 4×4, 64 [win.size 7×7, dim 64, head 2] $\times$ 1	[64,56,56]
2	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}$	[512,56,56]	concat 2×2, 128 [win.size 7×7, dim 128, head 4] $\times$ 1	[128,28,28]
3	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}$	[1024,28,28]	concat 2×2, 256 [win.size 7×7, dim 256, head 8] $\times$ 3	[256,14,14]
4	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}$	[2048,14,14]	concat 2×2, 512 [win.size 7×7, dim 512, head 16] $\times$ 1	[512,7,7]

and interlayer connections to form a global view. The expression of swin transformer block is as follows:

$$\hat{z}^l = W - \text{MSA}\left(\text{LN}\left(z^{l-1}\right)\right) + z^{l-1} \quad (1)$$

$$z^l = \text{MLP}\left(\text{LN}\left(\hat{z}^l\right)\right) + \hat{z}^l \quad (2)$$

$$\hat{z}^{l+1} = \text{SW} - \text{MSA}\left(\text{LN}\left(z^l\right)\right) + z^l \quad (3)$$

$$z^{l+1} = \text{MLP}\left(\text{LN}\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l+1} \quad (4)$$

where $\hat{z}^l$ and z^l denote the output features of the (S)WMSA module and the MLP module for block l, respectively. W-MSA and SW-MSA denote window-based multihead self-attention using regular and shifted window partitioning configurations, respectively. LN represents lay normalization. Compared with pure transformer, the swin transformer module optimizes the balance between local self-attention and global self-attention, greatly reduces the amount of calculation and exhibits better performance.

B. Cross Feature Fusion

The shape of cloud and cloud shadow is similar, which often leads to misjudgment in the detection process. In addition, the interference of ground objects (such as waters, ice, and snow, and other objects with similar attributes to clouds and cloud shadows) and noise can also cause misclassification. In order to obtain more accurate results, we use CFF (see Fig. 2) to

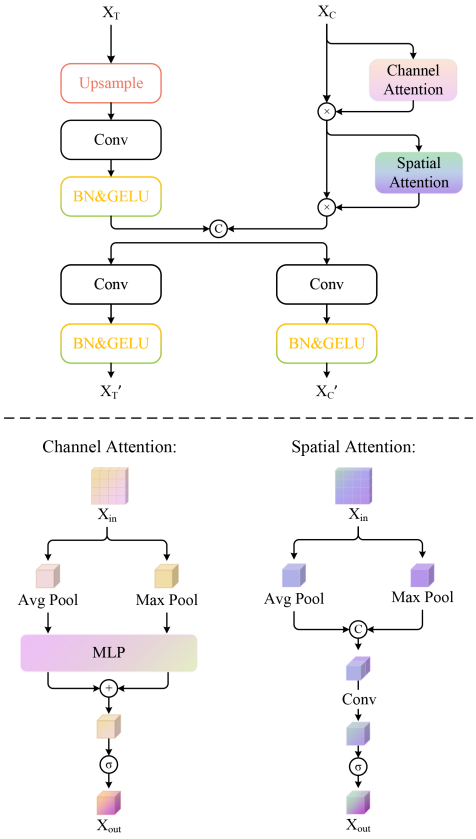

Fig. 2. Structure of CFF. Conv represents the convolution operation. BN represents batch normalization and GELU represents activation function GELU. σ is the sigmoid activation function. ©denotes concat operation.

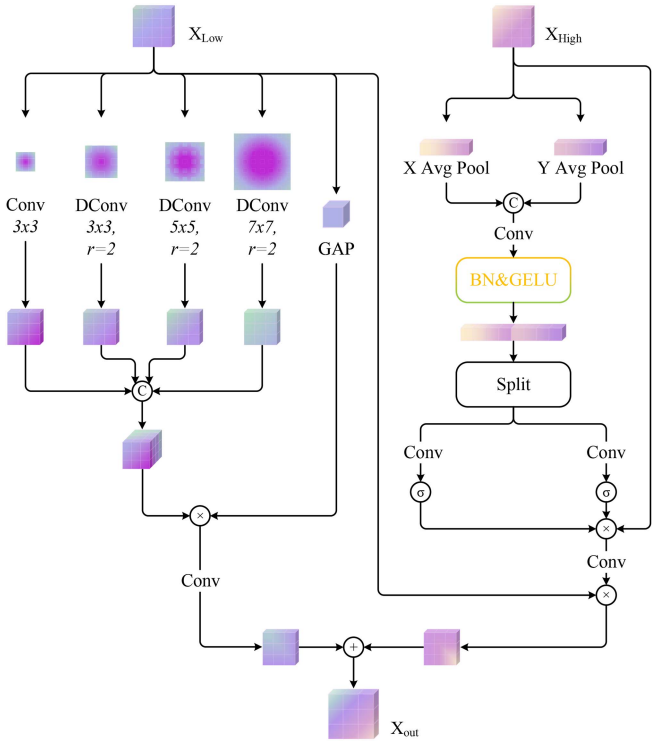

Fig. 3. Structure of CLF. Conv represents the convolution operation. DConv represents the atrous convolution operation, and r is the atrous rate. BN represents batch normalization, and GELU represents the activation function GELU. σ is the sigmoid activation function. ©denotes concat operation.

combine the low-level features extracted by the convolution branch with the high-level features extracted by the transformer branch. Compared with the high-level features, the low-level features extracted by convolution branches retain more spatial information. Therefore, the convolution branch can provide location information guidance for the deep semantic feature mining of the transformer branch. The module consists of two parallel branches, one of which applies attention to the features extracted by the convolution branch on the channel and space, respectively. By adaptively adjusting the weights on the channel and space, the network can better focus on the important features and regions in the input data and improve the expression ability of the convolution branch. Specifically, channel attention is to generate a channel weight vector by averaging and max pooling the spatial dimension of the feature map, and then sending it to a shared multilayer perceptron. Spatial attention performs average pooling and maximum pooling on the channel dimension of the feature map, then concatenates the processed features, and then uses a convolution layer to generate a spatial weight matrix. Multiply these weights with the original input feature map to achieve channel and spatial attention. The other branch uses bilinear interpolation to upsample the high-level feature to the same size of convolution branch at same stage, then concat the two features, and then adjusts the concatenated features to the same size as the input of each branch through 1×1 convolution, and feeds them to the subsequent network.

It can be seen from Fig. 7 that CFF can well fuse the feature information of convolution branch and transformer branch. Specifically, compared with Fig. 7(b)–(d), it can be seen that only using the convolution branch or using the convolution branch and transformer branch to extract features and simple stacking and upsampling are performed at the final stage, the internal attention to clouds and cloud shadows is not enough. For example, by comparing the (b), (c), and (d) of sample (1), it can be found that the convolution branch pays more attention to the cloud than to the cloud shadow. After adding transformer branch, the attention to cloud shadows has been improved, but the area inside the cloud has been ignored. After adding CFF, the feature information of the two branches is fused and enhanced, and the attention to cloud and cloud shadow is significantly improved, which shows that CFF has a positive effect on the feature fusion of the dual branches.

C. Cross Layer Fusion

Feature fusion can integrate multiple features to improve classification accuracy. However, if the feature extraction is not comprehensive enough, the fusion result will also be limited. For the semantic segmentation task of cloud and cloud shadow, its shape, scale, intensity, and spatial changes are extremely complex. It is difficult to capture these subtle differences only through single-scale feature extraction. Therefore, it is necessary to introduce multiscale feature extraction and fusion. We use CLF (see Fig. 3) to capture and analyze the complex characteristics of clouds and cloud shadows from different perspectives

TABLE II
COMPARISON OF DIFFERENT ATROUS KERNEL SIZE AND RATE

Kernel size		Atrous rate			MIoU
3,3,3	3,5,7	2,2,2	2,4,8	6,12,18	
✓		✓			77.21
✓			✓		76.98
✓				✓	76.75
	✓	✓			**77.32**
	✓		✓		77.18
	✓			✓	76.95

The bold values represent the most beneficial configurations for improving segmentation performance after exploring different kernel sizes and atrous rates in the hole convolution part of the CLF module. These configurations were also the parameters ultimately adopted by our network.

and levels to obtain more comprehensive and in-depth feature information.

Inspired by PSPNet and DeepLab, we use an improved atrous pyramid for low-dimensional features that contain more spatial and detailed information. In PSPNet, the pyramid module (PPM) obtains global context features at different scales by performing maximum pooling operations at multiple different scales. However, some important local details may be ignored or lost due to the simple pooling operation, especially for the complex edge segmentation tasks, such as cloud and cloud shadow. It can be seen from Table VI that the performance of the network is reduced after using the PPM module. DeepLab V3 has improved this by introducing hole convolution so that the convolution kernel can effectively obtain context information at different scales by adjusting the size of the hole without increasing the parameters. From Table II, we can see that the use of ASPP has improved the segmentation performance compared to PPM, but the increase is limited. This is because the ASPP in DeepLab V3 uses the 3×3 dilated convolution, and the atrous rate is set to (6,12,18). Simply increasing the atrous rate of the dilated convolution without changing the size of the convolution kernel may produce the so-called "grid effect." This is because although increasing the atrous ratio allows the convolution kernel to have a larger receptive field, in fact, the effective pixels perceived by the convolution kernel are not increased, which makes the model unable to accurately capture the local characteristics of the image. In order to solve this problem, we adjust the size of the convolution kernel and the void rate at the same time to explore the parameters that are most suitable for cloud and cloud shadow segmentation tasks. It can be seen from Table II that the most suitable convolution kernel size is $3 \times 3, 5 \times 5, 7 \times 7$, and the atrous rate is 2.

