Dual Encoder–Decoder Network for Land Cover Segmentation of Remote Sensing Image

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

***Abstract*—Although the vision transformer-based methods (ViTs) exhibit an excellent performance than convolutional neural networks (CNNs) for image recognition tasks, their pixel-level semantic segmentation ability is limited due to the lack of explicit utilization of local biases. Recently, a variety of hybrid structures of ViT and CNN have been proposed, but these methods have poor multiscale fusion ability and cannot accurately segment high-resolution and high-content complex land cover remote sensing images. Therefore, a dual encoder–decoder network named DEDNet is proposed in this work. In the encoding stage, the local and global information of the image is extracted by parallel CNN encoder and transformer encoder. In the decoding stage, the cross-stage fusion module is constructed to achieve neighborhood attention guidance to enhance the positioning of small targets, effectively avoiding intraclass inconsistency. At the same time, the multihead feature extraction module is proposed to strengthen the recognition ability of the target boundary and effectively avoid interclass ambiguity. Before outputting, the fusion spatial pyramid pooling classifier is proposed to merge the outputs of the two decoding strategies. The experiments demonstrate that the proposed model has superior generalization performance and can handle various semantic segmentation tasks of land cover.**

***Index Terms*—Dual encoder–decoder, image segmentation, land cover, vision transformer.**

I. Introduction

LAND cover segmentation is an important issue in remote sensing image processing. It involves the processing of massive remote sensing image data and requires algorithms with high-level semantic segmentation capacity at the pixel level [1]. Accurate land cover segmentation technology can provide reliable data support for urban planning, ecological environment detection [2], [3], and other fields [4], and help to improve the efficiency of urban resource utilization and prevent natural disasters [5], [6]. Therefore, the development of efficient and accurate land cover segmentation algorithm is one of the research hotspots of remote sensing image processing, and it is also an urgent need in practical application.

Before deep learning technology is applied to semantic segmentation, machine learning methods are the mainstream, mainly divided into two categories: feature-based [7] and pixel-based [8]. These traditional methods require manual design of features and rules, and are difficult to deal with complex scenes and multicategory problems. With the development of deep learning technology, CNN-based methods have shone in the field of computer vision. As an important downstream task of computer vision, semantic segmentation emphasizes the pixel-level classification ability of the model, which puts forward higher requirements for the feature extraction and judgment ability of the model [9], [10], [11], [12]. In 2015, Long et al. [13] proposed a fully convolutional neural network. It modifies the traditional convolutional neural network to support pixel-level output. However, it is still difficult to locate the edge of the object, and it is prone to segmentation defects at the pixel level, and the processing of occluded objects is still not ideal. In 2015, Ronneberger et al. [14] proposed a network UNet with encoder–decoder structure and introduced skip-connection. However, when dealing with excessive length-width ratio input images, cross-layer interaction loses its effect, resulting in low accuracy. In 2017, Zhao et al. [15] proposed PSPNet, which is characterized by a pyramid pooling structure that enables the network to effectively focus on scenes at different scales. However, it does not deal with the scale and rotation invariance in geometry, and it is easy to lose spatial information. In 2018, Li et al. [16] proposed DeepLabV3, which uses dilated convolution and atrous spatial pyramid pooling (ASPP) structure to process multiscale objects, but it has poor accuracy in segmenting vivid, moving, and deformed objects. In 2020, Khan et al. [17] combined the advantages of DenseNet [18] and UNet in multiscale feature extraction and retaining low-level features, respectively, and innovatively proposed a deep hybrid network based on two classical networks, which provides a new idea for semantic segmentation tasks.

In summary, CNN-based semantic segmentation methods generally have two ways to obtain global information. First, the receptive field is improved by using dilated convolution or multiscale pooling for deep features. Second, the overall architecture of encoder–decoder is adopted in the model and the context is connected by skip connection. The application of transformer [19] in the field of image has completely changed the method of obtaining global information. Transformer uses a self-attention mechanism to calculate the correlation between

Manuscript received 8 October 2023; revised 24 November 2023; accepted 20 December 2023. Date of publication 27 December 2023; date of current version 8 January 2024. This work was supported by the National Natural Science Foundation of PR China under Grant 42075130. *(Corresponding author: Min Xia.)*

Zhongchen Wang, Min Xia, Liguo Weng, 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: 202212490598@nuist.edu.cn; xiamin@nuist.edu.cn; 002311@nuist.edu.cn; 001600@nuist.edu.cn).