The results after each dilated convolution kernel are concatenated and multiplied by the features extracted by global pooling to complete the multiscale information extraction of low-dimensional features. For high-dimensional features, due to its multiple convolutions and attention operations, it has rich global semantic information, but loses some spatial context information. We restore it by introducing a coordinate attention mechanism. Then, we fuse the processed low-dimensional features and high-dimensional features and feed them to the back

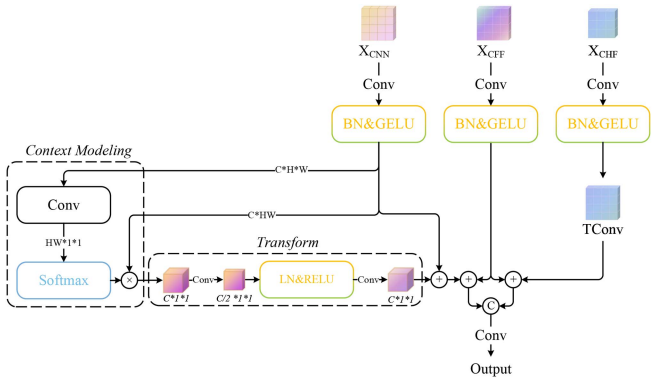

Fig. 4. Structure of CHF. Conv represents the convolution operation, and TConv represents the transposed convolution operation. BN represents batch normalization and LN represents layer normalization. GELU represents the activation function GELU. σ is the sigmoid activation function. ©denotes concat operation.

network. By comparing Fig. 7(d) and (e), it can be seen that after adding the CLF structure, the network pays more attention to the inside of the cloud and cloud shadow. This shows that CLF can effectively extract multiscale feature information and fuse low-dimensional and high-dimensional features.

D. Cross Hierarchy Fusion

The size and shape of clouds and cloud shadows are variable, which makes it very difficult to detect their boundaries. In the existing methods, the segmentation boundary of the feature map after multiple downsamplings and upsamplings is very rough and the detailed information is insufficient. The network proposed in this article uses CHF (see Fig. 4) in the decoding progress to gradually recover the segmentation boundary information.

The CHF utilizes encoder outputs, features from CFF modules by using skip connections and features from the convolution branch of certain stages to guide the model to repair the details of the segmented boundary, reducing the problem of serious loss of boundary details after deep downsampling. For the features from the convolution branch, CHF first uses context modeling to obtain the global information relationship vector, and then uses two layers of 1×1 convolution to further extract the information. The obtained attention map is applied to all the original feature maps to capture the global context information and integrate it into the original feature map. For features from the encoder output, or from the previous CHF module, transposed convolution is used to upsample to the same size as the other two input features. Then, the three feature inputs are fused and concatenated, respectively, and then 1×1 convolution is used to enhance the feature expression ability and feed it to the subsequent network. The calculation formula is as follows:

$$X_{\mathrm{CNN}}{}' = \sigma\{\mathrm{BN}\left[\mathrm{Conv}\left(X_{\mathrm{CNN}}\right)\right]\} \tag{5}$$

$$X_{\mathrm{CFF}}{}' = \sigma\{\mathrm{BN}\left[\mathrm{Conv}\left(X_{\mathrm{CFF}}\right)\right]\} \tag{6}$$

$$X_{\mathrm{CHF}}{}' = T\mathrm{Conv}\left(\sigma\{\mathrm{BN}\left[\mathrm{Conv}\left(X_{\mathrm{CHF}}\right)\right]\}\right) \tag{7}$$

$$\mathrm{Transform}(X) = \mathrm{Conv}(\delta\{\mathrm{LN}\left[\mathrm{Conv}(X)\right]\}) \tag{8}$$

$$\mathrm{CM}(X) = \odot[\mathrm{Conv}(X_{\mathrm{CNN}}{}')] \times X_{\mathrm{CNN}}{}' \tag{9}$$

$$\widehat{X_{\mathrm{CNN}}} = \mathrm{Transform}[\mathrm{CM}(X_{\mathrm{CNN}}{}')] + X_{\mathrm{CNN}}{}' \tag{10}$$

$$F_{\mathrm{CHF}} = \mathrm{Conv}\left[\mathrm{Concat}\left(\widehat{X_{\mathrm{CNN}}} + X_{\mathrm{CFF}}{}', X_{\mathrm{CFF}}{}' + X_{\mathrm{CHF}}{}'\right)\right] \tag{11}$$

where Conv (·) represents convolution operation, TConv (·) represents transposed convolution operation, BN (·) represents batch normalization operation, LN (·) represents layer normalization operation, σ (·) represents GELU activation function, δ (·) represents RELU activation function, $\odot$ represents softmax operation.

By comparing Fig. 7(e) and (f), it can be seen that the use of CHF further strengthens the model's attention to the segmentation edges and details, such as small area clouds and cloud shadows on the previous basis. This shows that CHF improves the network's ability to perceive edge details. By channel fusion of features, it can provide more accurate context information, help the model better locate and restore edges and other details, thereby improving the accuracy of segmentation results and detail restoration ability.

E. Cross Band Fusion

In the early stages of exploring the integration of auxiliary bands into the network, we attempted to directly overlay data from the visible light bands and auxiliary bands and input them into HyCloud for unified processing. As indicated by the ※ entry in Table VI, the experiments showed improvement compared to inputting only the visible light bands, but there was a certain gap from the anticipated effects. We attribute this to the fact that visible light bands and auxiliary bands typically contain different types of information. The visible light bands are more sensitive to details, such as color and shape, while auxiliary bands, such as the infrared band, provide information related to temperature, surface reflectance, and other factors. Therefore, we designed the CBF module, as illustrated in Fig. 5. By separately processing them, the network can focus more attentively on learning specific features from different channels. Subsequently, through the attention mechanism within the module, the network adaptively fuses information from these two types of bands. Ablation experiments demonstrated a further enhancement in the network's performance with the introduction of the CBF module. It also confirmed that separately processing the visible light and auxiliary bands, as opposed to a unified approach, has performance advantages.

Specifically, the CBF module processes the RGB channels and infrared auxiliary bands through convolution operations to form two key matrices W_q and W_k. The product of these two matrices generates a weight matrix. Then, the RGB channels and the infrared auxiliary bands are concatenated together to form W_v, which is equivalent to fusing the information of two different bands. Then, the weight matrix after softmax is multiplied by W_q to realize the weighted mixing of RGB channels and infrared auxiliary bands. Finally, the obtained results are residually connected with the original RGB channel to retain the original information of the RGB channel while fusing the

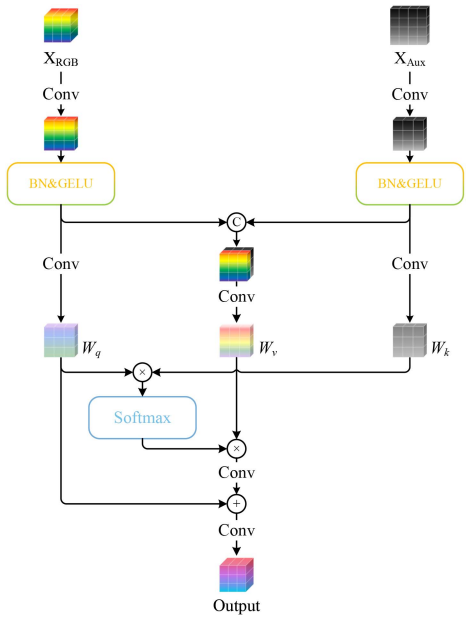

Fig. 5. Structure of CBF. Conv represents the convolution operation. BN represents batch normalization, and GELU represents the activation function GELU. ©denotes concat operation.

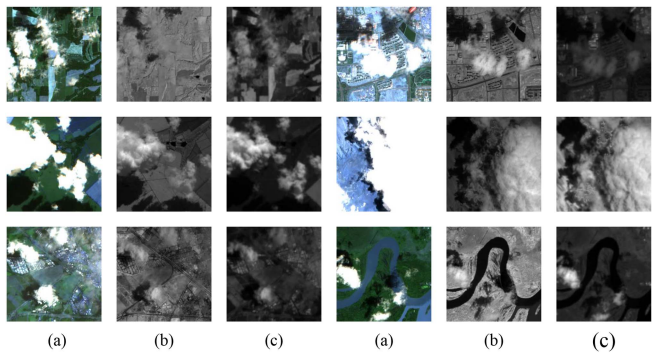

Fig. 6. Presentation of RGB channels and auxiliary bands of CloudSEN-12. (a) RGB channels. (b) B8 NIR band. (c) B11 SWIR band.

infrared information. The calculation formula is as follows:

$$X_{\mathrm{RGB}}{}' = \sigma\{\mathrm{BN}\left[\mathrm{Conv}\left(X_{\mathrm{RGB}}\right)\right]\} \tag{12}$$

$$X_{\mathrm{Aux}}{}' = \sigma\{\mathrm{BN}\left[\mathrm{Conv}\left(X_{\mathrm{Aux}}\right)\right]\} \tag{13}$$

$$W_q = \mathrm{Conv}\left(X_{\mathrm{RGB}}{}'\right) \tag{14}$$

$$W_k = \mathrm{Conv}\left(X_{\mathrm{Aux}}{}'\right) \tag{15}$$

$$W_v = \mathrm{Conv}\left[\mathrm{Concat}\left(X_{\mathrm{RGB}}{}', X_{\mathrm{Aux}}{}'\right)\right] \tag{16}$$

$$F_{\mathrm{CBF}} = \mathrm{Conv}\{\mathrm{Conv}\left[\odot\left(W_q \times W_k\right) \times W_v\right] + W_q\} \tag{17}$$

where Conv (·) represents convolution operation, BN (·) represents batch normalization operation, and σ (·) represents GELU activation function. $\odot$ represents softmax operation.

By comparing Fig. 7(b) and (g), we can clearly see that after using the CBF module to integrate the infrared band information, the model's ability to identify and distinguish clouds and cloud shadows has been significantly improved. Specifically, this process strengthens the network's attention to cloud shadow

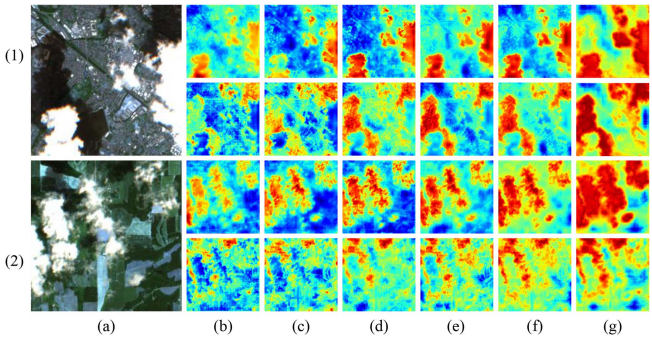

Fig. 7. Heat map representation. The first line of each sample is the attention to the cloud, and the second line is the attention to the cloud shadow. (a) Test image. (b) Convolution branch. (c) Convolution branch + transformer branch. (d) Convolution branch + transformer branch + CFF. (e) Convolution branch + transformer branch + CFF + CLF. (f) Convolution branch + transformer branch + CFF + CLF + CHF. (g) Convolution branch + transformer branch + CFF + CLF + CHF + CBF.

features, making the model more effective in capturing and processing important environmental information, and can have a deeper understanding of more subtle texture and shape differences, especially in complex meteorological conditions. This not only improves the network's ability to perceive local details, but also improves the generalization and robustness of the network. The design of CBF fully considers the complementarity of two different band information, which can effectively improve the expression ability of the network and provide a new possibility for high-quality processing multimodal image data.

III. Experiments

A. Datasets

1) CloudSEN-12: This is a large dataset for semantic segmentation of cloud and cloud shadow [41]. The dataset consists of 49 400 image patches (IPs), each IP covers 5090 × 5090 m, and is evenly distributed on all continents except Antarctica, covering multiple time points. The dataset contains data from Sentinel-2 1 C and 2 A levels, as well as manually labeled high-quality, scribble, and no-annotation three types of thick, thin clouds, and cloud shadow annotations. Table III shows the information of all channels and bands of the dataset.

We selected high-quality manually labeled data items with few, medium, cloudy and almost cloudless from the original dataset, a total of 9999 samples, and we selected 5633 samples. We choose the TCI_R, TCI_G, TCI_B of these samples as the visible light channels, stacked into a three-channel 24-bit image. We chose the B8 near-infrared (NIR) band and the B11 shortwave infrared (SWIR) band as the auxiliary band, because the wavelength of the B8 band covers the infrared band, and the clouds and cloud shadows have obvious reflectivity changes in the infrared band, so the B8 band can better detect clouds and cloud shadows. High clouds have obvious reflectivity changes in the wavelength range of about 1.6 microns, and the B11 band just covers this wavelength range, so the B11 band is suitable for detecting high clouds. Fig. 6 shows that these two bands can provide additional information different from the visible light band. We extract the data of these two bands and map them into 8-bit grayscale images, respectively. So far, we have a total of 5633 samples, each of which contains an RGB channel image and two grayscale images with a size of 509 × 509 pixels. Then we divided each image into nine subimages with 224 × 224 pixels, and obtained a total of 50 697 samples. We divided it into training set and validation set according to the ratio of 9:1, and finally obtained 45621 training sets and 5076 validation sets.

TABLE III
CloudSEN-12 Dataset

Band	Wavelength (μm)	Resolution (m)
B1 (Aerosols)	0.43–0.45	60
B2 (Blue)	0.44–0.51	10
B3 (Green)	0.52–0.59	10
B4 (Red)	0.64–0.67	10
B5 (Red Edge 1)	0.68–0.7	20
B6 (Red Edge 2)	0.73–0.74	20
B7 (Red Edge 3)	0.77–0.79	20
B8 (NIR)	0.78–0.9	10
B8A (Red Edge 4)	0.85–0.87	20
B9 (Water Vapor)	0.93–0.95	60
B10 (Cirrus)	1.36–1.38	60
B11 (SWIR1)	1.56–1.65	20
B12 (SWIR2)	2.1–2.28	20
AOT (aerosol optical thickness)	–	–
WVP (water vapor pressure)	–	–
TCI_R (true color image, red)	–	–
TCI_G (true color image, green)	–	–
TCI_B (true color image, blue)	–	–

2) 38-Cloud: The dataset is derived from 38 Landsat 8 scene images and their manually extracted pixel-level ground truths for cloud detection [42]. The dataset contains four channels: red, green, blue, and NIR. We cut the image into subimages with 224 × 224 pixels. After removing the subimages which contain 20% or more of the black edge region, the training set and the test set are divided according to the ratio of 9:1. Finally, we obtained 4639 datasets and 515 training sets.

3) SPARCS-Val: The dataset contains 80 Landsat 8 scenes with a size of 1000 × 1000 pixels and manually labeled mask labels. It includes seven categories: cloud, cloud shadow, cloud shadow over the water, water, ice and snow, land and flooded areas. The dataset contains all bands from the original Landsat Level-1 data product as shown in Table IV. We selected B1, B6, and B9 as auxiliary bands. We cut each image into subimages with a size of 224 × 224, and a total of 2000 images are obtained. In order to enhance the generalization ability of the model, we expanded the dataset by horizontal flipping, vertical flipping, and random rotation, and then divided the training set and test set according to the ratio of 9:1. Finally, we obtained 7200 training sets and 800 training sets.

B. Experiment Details

1) Basic Information: The experiment was performed on the NVIDIA RTX 4090 GPU using PyTorch. Because most of the networks in this experiment converge after 250 iterations, we

TABLE IV
SPARCS-VAL DATASET

Band	Wavelength (μm)	Resolution (m)
B1 (Coastal)	0.43–0.45	30
B2 (Blue)	0.45–0.51	30
B3 (Green)	0.53–0.59	30
B4 (Red)	0.64–0.67	30
B5 (NIR)	0.85–0.88	30
B6 (SWIR-1)	1.57–1.65	30
B7 (SWIR-2)	2.11–2.29	30
B8 (Pan)	0.5–0.68	15
B9 (Cirrus)	1.36–1.38	30
B10 (TIRS-1)	10.6–11.19	100

fixed the epoch number to 300 with the batch size of 16. We used cross entropy loss as the loss function and AdamW as the optimizer, and the weight attenuation coefficient is 0.0001. We used the poly LR strategy in training progress. The initial LR is set to 0.0001, and the poly power is set to 2. The learning rate (LR) of each round of training is described as follows, where epoch is the number of current iteration:

$$\mathrm{LR} = 0.0001 \times \left(1 - \frac{\mathrm{epoch}}{300}\right)^2. \tag{18}$$

2) Metrics: We choose precision (P), recall (R), $F1$ score, pixel accuracy (PA), average pixel accuracy (MPA), and average intersection–union ratio (MIoU) to evaluate the performance of the method in cloud and cloud shadow segmentation tasks. The calculation formula of each evaluation metric is as follows:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{19}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{20}$$

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{21}$$

$$\mathrm{PA} = \frac{\sum_{i=0}^{\mathrm{k}} \rho_{\mathrm{i},j}}{\sum_{i=0}^{k} \sum_{j=0}^{k} \rho_{i,j}} \tag{22}$$

$$\mathrm{MPA} = \frac{1}{k} \sum_{i=0}^{k} \frac{\rho_{i,j}}{\sum_{j=0}^{k} \rho_{i,j}} \tag{23}$$

$$\mathrm{MIoU} = \frac{1}{\mathrm{k}+1} \sum_{i=0}^{k} \frac{\rho_{i,j}}{\sum_{j=0}^{k} \rho_{i,j} + \sum_{j=0}^{k} \rho_{j,i} - \rho_{i,i}} \tag{24}$$

where true positive (TP) represents the number of correctly predicted cloud (cloud shadow) pixels, the false positive (FP) represents the number of incorrectly predicted cloud (cloud shadow) pixels, the true negative (TN) represents the number of correctly classified noncloud pixels, and the false negative (FN) represents the incorrectly classified cloud (cloud shadow). k denotes the number of categories (excluding background), $p_{i,i}$ denotes the number of TP, $p_{i,j}$ denotes the number of categories belonging to category i but predicted as category j.