Haifeng Lin is with the College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China (e-mail: haifeng.lin@njfu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3347595

each element and other elements in the sequence data [20], and through the whole image processing, thereby eliminating the limitations of local features and obtaining more global context information. ViT, proposed by Dosovitskiy et al. [21] in 2020, is an image classification model based on transformer, which applies self-attention mechanism to image recognition for the first time. However, its self-attention is always carried out on the largest window, which causes its complexity to increase squarely with the size of the image, so it is not suitable for processing dense predictive tasks. Based on the idea of ViT, Chen et al. [22] proposed SETR in 2021, which uses transformer as an encoder in a seq2seq framework and uses small kernel convolutions, allowing it to extract features without increasing the depth of the feature maps. Wang et al. [23] proposed PVT in 2021. It is based on the feature pyramid structure and uses a hierarchical form to aggregate multiscale information of the network from low to high. With the further exploration of transformer in the field of vision, researchers have found that the simple transformer structure is difficult to adapt to the detection and segmentation tasks that require higher accuracy, and the larger computational complexity is difficult to cope with high-resolution data input. Therefore, based on these challenges, Liu et al. [24] proposed swin transformer in 2021. It can compute self-attention within a partitioned window and associate global information using window shift operations.

In recent years, hybrid models based on CNN and transformer have emerged in an endless stream [25], [26], [27]. In 2021, Graham et al. [28] proposed a lightweight vision transformer model LeViT through adaptive multiscale feature fusion and cross-layer information transmission. Wu et al. [29] proposed CVT in 2021, which introduces the characteristics of convolutional neural networks into the ViT architecture. The adaptive ability of transformer methods in different tasks is strengthened. In 2022, Chen et al. [30] proposed mobile-former, which uses a bidirectional bridge connection between MobileNet and transformer. The bridge is modeled using a proposed lightweight cross-attention mechanism, which achieves high accuracy with minimal computational cost. In 2022, Lee et al. [31] proposed MPViT, which can independently encode different scale tokens through multiple paths, so as to achieve fine and rough feature representation at the same feature level. The hybrid model can retain the ability of CNN to extract local features while utilizing the global attention mechanism of transformer to capture long-range semantic relationships.

In semantic segmentation research on land cover, remote sensing images are diverse in terms of height and angle of capture, and complex object occlusion relationships within the images severely affect the coherence of semantic information, leading to misdetection or omission [32]. Existing hybrid models mostly pursue lightweight design, which results in shallow depth and insufficient multiscale capability, lacking recognition of small targets and edge details, and prone to intraclass inconsistency and interclass ambiguity [33]. In view of these problems, we design a hybrid structure network of dual encoder–decoder. In the encoding stage, the network uses two strategies, convolution and transformer, to extract features from the image in stages. In the decoding stage, on the one hand, same-stage and adjacent-stage features are fused across stages to achieve multiscale fusion. On the other hand, multihead extraction of deep features is performed to expand the receptive field. Finally, an improved spatial pyramid pooling (SPP) classifier is used to integrate the outputs of the two decoders. The main contributions of our work are listed as follows.

1) In order to achieve accurate segmentation with high resolution, we propose a dual encoder–decoder network DEDNet. This method can combine the advantages of CNN and transformer to make full use of global information. It can complete end-to-end training without any manual parameter adjustment, simplifying the process of land cover detection.
2) We propose the cross-stage fusion (CF) module between encoders with its submodule: the neighbored-stage attention guidance (NAG) unit, the multihead feature extraction (MFE) module, and the fusion spatial pyramid pooling (FSPP) classifier. DEDNet can effectively fuse dual encoder and cross-scale features to improve the understanding of global context information. MFE can perform multiple fusion in deep channels, improve the receptive field of the model, and effectively deal with interclass ambiguity and intraclass inconsistency. FSPP can reduce the influence of feature map attributes on classification performance and improve the generalization performance of the model.
3) Quantitative and qualitative experiments on three datasets show that the comprehensive performance of our proposed DEDNet is always superior to other state-of-the-art methods, achieving high-precision land cover semantic segmentation.

II. Methodology

This article proposes a network DEDNet with dual encoder–decoder. It combines CNN-based and transformer-based feature extraction, extended receptive field, and attention guidance strategies in both encoding and decoding stages, which enables it to effectively fuse local features and global features and accurately identify land cover. The overall architecture of the DEDNet network is shown in Fig. 1. In this section, we first introduce the dual encoder stage of ResNet [34] and swin transformer. Then, we introduce the CF module with its submodule: NAG unit, and the MFE module in the dual decoder stage. Finally, we describe the fusion pyramid pooling (FSPP) classifier for fusing the output features of the dual decoder.

A. Dual Encoder

ResNet is used as the convolution encoder. The translation invariance of CNN gives it strong image recognition and generalization abilities, and it is robust to small changes in the image. The residual structure of ResNet prevents gradient disappearance even when it has many convolutional layers. At the same time, it has more channels than the same-stage swin transformer, giving it stronger model expression and local feature extraction capabilities. The residual unit can be mathematically represented