TABLE V
COMPARISON OF DIFFERENT CONVOLUTION NETWORKS

Network	GFLOPs	Params	MIoU(%)
MobileNetV2 [43]	5.51	18.62	71.89
VGG16 [44]	62.23	32.06	72.45
VGG19 [44]	78.88	42.68	72.98
ResNet-34	4.52	23.91	72.73
ResNet-50	**17.36**	**65.48**	**73.22**
ResNet-101	21.09	84.47	73.46

The bold values represents our experiments with different models. After considering the parameters, computational complexity, and segmentation accuracy represented by MIoU, we selected ResNet-50 as the backbone for our model. This selection is highlighted in bold.

TABLE VI
ABLATION FOR DIFFERENT MODULES IN THE NETWORK

Module		MIoU
Encoder	Decoder	
Convolution	CUP	73.22
Convolution+Transformer	**CUP**	**75.28 (2.06↑)**
Convolution+Transformer+CFF	**CUP**	**76.62 (1.34↑)**
Convolution+Transformer+CFF+PPM	CUP	76.58
Convolution+Transformer+CFF+ASPP	CUP	76.75
Convolution+Transformer+CFF+CLF	**CUP**	**77.32 (0.7↑)**
Convolution+Transformer+CFF+CLF	**CHF**	**77.85 (0.53↑)**
Convolution+Transformer+CFF+CLF ※	**CHF**	**78.44 (0.59↑)**
Convolution+Transformer+CFF+CLF+CBF	**CHF**	**78.86 (1.01↑)**

The bold values represent modules or structures that have been proven to improve segmentation performance compared to previous iterations through a series of ablation experiments.

C. Ablation Experiments on CloudSEN-12 Dataset

First, we compared the performance of various convolution networks and different versions to determine the network structure for the convolution branch in the backbone. We employed cascaded upsampling (CUP) as the fundamental decoder, which consists of multiple upsampling steps, to decode hidden features for producing the final segmentation output. Table V shows the results. We ultimately chose ResNet-50 as the convolution branch.

Then, we progressively added the transformer branch and the proposed modules to the network to verify the performance of each module and the entire network. The research in this section mainly uses MIoU metric to evaluate. Table VI shows the results, and it can be seen that our proposed network contains all the modules that achieve the best results.

1) Dual-branch ablation: The convolution structure can effectively capture local features when processing images, while the transformer structure can better capture global context information. By combining convolution and transformer into a dual-branch structure, their advantages can be used at the same time to obtain rich feature representation, thereby improving the accuracy of semantic segmentation. Experiments show that the MIoU is increased from 73.22% to 75.28% by using the dual-branch model, which indicates that the dual-branch structure is very effective in extracting spatial and semantic information.

TABLE VII
COMPARISON OF DIFFERENT NETWORKS ON CLOUDSEN-12 DATASET

Structure	Network	Overall			Cloud			Cloud shadow		
		MIoU (%)	PA (%)	MPA (%)	*P* (%)	*R* (%)	*F1* (%)	*P* (%)	*R* (%)	*F1* (%)
CNN	FCN-32 s [10]	71.23	86.98	84.8	87.55	89.85	88.68	78.7	61.31	68.92
	DANet [45]	71.79	87.02	84.14	88.59	88.44	88.51	75.51	66.02	70.44
	BiSeNetV2 [46]	74.19	88.12	85.47	90.08	89.81	89.94	77.26	71.03	74.01
	PAN [17]	74.83	88.48	86.48	91.43	88.67	90.02	79.97	70.48	74.92
	CGNet [47]	74.98	88.59	86.23	90.27	90.39	90.32	79	71.04	74.8
	LinkNet [48]	75.19	88.63	86.64	91.96	88.38	90.13	80	71.5	75.51
	DenseASPP [49]	75.32	88.77	86.61	91.44	89.23	90.32	79.71	71.36	75.3
	DeeplabV3 [15]	75.33	88.71	86.79	89.79	90.66	90.22	81.03	70.98	75.67
	HRNet [50]	76.45	89.25	87.03	91.35	90.04	90.69	80	74.27	77.02
	OCRNet [51]	76.74	89.5	87.66	91.44	90.32	90.87	81.91	72.8	77.08
	CDUNet [52]	76.99	89.59	86.57	91.41	90.52	90.96	81.73	73.03	77.13
	SegNet [12]	77.01	89.62	87.56	91.22	90.63	90.92	81.5	73.14	77.09
Transformer	SETR [53]	73.9	87.78	85.38	88.92	89.96	89.43	78.07	71.29	74.52
	PVT [21]	76.62	89.03	86.59	89.67	90.43	90.04	78.98	72.34	75.51
	SwinUNet [54]	77.53	89.78	87.61	91.16	**91.19**	91.17	80.88	75.96	78.34
Hybrid	CvT [23]	73.93	87.93	85.16	89.62	89.44	89.52	76.55	71.44	73.9
	MPViT [24]	77.22	88.89	87.37	91.66	89.67	90.65	78.28	73.79	75.96
	DBNet [36]	77.37	89.71	87.4	91.7	90.51	91.1	80.03	**76.17**	78.05
	HyCloud (ours)	**77.85**	**90.03**	**88.33**	**91.82**	**92.1**	**91.95**	**82.37**	76.12	**79.12**
	HyCloudX (ours)	**78.86**	**90.43**	**88.88**	**92.7**	91.1	**91.89**	**83.03**	**78.43**	**80.66**

The bold values represent the top two ranked items for each metric in our series of comparative and generalization experiments.

2) Ablation of CFF: This module is used to fuse the spatial location information obtained by the convolution branch and the global context information extracted by the transformer branch, combine the high-level features and the low-level features with rich information, thereby improving the recognition accuracy. The results of Table VI show that the CFF module is an effective module, which can increase the MIoU of the model by 1.34%.
3) Ablation of CLF: Since clouds and cloud shadows have different shapes, sizes, and complex edges, the boundaries generated by the existing networks are relatively rough, so it is necessary to introduce multiscale feature extraction. The ASPP module in DeepLab and the PPM module in PSPNet are also known to extract multiscale information. We compare the proposed CLF module with the PPM module and the ASPP module. It can be seen from Table VI that the proposed CLF module is an effective module that can increase the MIoU of the network by 0.7%.
4) Ablation of CHF: After the encoding stage, the network obtains deep features containing global semantic information, but it loses some spatial details, particularly boundary information. To enhance the segmentation accuracy of the boundary details, we replaced the CUP module with the CHF module. This module can increase the MIoU of the model from 77.32% to 77.85%.
5) Ablation of CBF: The introduction of infrared bands into the semantic segmentation of clouds and cloud shadows can provide rich information, such as temperature, absorptivity, and reflectivity. We conducted experiments, comparing the direct concatenate of all bands as input into the network (denoted by ※) with the separate processing of auxiliary bands using the CBF module. We found that the latter further improves segmentation performance. From Table VI, we can see that the CBF module we proposed is an effective module, which can further increase the MIoU of the network from 77.85% to 78.86%.

D. Comparison Experiments on CloudSEN-12

In this section, our proposed network is compared with various state-of-the-art networks. These networks are mainly divided into three categories according to their architecture: based on convolution structure, such as FCN, DeepLab, and OCRNet. Based on transformer structure, such as SETR, PVT, and SwinUNet. Based on convolution–transformer hybrid architecture, such as CvT, MPViT, and DBNet.

Table VII is the comparison result of different networks. According to the overall ranking of MIoU metric, FCN-32S and DANet have the worst performance, and then according to the MIoU metric, they are SETR, CvT, BiSeNet V2, PAN, CGNet, LinkNet, DenseASPP, DeepLab V3, HRNet, PVT, OCRNet, Se-gNet, MPViT, DBNet, and SwinUNet, among which DBNet and SwinUNet performed better. According to the structure of the network, the transformer network usually has better performance than the CNN network. However, the network using the convolution-transformer hybrid structure has not been further improved on the basis of transformer. This is because the previous hybrid structure networks are too simple in dealing with the features extracted by different structural branches, or lack effective multiscale feature extraction capabilities, resulting

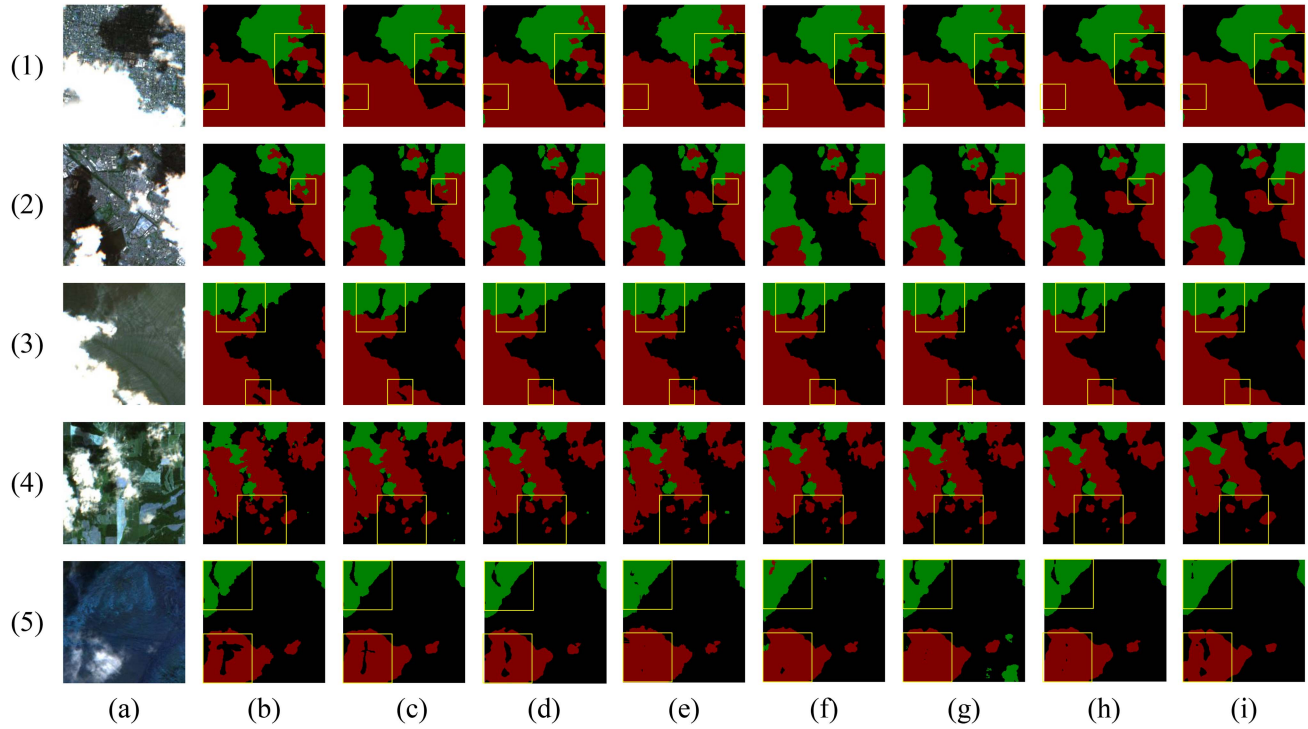

Fig. 8. Comparison of networks in different scenarios. (a) Test images. (b) Label. (c) Segmentation results of HyCloudX. (d) Segmentation results of HyCloud. (e) Segmentation results of SwinUNet. (f) Segmentation results of DBNet. (g) Segmentation results of SegNet. (h) Segmentation results of OCRNet. (i) Segmentation results of DenseASPP.

in not making full use of the advantages of this architecture. Our HyCloud is ahead of CNN structure, transformer structure, and the same type of hybrid structure in MIoU, PA, and MPA metrics, reaching 77.85%, 90.03%, and 88.33%, respectively. On this basis, HyCloudX has a greater improvement, with MIoU, PA, and MPA reaching 78.86%, 90.43%, and 88.88%, respectively. According to the categories of the dataset, HyCloud(X) is basically ahead of other networks in terms of P, R, and $F1$ metrics of all categories. Although the R metric of HyCloud on cloud shadow and the R metric of HyCloudX on cloud are not the highest, there is only a slight gap compared with the well-behaved methods.

We randomly selected five images according to the categories of cities, villages, open spaces, waters, and selected several networks with the highest MIoU in the convolution structure, transformer structure, and hybrid structure as representative networks, and used them to compare the segmentation results. Fig. 8 is the relust of comparison. The red area of the label image represents the cloud, the green area represents the shadow of the cloud and the black area is the background. For the segmentation results of cloud and cloud shadow, we generally evaluate the segmentation accuracy and local details of the edge, such as small area cloud and cloud shadow misjudgment (less or more judgment) and the segmentation of thick and thin clouds, cloud shadow or background inside the cloud. The segmentation results of DenseASPP are relatively rough. For example, small clouds in (1) are not detected, and there are many missed and false pixels in the cloud and cloud shadow boundaries. Compared with DenseASPP, OCRNet, and SegNet and DBNet have improved the accuracy of cloud and cloud shadow boundary segmentation, but there are still deficiencies in the identification of small clouds and inside clouds. For example, DBNet and SegNet missed the detection of small clouds in (1) and (4), respectively. Compared with the previous networks, SwinUNet can better deal with the identification of small clouds and the situation of thick and thin clouds, cloud shadows and ground inside clouds in most cases, but in some cases [such as the lower left area in (5)] there are still problems. This is because from the perspective of visible light, the hickness of the cloud is difficult to capture. When facing a complex background, the texture of the cloud and cloud shadow is similar to it, and the complex atmospheric optics interfere with the image. These factors make the segmentation of clouds and cloud shadows challenging. Our HyCloudX achieved the best results, which is due to the use of convolution and transformer dual-branch structure, which combines the advantages of these two network structures for local feature extraction and global information modeling, respectively. The CFF module is used to guide the extraction of features between the two branches, and the CLF module is used to extract and fuse multiscale feature information. The feature information of the two branches is deeply multiscale feature extracted and fused. In the decoding process, the CHF module uses the multiscale features of the above-mentioned backbone and modules to guide the network to repair the detailed information of the segmented boundary, which reduces the problem of serious loss of boundary details after deep down sampling. It plays a crucial role in improving the final effect of cloud and cloud shadow segmentation tasks with complex boundaries. More importantly, our network introduces an infrared band to assist the visible light channel for segmentation, so that

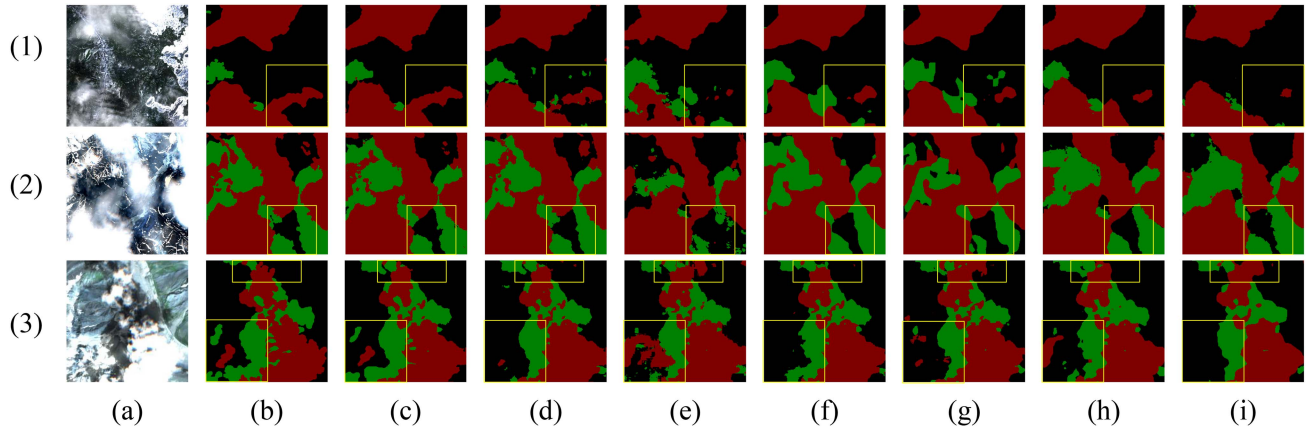

Fig. 9. Comparison of different networks under noise. (a) Test images. (b) Label. (c) Segmentation results of HyCloudX. (d) Segmentation results of HyCloud. (e) Segmentation results of SwinUNet. (f) Segmentation results of DBNet. (g) Segmentation results of SegNet. (h) Segmentation results of OCRNet. (i) Segmentation results of DenseASPP.

the network can accurately locate clouds and cloud shadows. Experiments show that the proposed method not only improves the accuracy of cloud and cloud shadow segmentation, but also retains rich boundary details. The judgment of small clouds and the processing inside the cloud show the best performance compared with other networks.

Fig. 9 is the segmentation results of various networks under noise interference (such as snow and other backgrounds similar to clouds). It can be seen that under the interference of noise, DenseASPP, OCRNet, SegNet, DBNe, and SwinUNet have different degrees of missed detection and false detection. Compared with the previous segmentation results of scenes with less noise, these networks cannot achieve accurate segmentation in the case of more noise interference, especially the performance of segmentation boundaries, small-size clouds and cloud shadows, and distinguishing backgrounds will be drastically reduced. This is because a large amount of noise will mislead the network's judgment. In addition, the semantic information and spatial information of cloud and cloud shadow extracted by these networks are also insufficient. In contrast, the HyCloud network has stronger anti-interference ability, fewer false detection pixels, and more refined segmentation results, and HyCloudX network shows amazing excellent segmentation performance under the assistance of the infrared auxiliary band.