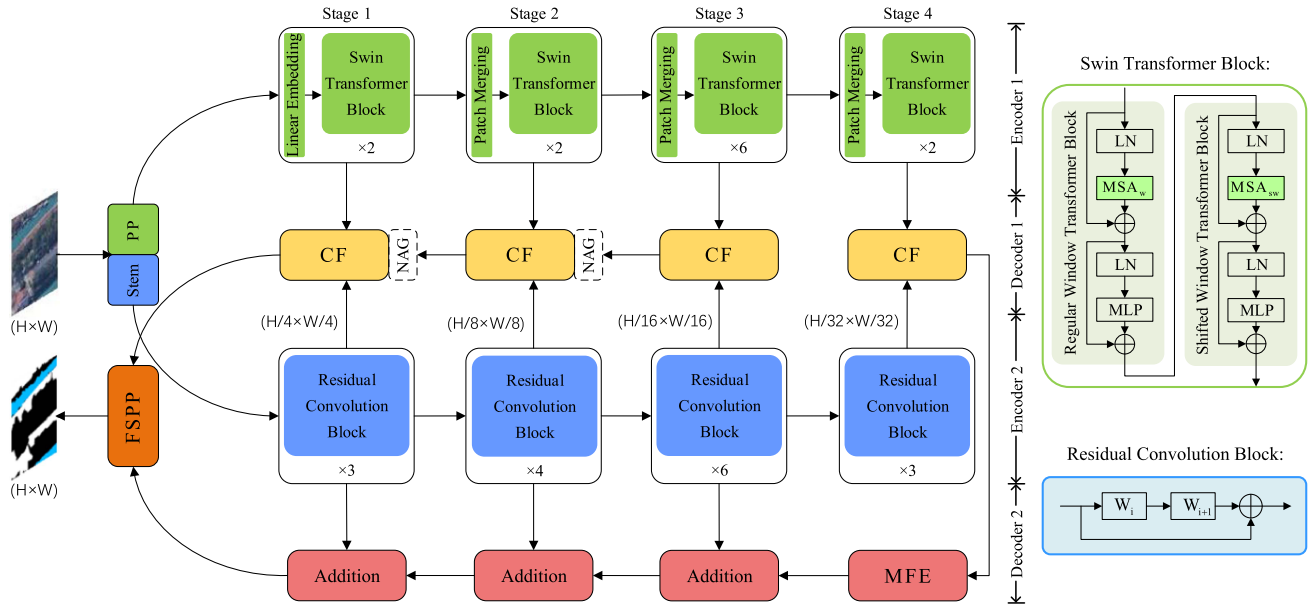

Fig. 1. Overall structure of DEDNet.

as follows:

$$r_{i+1} = W_{i+1}\gamma(W_i r_i) + r_i \tag{1}$$

where r_i is the input vector of the ith residual unit, the function $\gamma(\cdot)$ represents ReLu function, and W_i represents weight matrices.

Swin-T is used as the transformer encoder. First, the input RGB image is partitioned into nonoverlapping patches through patch partition, and each patch is treated as a token. Then, the vector dimension is set using a linear embedding layer. Next, by employing window-based self-attention computation, each window contains the same patches, addressing the issue of quadratic growth in computational complexity with image size, as observed in ViT [35]. Meanwhile, patches originally belonging to different windows can interact after the shift window operation. The continuous swin transformer block using the shift window partition method can be mathematically represented as follows:

$$\hat{s}_i = \mathrm{MSA}_w(\mathrm{LN}(s_{i-1})) + s_{i-1} \tag{2}$$

$$s_i = \mathrm{MLP}(\mathrm{LN}(\hat{s}_i)) + \hat{s}_i \tag{3}$$

$$\hat{s}_{i+1} = \mathrm{MSA}_{\mathrm{sw}}(\mathrm{LN}(s_{i+1})) + s_i \tag{4}$$

$$s_{i+1} = \mathrm{MLP}(\mathrm{LN}(\hat{s}_{i+1})) + \hat{s}_{i+1} \tag{5}$$

where MSA_w and $\mathrm{MSA}_{\mathrm{sw}}$ represent multihead self-attention modules with regular and shifted windowing configurations, respectively, MLP represents multilayer perceptron, $\hat{s}_i$ and s_i represent the output characteristics of $\mathrm{MSA}_{(\mathrm{s})\mathrm{w}}$ module and MLP module, respectively, LN represents LayerNorm layer.

To achieve the construction of multilevel features, swin transformer uses patch merging before the stages, while ResNet uses convolution groups in stages. Their purposes are the same, which is to form a hierarchical structure through stage-by-stage downsampling to extract multiscale features, which is crucial for semantic segmentation tasks. The specific composition of each stage is shown in Table I. In the hybrid encoder structure, we refer to two common encoder mixing structures shown in Fig. 2. Currently, existing hybrid structures can be broadly categorized into two types: serial structure and parallel structure. Models with the latter architecture have gained prominence in the field of semantic segmentation. Sgformer [36] has achieved excellent boundary segmentation results in land cover detection tasks. ST-UNet [35] has demonstrated high accuracy in multiclass, large-scale remote sensing image datasets. In addition, as shown in Tables VI, VII, and VIII, [37], [38] of parallel structure generally exhibit higher segmentation accuracy on land cover datasets compared to [28], [39] of serial structure. Therefore, we adopt the strategy of parallel extraction of features using ResNet and swin transformer in the encoding stage.