Fig. 10 shows the heat maps of attention of different networks, where the first row of each sample is the attention of the network to the cloud, and the second row is the attention to the cloud shadow. The attention degree of a region from strong to weak shows a gradual change from red–orange–yellow–green–blue. It can be seen from the heat map results that CvT does not pay enough attention to clouds and cloud shadows, which also leads to poor segmentation results. The attention of DBNet, OCRNet, and DenseASPP is more concentrated than CvT, but the attention to the boundary is still insufficient, resulting in relatively rough segmentation of the edges of clouds and cloud shadows, as well as small clouds and thin clouds and backgrounds in clouds. In particular, in noisy scenes [e.g., Fig. 10 (3)], other networks have different degrees of bias in the attention of clouds and

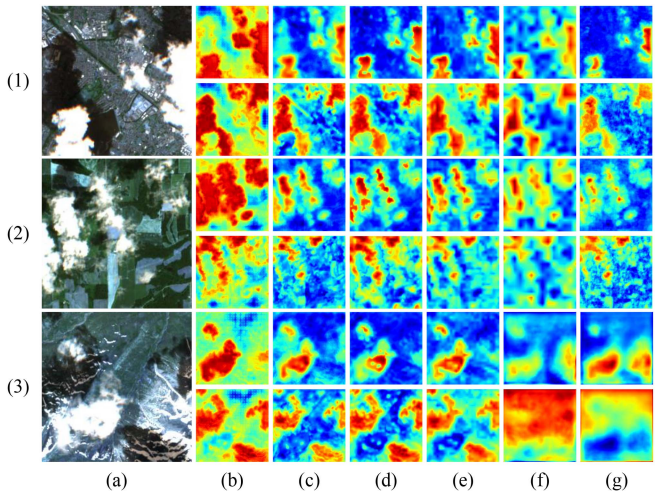

Fig. 10. Visual comparison of attention of different networks. (a) Test images. (b) HyCloudX. (c) DBNet. (d) OCRNet. (e) DenseASPP. (f) CvT. (g) SegNet.

cloud shadows, and CvT and SegNet have even lost a large area of attention to cloud shadows. Our network provides highly focused attention to the subject of clouds and cloud shadows. Besides, due to the complexity of the cloud and cloud shadow boundary, our proposed network also pays some attention to the area near the boundary. Experiments show that it is beneficial for obtaining more boundary information and achieving more precise segmentation results.

E. Generalization Experiments on 38-Cloud Dataset

Table VIII shows the comparison between HyCloud(X) and the current excellent network on the 38-Cloud dataset. From the perspective of network structure category, the network with hybrid structure has better performance than the network with convolution and transformer structure. From the comprehensive performance point of view, SETR and DANet have the worst performance, and then according to the MIoU metric, they are CvT, FCN-32S, BiSeNet V2, LinkNet, DenseASPP,

TABLE VIII
COMPARISON OF DIFFERENT NETWORKS ON 38-CLOUD DATASET

Structure	Network	Overall			Cloud			Cloud Shadow		
		MIoU (%)	PA (%)	MPA (%)	*P* (%)	*R* (%)	*F1* (%)	*P* (%)	*R* (%)	*F1* (%)
CNN	DANet [45]	87.69	93.44	93.45	93.69	93.08	93.38	93.2	93.79	93.5
	FCN-32s [10]	88.67	94	93.99	93.78	94.17	93.98	94.21	93.82	94.02
	BiSeNetV2 [46]	91.28	95.44	95.45	94.68	96.24	95.46	96.22	94.65	95.43
	LinkNet [48]	91.48	95.55	95.55	95.63	95.41	95.52	95.47	95.68	95.58
	DenseASPP [49]	91.62	95.62	95.63	95.35	95.87	95.61	95.9	95.38	95.64
	PAN [17]	91.69	95.66	95.66	95.33	95.99	95.66	96	95.35	95.67
	DeeplabV3 [15]	91.86	95.75	95.77	94.75	96.83	95.79	96.79	94.69	95.74
	CGNet [47]	92.24	95.96	95.98	94.88	96.12	95.49	97.08	94.81	95.95
	PSPNet [14]	92.34	96.02	96.01	95.44	96.61	96.02	96.61	95.43	96.02
	SegNet [12]	92.58	96.14	96.16	96.22	96.02	96.12	96.07	96.27	96.17
	HRNet [50]	92.63	96.17	96.17	96.02	96.29	96.16	96.32	96.06	96.19
	CDUNet [52]	92.64	96.18	96.19	95.52	**96.86**	96.19	96.85	95.5	96.18
	OCRNet [51]	92.69	96.2	96.21	96.48	95.87	96.17	95.94	96.54	96.24
Transformer	SETR [53]	82.65	90.5	90.51	89.86	91.2	90.53	91.16	89.82	90.49
	PVT [21]	87.89	93.65	93.7	93.76	93.8	93.78	94.53	94.62	94.57
	SwinUNet [54]	93.1	96.42	96.42	96.4	96.41	96.41	96.45	96.44	96.44
Hybrid	CvT [23]	87.92	93.57	93.56	93.51	93.56	93.54	93.62	93.58	93.6
	MPViT [24]	92.86	95.96	95.97	95.97	95.93	95.94	95.23	95.74	95.48
	DBNet [36]	93.27	96.52	96.51	96.82	96.16	96.49	96.22	96.67	96.44
	HyCloud (ours)	**93.83**	**96.82**	**97.06**	**97.01**	96.85	**96.76**	**97.11**	**96.72**	**96.91**
	HyCloudX (ours)	**94.71**	**97.28**	**97.54**	**97.3**	**96.97**	**97.13**	**97.79**	**97.29**	**97.53**

The bold values represent the top two ranked items for each metric in our series of comparative and generalization experiments.

PAN, DeepLab V3, CGNet, PSPNet, SegNet, HRNet, CDUNet, OCRNet, SwinUNet, and DBNet. SwinUNet and DBNet perform well, but they are not as good as HyCloud(X). HyCloud is ahead of other networks in MIoU, PA, and MPA metrics, with MIoU, PA, and MPA reaching 93.83%, 96.82%, and 97.06%, respectively. HyCloudX reached 94.71%, 97.28%, and 97.54%, respectively. According to the classification, whether it is cloud or cloud shadow, our HyCloud(X) is basically ahead of other networks in terms of *P*, *R*, and *F*1 metrics of subclassification. Although HyCloud's R metric on cloud detection is not as good as CDUNet's, it has only a slight gap.

Fig. 11 shows the segmentation results of representative networks of convolution, transformer, and hybrid structures in few-cloud, multicloud, and no-cloud scenarios on 38-Cloud dataset. The white area of the label image represents the cloud, and the black area is the background. It can be seen from the comparison that the edges of clouds and cloud shadows of CDUNet, DBNet, and OCRNet are relatively rough. There are a large number of missed and false detection pixels in complex backgrounds, such as snow [such as (2)]. In contrast, SwinUNet shows stronger anti-interference ability, the number of false detection and missed detection points is reduced, and the edge accuracy of the cloud is improved. But compared to our HyCloud, the details are still slightly inadequate. Our proposed network HyCloud uses convolution and transformer dual-branch structure to fully extract and fuse multiscale contextual feature information, and uses multiscale feature information to guide the network to repair the details of the segmentation boundary in the decoding stage, thus achieving better cloud detection and segmentation. However, in cloudless and complex scenes [such as (6)], all networks (including HyCloud) except HyCloudX have different degrees of false detection. HyCloudX can show such an excellent effect, which is benefited from the assistance of the infrared band. By capturing the different absorption rates and reflectivity of clouds, cloud shadows and backgrounds in the infrared band, which cannot be obtained in the visible light channel, and integrating them into the features extracted by the dual-branch structure, the final segmentation performance is improved.

F. Generalization Experiments on SPARCS-Val Dataset

In order to further evaluate the segmentation performance and generalization ability of our network, we also conducted generalization experiments on the SPARCS-Val dataset. The experimental results are shown in Table IX, where the left-hand side shows the overall metrics, and the right-hand side shows the PA of different networks for each category, where CS refers to cloud shadow category, CS OW refers to cloud shadow over water category, W refers to water category, I/S refers to ice and snow category, L refers to land category, C refers to cloud category, F refers to flooded category.

According to the categories of network structure, the transformer structure network performs better than the convolution structure network on more categories and more complex datasets, and the segmentation performance of the hybrid structure network is further improved on the basis of transformer. This