TABLE I
HIERARCHICAL STRUCTURE OF DUAL ENCODER

Stage	Transformer encoder	Convolution encoder	downsp. rate (output size)
S1	Swin. $\begin{bmatrix} 7\times7, 96 \\ head\ \ 3 \end{bmatrix} \times 2$	Res. $\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$	$4\times$ (56×56)
S2	Swin. $\begin{bmatrix} 7\times7, 192 \\ head\ \ 6 \end{bmatrix} \times 2$	Res. $\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$	$8\times$ (28×28)
S3	Swin. $\begin{bmatrix} 7\times7, 384 \\ head\ \ 12 \end{bmatrix} \times 6$	Res. $\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 6$	$16\times$ (14×14)
S4	Swin. $\begin{bmatrix} 7\times7, 768 \\ head\ \ 24 \end{bmatrix} \times 2$	Res. $\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$	$32\times$ (7×7)

B. CF Module

To make reasonable use of the advantages of ResNet and swin transformer, we propose a CF module that can fuse the dual

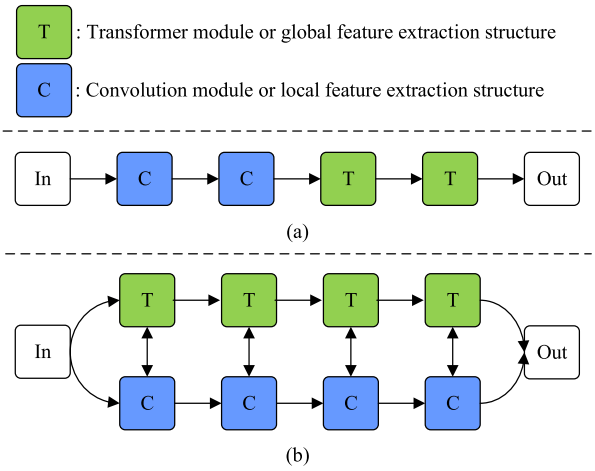

Fig. 2. Two main hybrid structures. (a) serial structure; (b) parallel structure.

encoder features. As shown in Fig. 3, this module can fully integrate the outputs of the two strategies at the same stage during decoding, as well as achieve CF, thus completing the fusion of the encoder with the encoder and the encoder with the decoder within one module.

Given stage n, the output $s \in \mathbb{R}^{(h\times w)\times c_s}$ by the swin transformer encoder is reshaped into $f_s \in \mathbb{R}^{c_s\times h\times w}$, and the output by the ResNet encoder is $f_r \in \mathbb{R}^{c_r\times h\times w}$. First, f_s enhances the local information understanding ability of the feature map through a 3×3 convolution, and then the channel is increased to c_r through a 1×1 convolution. At the same time, f_r and f_s are concatenated to obtain rich channel dimension information, and the channel is then reduced to c_r through global pooling and 1×1 convolution to fully fuse multichannel and global information. Then, the sigmoid activation function is applied to map the value of each weight in the feature map to the interval [0, 1] to obtain ω, which can reflect the importance of the corresponding channel. f_r will be merged with f_s by addition after weighting the attention parameters, resulting in $f_d \in \mathbb{R}^{c_r\times h\times w}$. In the decoding stage, the next stage of the CF output, which has a richer semantic information feature map, will guide attention on f_d. The calculation formula for the abovementioned process can be expressed as

$$\omega = \sigma(\text{Conv}_{1\times1}(\text{Conv}_{1\times1}(\text{Avg}(\text{Cat}[f_r, f_s])))) \tag{6}$$

$$f_d = \text{Conv}_{1\times1}(\text{Conv}_{3\times3}(f_s)) \odot \omega + f_r \tag{7}$$

where $\text{Conv}_{1\times1}(\cdot)$ represents 1×1 convolution layer with batch normalization. $\text{Conv}_{3\times3}(\cdot)$ represents 3×3 convolution layer with batch normalization and ReLU, $\text{Avg}(\cdot)$ represents adaptive average pooling, $\text{Cat}[\cdot]$ represents concatenation, $\sigma(\cdot)$ represents sigmoid activation function, and $\odot$ represents element-level multiplication.

C. NAG Unit

Decoder is often used to reconstruct the original input image or complete the corresponding task from the feature vectors obtained from the encoder. However, some traditional decoders often lose detailed information when processing large input images, resulting in difficulties for the model to accurately capture details of images [40]. To address these issues, we combine the self-attention mechanism and the ideas of BiseNet [41], [42] to design the NAG unit in CF, as shown in Fig. 3. It can dynamically adjust the attention of the network in different regions to enhance the attention to targets in complex backgrounds, weaken interference from nontarget information thereby further improving the performance and effectiveness of the decoder.

Depth-wise convolution (DWConv) is used to process the output $f_{\text{out}}^{n+1} \in \mathbb{R}^{2c_s\times h/2\times w/2}$ of the CF module in stage $(n+1)$ to obtain the high-level features $f_h \in \mathbb{R}^{c_s\times h/2\times w/2}$. At the same time, DWConv also process f_d of the CF module in stage n, as the low-level feature $f_l \in \mathbb{R}^{c_r\times h\times w}$. Then, three linear transformations are performed to map f_h into query matrix q, key matrix k, and value matrix v. Then, by matrix multiplication of q and k and applying the softmax activation function, the attention score matrix is obtained. This matrix is then multiplied by the v to obtain a feature map that considers global information. This feature map is then activated by the sigmoid function to serve as the guide feature map $f_g \in \mathbb{R}^{c_r\times h/2\times w/2}$, guiding the low-level feature map. There are two guiding methods. The first method is that f_l is globally pooled to obtain a feature map with a same size as f_g, so that it can be multiplied by f_g and then upsampled to restore its original size. The second method is that f_l is processed through 1×1 convolution without changing channels, and f_g is upsampled to the same size as f_l, and then multiplied by it. Finally, the outputs of these two guiding methods are added and fused together. The calculation formula for the abovementioned process can be expressed as

$$f_g = \delta(\text{Conv}_{1\times1}^{q}(f_h) \otimes \text{Conv}_{1\times1}^{k^T}(f_h)) \otimes \text{Conv}_{1\times1}^{v}(f_h). \tag{8}$$

$$f_{g1} = \text{Up}(\text{Avg}(f_l) \odot \sigma(f_g)) \tag{9}$$

$$f_{g2} = \text{Conv}_{1\times1}(f_l) \odot \sigma(\text{Up}(f_g)) \tag{10}$$

$$f_{\text{out}} = \text{Conv}_{3\times3}(f_{g1} + f_{g2}) \tag{11}$$

where $\delta(\cdot)$ represents softmax activation function, $\text{Up}(\cdot)$ represents the upsampling, and $\otimes$ represents matrix multiplication.

D. MFE Module

Low-view building images exhibit complex occlusion relationships and few smooth contours, being characterized by a multitude of small protrusions or depressions, such as boundaries, chimneys, and corners obstructed by vegetation [43]. The model tends to blur these small angular boundaries and targets, impacting the accuracy of predictions. Therefore, to improve edge localization and segmentation, the model adopts a multiscale information fusion strategy after preliminary feature extraction. For instance, the pyramid pooling module (PPM) of PSPNet utilizes pooling layers of different sizes to extract features of four different scales. In addition, the ASPP of DeepLabV3plus employs multirate dilated convolutions to create significantly different receptive fields between features. Taking inspiration from these approaches, we introduce the MFE module, illustrated in Fig. 4, for performing multiscale fusion of

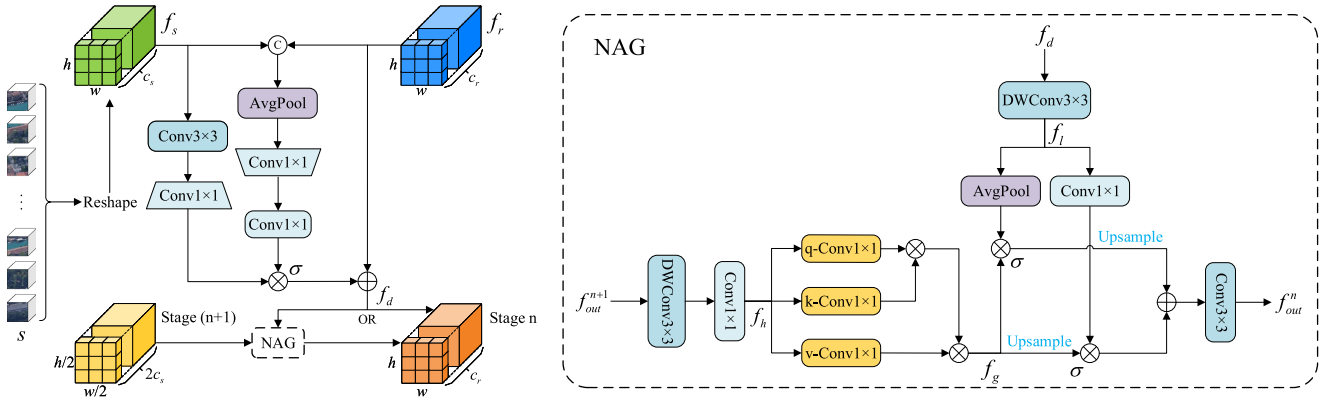

Fig. 3. Structure of CF module and NAG unit.