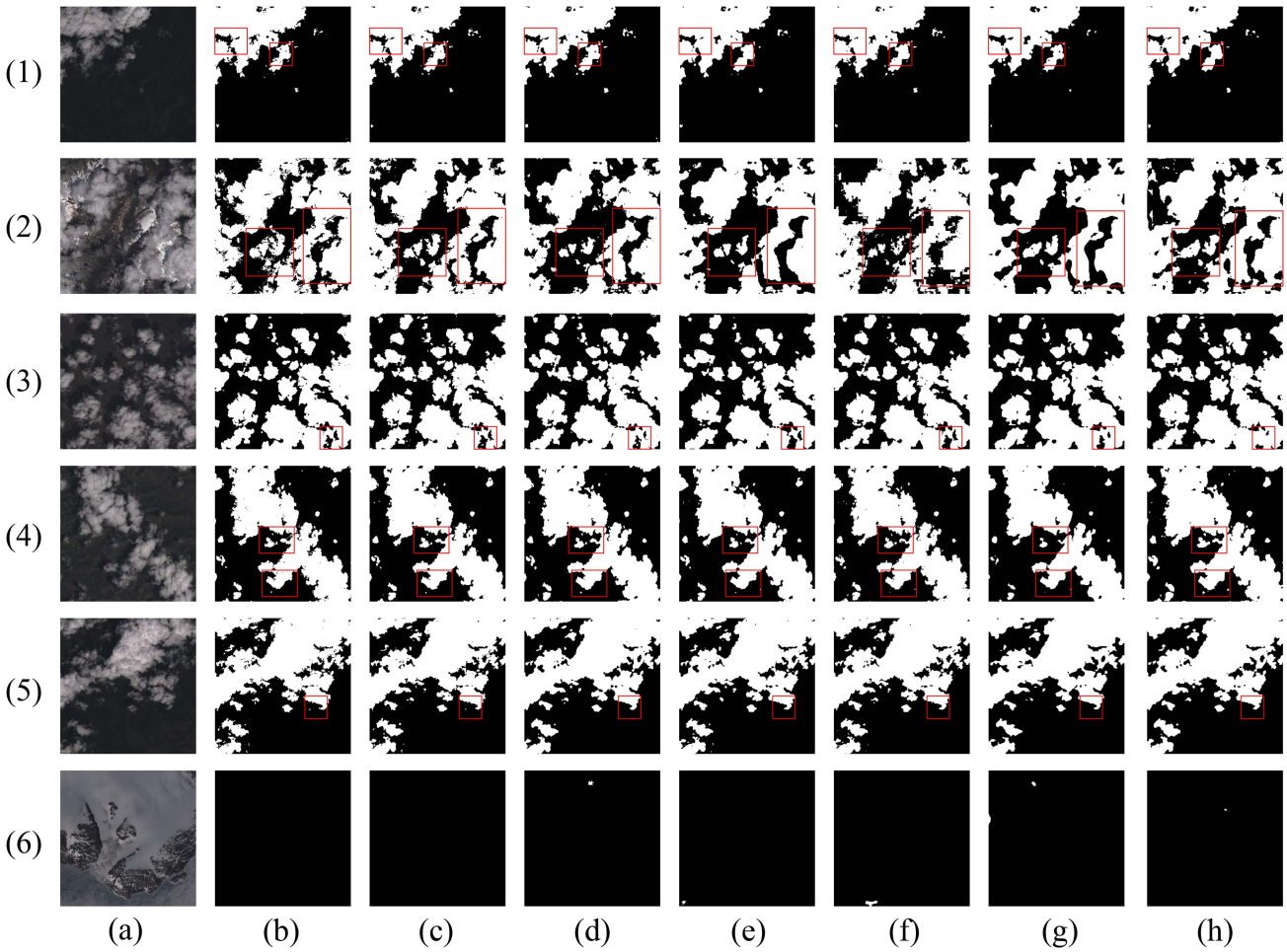

Fig. 11. Comparison of different networks on the 38-Cloud dataset. (a) Test images. (b) Label. (c) Segmentation results of HyCloudX. (d) Segmentation results of HyCloud. (e) Segmentation results of DBNet. (f) Segmentation results of SwinUNet. (g) Segmentation results of OCRNet. (h) Segmentation results of CDUNet.

TABLE IX
COMPARISON OF DIFFERENT NETWORKS ON SPARCS-VAL DATASET

Structure	Network	Overall					Class PA						
		MIoU (%)	PA (%)	MPA (%)	*R* (%)	*F1* (%)	CS (%)	CS OW (%)	W (%)	I/S (%)	L (%)	C (%)	F (%)
CNN	DANet [45]	55.61	85.04	70.28	67.12	66.76	55.44	30.37	89.21	87.63	91.07	76.89	61.36
	FCN-32s [10]	61.38	88.03	75.41	71.2	72.4	64.63	37.51	89.57	89.56	92.36	83.1	71.17
	BiSeNetV2 [46]	64.38	88.57	80.33	73.26	75.8	73.24	57.35	90.2	92.17	91.65	83.18	74.52
	SegNet [12]	65.86	89.3	80.74	75.18	77.53	76.98	55.74	86.82	92.49	92.66	83.5	77.02
	CGNet [47]	66.82	89.93	80	76.37	77.31	72.47	50.32	93	90.53	94.02	84.17	75.49
	PSPNet [14]	67.23	89.92	82.5	75.23	77.81	76.72	56.59	93.06	92.59	92.75	83.81	82.02
	DenseASPP [49]	67.73	89.81	82.42	76.21	78.63	77.13	57.26	94.13	93.23	93.46	81.2	80.58
	DeeplabV3 [15]	68.26	90.06	82.94	76.9	79.05	79.87	60.99	91.28	91.56	93.71	81.97	81.21
	LinkNet [48]	68.62	90.84	83.38	76.8	79.24	79.3	60.06	87.36	91.46	93.2	88.7	83.58
	HRNet [50]	69.74	90.98	84.61	77.3	80.51	82.77	65.31	89.36	93.22	93.23	85.91	82.52
	CDUNet [52]	69.88	90.79	85.59	77.23	80.34	82.26	67.52	91.39	93.43	93.08	86.9	84.6
	OCRNet [51]	69.91	90.94	86.21	77.15	80.04	81.64	68.4	94.27	93.53	92.46	87.4	85.78
Transformer	SETR [53]	63.59	87.73	79.58	72.89	75.38	71.16	55.83	89.85	92.31	91.89	78.9	77.14
	PVT [21]	68.54	89.28	83.02	77.57	80.2	76.34	62.57	90.42	93.23	92.46	84.6	81.55
	SwinUNet [54]	73.0	91.86	86.44	80.44	83.1	80.14	70.09	92.34	**94.17**	94.15	88.27	85.94
Hybrid	CvT [23]	62.68	87.03	78.3	72.42	74.6	67.11	52.85	88.78	93.11	91.64	77.79	76.85
	MPViT [24]	72.98	90.02	85.26	79.49	82.27	80.42	68.95	91.03	93.77	92.96	84.56	85.18
	DBNet [36]	74.04	92.54	87.26	81.01	83.65	81.74	70.21	93.3	94.15	94.44	89.38	**87.62**
	HyCloud (ours)	**76.6**	**93.27**	**88.18**	**84.79**	**85.72**	**83.95**	**70.22**	**94.75**	94.07	**95.45**	**91.5**	87.35
	HyCloudX (ours)	**79.13**	**94.1**	**89.22**	**86.53**	**87.55**	**84.28**	**74.65**	**94.86**	**94.3**	**95.85**	**92.94**	**87.72**

The bold values represent the top two ranked items for each metric in our series of comparative and generalization experiments.

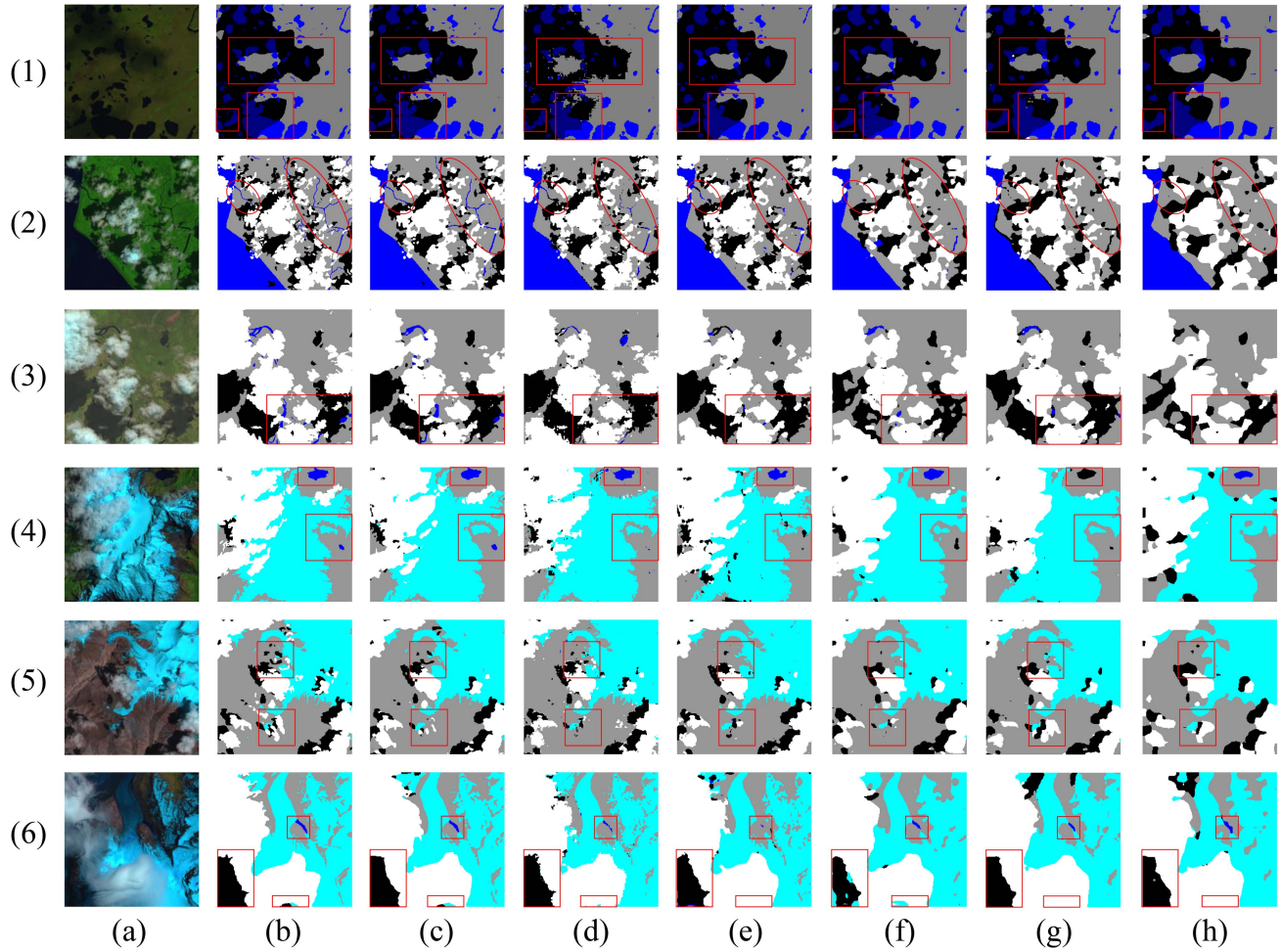

Fig. 12. Comparison of different networks on the SPARCS-Val dataset. (a) Test images. (b) Label. (c) Segmentation results of HyCloudX. (d) Segmentation results of SwinUNet. (e) Segmentation results of DBNet. (f) Segmentation results of OCRNet. (g) Segmentation results of LinkNet. (h) Segmentation results of DeepLab V3.

shows that the dual-branch structure composed of transformer and convolution is effective in feature extraction. According to the overall metrics, our proposed HyCloud is superior to other networks in MIoU, PA, MPA, R, and $F1$ metrics, especially the DBNet, which is second in the overall lead and also uses the convolution–transformer hybrid structure. And HyCloudX is further improved on this basis. From the perspective of class PA, HyCloud has achieved the second place in addition to ice and snow and flood categories, and HyCloud's performance in these two categories is also very close to the second place network. HyCloudX achieves the highest PA in all categories, which is benefited from the additional information of the infrared auxiliary band, our network has strong segmentation performance and generalization ability, and still has excellent performance in some complex situations (such as clouds and shadows on the water, and clouds and snow).