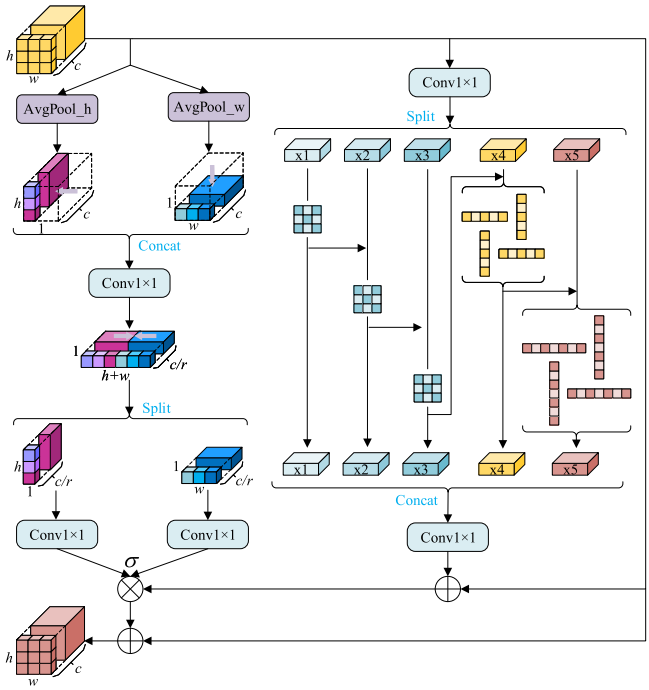

Fig. 4. Structure of MFE module.

low-resolution high-channel feature maps after preliminary feature aggregation. MFE utilizes strip convolution and a residual structure to achieve a wide range of equivalent receptive fields at a low computational cost.

The MFE unit divides the input feature into five heads evenly along the channel dimension, denoted as x_i, $i \in \{1,2,3,4,5\}$. Each head has a corresponding convolutional layer denoted as $L_i(\cdot)$, and its output y_i can be represented as follows:

$$y_i = \begin{cases} L_i(x_i), & i = 1 \\ L_i(x_i + y_{i-1}), & i = 2, 3, 4, 5. \end{cases} \tag{12}$$

For each head, when $i \in \{1,2,3\}$, it corresponds to a 3×3 convolution. To reduce the number of model parameters, we use stripe convolutions for $i = 4$ and $i = 5$, which are equivalent to 5×5 and 7×7 convolution kernels, respectively. We use a split design to realize these convolution kernels, which operate in parallel cascade while maintaining the same receptive field as the original convolution kernels. Here, $\mathrm{Conv}_{h\times w}(\cdot)$ represents the convolution kernel, h and w represent the number of rows and columns of the kernel, and $C_i(\cdot)$ is the specific composition in each head, which can be expressed as follows:

$$L_i(\cdot) = \begin{cases} \mathrm{Conv}_{3\times 3}(\cdot), & i = 1, 2, 3 \\ \mathrm{Conv}_{5\times 1}(\cdot) + \mathrm{Conv}_{1\times 5}(\cdot), & i = 4 \\ \mathrm{Conv}_{7\times 1}(\cdot) + \mathrm{Conv}_{1\times 7}(\cdot), & i = 5. \end{cases} \tag{13}$$

When the input features pass through these filters, their filtered output features can be considered as a result of increased receptive field. Due to the combinatorial effect, this design can generate many equivalent feature scales, thereby enhancing the feature selection ability of the convolution kernel at different scales. At the same time, we use coordinate attention [44] outside of the MFE unit to dynamically learn the relationships between multiple channels and adaptively allocate different weights to enhance the perception efficiency of the module. This improves the performance of the model when processing low-angle building images with rich details and occlusions.

E. Fusion Spatial Pyramid Pooling Classifier

The two decoders employ different feature processing strategies. Decoder 1 consists of CF and NAG, producing shallow features enhanced through attention mechanisms. Decoder 2, on the other hand, comprises MFE and skip connections, resulting in features processed at multiple scales in the deep layers. Therefore, fusing the output features from these two decoders can complement contextual semantic information. However, simple concatenation and dimension reduction may lead to the loss of diversity between low-level spatial information and high-level semantic information, potentially increasing false positive rates [45].

To interact with semantic information across channel dimensions and reduce false positives, we draw inspiration from the commonly used SPP approach in object detection and create fusion spatial pyramid pooling (FSPP) classifier, as shown in Fig. 5. In FSPP, the feature maps are sequentially passed through three maxpooling layers using a kernel size of 5 and padding of 2. This constructs a pyramid-shaped spatial pooling layer

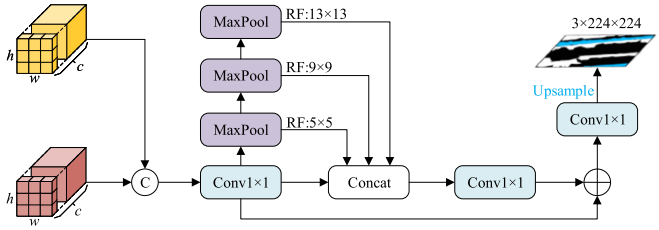

Fig. 5. Structure of fusion spatial pyramid pooling classifier.

at different scales. This design, without adding parameters, effectively achieves the receptive field of maxpooling layers with kernel sizes of 5, 9, and 13. Unlike traditional symmetric receptive fields, the receptive field of FSPP adapts to different scales in different positions for pooling operations. With the introduction of FSPP, the semantic information differences caused by the channel fusion operation of the dual decoder are flexibly alleviated by the multiscale pooling layers. This helps mitigate the risk of model overfitting and reduces the false positive rate, enabling the model to better adapt to targets or scenes at various scales.

III. Experiments

A. Datasets

To test the performance of our proposed model for different land cover scenes, we select three datasets with different numbers of classes, different focus objects, and different resolutions for comparative experiments. Moreover, the cropped datasets are all subjected to data augmentation to improve the model's generalization ability and anti-interference ability, including the following three different types: horizontal and vertical rotation (50%), random rotation (-10° to 10°). After obtaining the raw dataset, we use hold out cross-validation [46] with an 8:2 ratio to split the data into a training set and a validation set. By using these three different datasets, we can evaluate the proposed model's adaptability and accuracy under different resolutions, target categories, and land cover conditions, helping to verify the model's generalization ability. Sliced images in the datasets and their labels are shown in Fig. 6.

1) Building and Water Dataset: We created a building and water dataset (BWD) to test the model's comprehensive performance. The dataset was created from remote sensing images captured by Google Earth (GE), with a total of 300 original images with a size of 1600×900, including rural parks in Asia and Europe, private residences in the U.S., and coastal and lakeside residential areas. In terms of dataset processing, first, the photos were divided into 224×224 images and labeled as three classes of objects: water, building, and background. Then, a total of 10 000 raw datasets were obtained through data augmentation and selection. The remote sensing images in this dataset have a low viewing angle, rich object details, and complex occlusion relationships, which pose a great challenge to the model's ability to handle edge details and small targets.

2) Gaofen Image Dataset (GID): This dataset is the large-scale classification set of the publicly available GID [47] to test the ability to handle multiclass high-resolution images. This dataset is a large-scale land cover dataset constructed from remote sensing images from the GF-2 satellite. It contains 150 GF-2 images with pixel-level annotations, with a size of 7200×6800, including towns, villages, forests, and rivers in various regions of China. The images were labeled into six classes: buildings, farmland, forests, meadow, water, and background. In terms of dataset processing, we first selected 30 raw images with single-image classification of three or more classes and divided the photos into 512×512 images. Then, a total of 11 400 raw datasets were obtained through data augmentation. This dataset has diverse categories and high resolution, which poses a great challenge to the model's ability to handle interclass ambiguity and intraclass inconsistency.

3) L8SPARCS Dataset: This multispectral public dataset was developed by M. Joseph Hughes, Oregon State University, and was derived manually from precollection Landsat-8 scenes, which contains 80 1000×1000 pixel subsets of Landsat 8 OLI/TIRS scenes. Each scene comes with manually created cloud-realistic masks and color composite preview images. Each image is provided as a thematic raster, including categories, such as clouds, cloud shadows, snow/ice, water, and land. This dataset can verify the generalization ability of the model in land cover segmentation under the influence of clouds and cloud shadows.

B. Implementation Details

In terms of hardware, our experiments were conducted on an Intel Core i5-13600KF CPU and an NVIDIA RTX 3090 GPU. The software used was based on PyTorch (version 1.13.1), with adaptive moment estimation (Adam) [48] employed as the optimizer, and cross-entropy loss used as the loss function. During the network training period, the Poly strategy is applied to dynamically adjust the learning rate, starting at 0.001 with a decay exponent of 0.9, and a maximum training iteration of 200. Due to memory constraints, the batch size is set to 32 for BWD and L8SPARCS, and 4 for GID. The evaluation metrics in this study encompass precision (P), recall (R), F1 score, pixel accuracy (PA), mean pixel accuracy (MPA), and mean intersection over union (MIoU)

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{14}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{15}$$

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{16}$$

$$\mathrm{PA} = \frac{\sum_{i=0}^{\mathrm{k}} \rho_{\mathrm{i},j}}{\sum_{i=0}^{k} \sum_{j=0}^{k} \rho_{i,j}} \tag{17}$$

$$\mathrm{MPA} = \frac{1}{k} \sum_{i=0}^{k} \frac{\rho_{i,j}}{\sum_{j=0}^{k} \rho_{i,j}} \tag{18}$$

$$\mathrm{MIoU} = \frac{1}{\mathrm{k}+1} \sum_{i=0}^{k} \frac{\rho_{i,j}}{\sum_{j=0}^{k} \rho_{i,j} + \sum_{j=0}^{k} \rho_{j,i} - \rho_{i,i}} \tag{19}$$