Fig. 12 shows the segmentation results of representative networks of three structures in different scenarios of this dataset. The first group is the segmentation result of cloud shadow on the water area, and the second and third groups are the segmentation results of cloud, cloud shadow, and water area. It can be seen from the graph that the segmentation results of the traditional convolution structure network DeepLabV3, LinkNet, and OCRNet are not good, the edges are rough, and there are a lot of false detections in the water areas category. The segmentation results of transformer structure SwinUNet and hybrid structure DBNet are relatively good, but there is still a certain range of false detection. HyCloudX has the best effect, because the CFF and CLF modules can make the dual-branch structure extract and fuse the multiscale information of the image efficiently and accurately, which reduces the occurrence of false detection. The fourth, fifth, and sixth groups in Fig. 12 mainly show the segmentation results of clouds, cloud shadows and waters by different networks under the noise of ice and snow. It can be seen that due to the interference of ice and snow, other networks have a large area of false detections in the segmentation of water, clouds, and cloud shadow areas. With the assistance of the infrared band, our HyCloudX network can easily obtain the information that the visible light channel cannot obtain, such as the reflectivity and absorptivity of ice and snow, clouds, shadows and waters, which are completely different from those of clouds, clouds and waters. It greatly reduces the occurrence of false detection and improves the segmentation effect under noise interference.

IV. Conclusion

In this article, we propose a multibranch cloud and cloud shadow semantic segmentation network HyCloud(X), which is composed of traditional convolution and transformer branches, and use their ability to extract local features and network global information, respectively. Different from the simple processing of branch features in the same type of hybrid network, we introduce the CFF module to guide each other between the two branches, and use the CLF module to extract and fuse multiscale features to further improve the feature extraction and processing capabilities. In the decoding process, we use the CHF module to utilize the multiscale features of the backbone network and other modules to help the network repair the details of the segmented boundary. This design effectively reduces the loss of boundary details caused by deep downsampling. According to our experimental results, HyCloud has better segmentation performance than other networks using convolution structure, transformer structure or hybrid structure. Moreover, HyCloudX has further improved segmentation performance based on this foundation. This is mainly due to the fact that it not only integrates information from the visible light channels, but also cleverly introduces the infrared auxiliary band, to more accurately understand complex scenes. This enables HyCloudX to achieve remarkable robustness when handling various complex scenarios. Nevertheless, there is still much room for improvement in our network. In the future, we plan to further optimize HyCloudX under the premise of maintaining segmentation performance, reducing the number of parameters and computational complexity of the network, and improve the inference speed of the network. Most importantly, the research in this article shows that the introduction of infrared and other bands into semantic segmentation can greatly improve the segmentation performance. This breakthrough idea opens up new possibilities for the segmentation network in the application of atmospheric science, and has far-reaching significance for cloud detection related work and extensive downstream applications in the field of atmospheric science.

References

[1] Z. Ma, M. Xia, H. Lin, M. Qian, and Y. Zhang, "FENet: Feature enhancement network for land cover classification," *Int. J. Remote Sens.*, vol. 44, no. 5, pp. 1702–1725, 2023.

[2] H. Ren, M. Xia, L. Weng, K. Hu, and H. Lin, "Dual attention-guided multiscale feature aggregation network for remote sensing image change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 4899–4916, 2024.

[3] 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, 2023, Art. no. 1536.

[4] J.-M. Chassery and C. Garbay, "An iterative segmentation method based on a contextual color and shape criterion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 6, pp. 794–800, Nov. 1984.

[5] H. Li, Y. Wang, K. R. Liu, S.-C. Lo, and M. T. Freedman, "Computerized radiographic mass detection. I. Lesion site selection by morphological enhancement and contextual segmentation," *IEEE Trans. Med. Imag.*, vol. 20, no. 4, pp. 289–301, Apr. 2001.

[6] Ren and Malik, "Learning a classification model for segmentation," in *Proc. IEEE 9th Int. Conf. Comput. Vis.*, 2003, pp. 10–17.

[7] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.

[8] K. Chen, X. Dai, M. Xia, L. Weng, K. Hu, and H. Lin, "MSFANet: Multi-scale strip feature attention network for cloud and cloud shadow segmentation," *Remote Sens.*, vol. 15, no. 19, 2023, Art. no. 4853.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

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

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.

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

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

[14] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.

[15] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

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

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

[18] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[19] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2021, *arXiv:2010.11929*.

[20] L. Weng, K. Pang, M. Xia, H. Lin, M. Qian, and C. Zhu, "SGFormer: A local and global features coupling network for semantic segmentation of land cover," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 6812–6824, 2023.

[21] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 548–558.

[22] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.

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

[24] Y. Lee, J. Kim, J. Willette, and S. J. Hwang, "MPVIT: Multi-path vision transformer for dense prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7277–7286.

[25] 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.

[26] 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.

[27] Y. Qu, M. Xia, and Y. Zhang, "Strip pooling channel spatial attention network for the segmentation of cloud and cloud shadow," *Comput. Geosci.*, vol. 157, 2021, Art. no. 104940.

[28] H. Ji, M. Xia, D. Zhang, and H. Lin, "Multi-supervised feature fusion attention network for clouds and shadows detection," *ISPRS Int. J. Geo-Inf.*, vol. 12, 2023, Art. no. 247.

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

[30] L. Ding, M. Xia, H. Lin, and K. Hu, "Multi-level attention interactive network for cloud and snow detection segmentation," *Remote Sens.*, vol. 16, no. 1, 2024, Art. no. 112.

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

[32] C. Zhang, L. Weng, L. Ding, M. Xia, and H. Lin, "CRSNet: Cloud and cloud shadow refinement segmentation networks for remote sensing imagery," *Remote Sens.*, vol. 15, no. 6, 2023, Art. no. 1664.

[33] X. Dai, K. Chen, M. Xia, L. Weng, and H. Lin, "LPMSNet: Location pooling multi-scale network for cloud and cloud shadow segmentation," *Remote Sens.*, vol. 15, 2023, Art. no. 4005.

[34] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, 2023, Art. no. 113856. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425723004078
[35] D. Hong et al., "SpectralGPT: Spectral foundation model," 2024, *arXiv:2311.07113*.
[36] 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. 5410012.
[37] Z. Wang, M. Xia, L. Weng, K. Hu, and H. Lin, "Dual encoder–decoder network for land cover segmentation of remote sensing image," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 2372–2385, 2024.
[38] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE Proc. IRE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[39] 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. 2018.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[41] C. Aybar et al., "CloudSEN12, a global dataset for semantic understanding of cloud and cloud shadow in Sentinel-2," *Sci. Data*, vol. 9, no. 1, 2022, Art. no. 782.
[42] 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.*, 2019, pp. 1029–1032.
[43] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.
[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, *arXiv:1409.1556*.
[45] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3141–3149.
[46] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "BiseNet V2: Bilateral network with guided aggregation for real-time semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, pp. 3051–3068, 2021.
[47] 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.
[48] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Vis. Commun. Image Process.*, 2017, pp. 1–4.
[49] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3684–3692.
[50] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5686–5696.
[51] Y. Yuan, X. Chen, X. Chen, and J. Wang, "Segmentation transformer: Object-contextual representations for semantic segmentation," 2021, *arXiv:1909.11065*.
[52] K. Hu, D. Zhang, and M. Xia, "CDUNet: Cloud detection Unet for remote sensing imagery," *Remote Sens.*, vol. 13, no. 22, 2021, Art. no. 4533.
[53] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6877–6886.
[54] H. Cao et al., "SWIN-UNet: Unet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 205–218.

Ziwei Hu received the B.S. degree in automation from Jinan University, Guangzhou, China, in 2022. He is currently working toward the graduation degree in electronic information from the 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 in electrical engineering 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 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 the Nanjing University of Information Science and Technology, Nanjing, China, where he is also the Deputy Director of Jiangsu Key Laboratory of Big Data Analysis Technology. His research interests include machine learning theory and its application.

Kai Hu received the bachelor's degree from the China University of Metrology, Hangzhou, China, in 2003; the master's degree from the 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, China, in 2015.

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

Haifeng Lin received the Ph.D. degree in forest engineering from Nanjing Forestry University, Nanjing, China, in 2019.

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