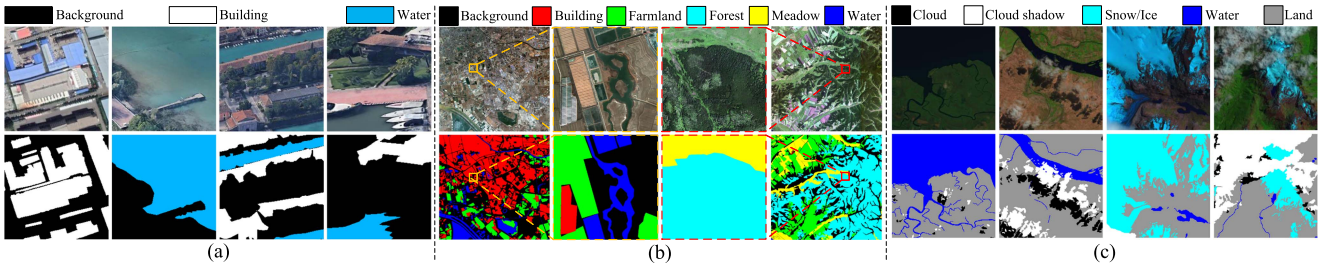

Fig. 6. Sliced images in the datasets and their labels. (a) BWD. (b) GID. (c) L8SPARCS Dataset.

TABLE II
COMPARISON OF DIFFERENT DUAL ENCODER STRATEGIES

Method	Dual Encoder	PA(%)	MPA(%)	MIoU(%)
Variant_1	VGG16+PVT-T	93.03	94.24	87.72
Variant_2	VGG16+Swin-T	93.73	93.28	88.46
Variant_3	ResNet50+PVT-T	94.57	94.33	90.18
Variant_4	ResNet50+Swin-T	**94.78**	**94.40**	**90.52**

The bold entities indicate the optimal.

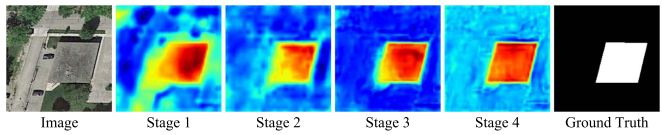

Fig. 7. Visualization of dual encoder fusion features in each stage.

TABLE III
ABLATION OF NAG CONNECTION POSITION IN DECODER1

Method	PA(%)	MPA(%)	MIoU(%)
1←2	94.46	93.91	89.90
1←2←3	**94.78**	**94.40**	**90.52**
1←2←3←4	94.61	94.17	90.22

The bold entities indicate the optimal.

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. k represents the object segmentation class. The true class is represented by $\rho_{i,i}$. $\rho_{i,j}$ represents the number of pixels related to class i, but predicted as class j.

C. Dual Encoder–Decoder Strategy Analysis

1) Encoder Strategy Analysis: The primary role of the encoder is to capture local and global information within the image, enabling subsequent layers of the network to understand the semantic content of different regions in the image. A reliable hierarchical encoder can progressively extract features and encode them into higher level representations. Our model adopts a parallel structure with both CNN and transformer encoders. VGG [49] and ResNet are common choices for CNN encoder, capable of multiscale feature extraction. PVT and swin transformer, on the other hand, are effective at capturing global features hierarchically. To select the most suitable encoder combination for our dual branch parallel structure and efficiently achieve cross-encoder fusion of local and global features, we conduct comparative experiments with various combinations, while keeping other training parameters at their default values in Table II. Experimental results suggest that VGG struggles with high depth, large-scale semantic segmentation tasks, whereas ResNet can effectively constrain the performance of output features at each stage using residual blocks, preventing feature divergence. PVT, based on patch-wise image block encoding, tends to lose global information during feature extraction, whereas swin transformer achieves global information association for the entire image using shifted windows. Therefore, our proposed model utilizes ResNet and swin transformer as the dual encoder.

2) Decoder Strategy Analysis: In order to visually compare the multiscale features of the fusion, Fig. 7 shows the typical channel characteristics of the CF module output at each stage. In the low-level stage of feature fusion, feature maps can only encode low-level information, such as color and texture, so vegetation areas with huge differences from house features will be suppressed, but roads and shadows similar to house colors are still difficult to distinguish. As the number of layers further deepens, the network focuses more on capturing abstract semantic information in the image, such as the concepts of objects, object parts, and overall scenes. The resolution of these feature maps is relatively low, but for the dense prediction task of semantic segmentation, they contain more representative information. It is worth noting that although the output feature map of CF at stage 4 constructs good edge details, the interference of background information increases. This is because as the stage increases, the channel gap between the output features of the CNN encoding block and the transformer encoding block in the same stage gradually widens, resulting in global attention loss after feature fusion.

We repeat the use of CF in Decoder 1, but the NAG submodule in CF might impact the results when it connect between different adjacent CF stages. Therefore, we conduct ablation experiments on the NAG connections between different CF stages. The experimental results are shown in Table III. The numbers in 1←2, 1← 2 ←3, and 1← 2 ← 3 ←4 represent the corresponding CF stages, while the "←" indicates the NAG connections and their directions. As the table shows, the best results are obtained when the NAG is applied between adjacent CF stages, specifically between CF Stage 1 and CF Stage 3.

In Decoder 2, we replace MFE with ASPP and PPM to compare their deep processing capabilities. The experimental results are shown in Table V. Meanwhile, to verify the rationality of the skip-connection method used in the upsampling stage, we compare it with no skip-connection and concatenation-based

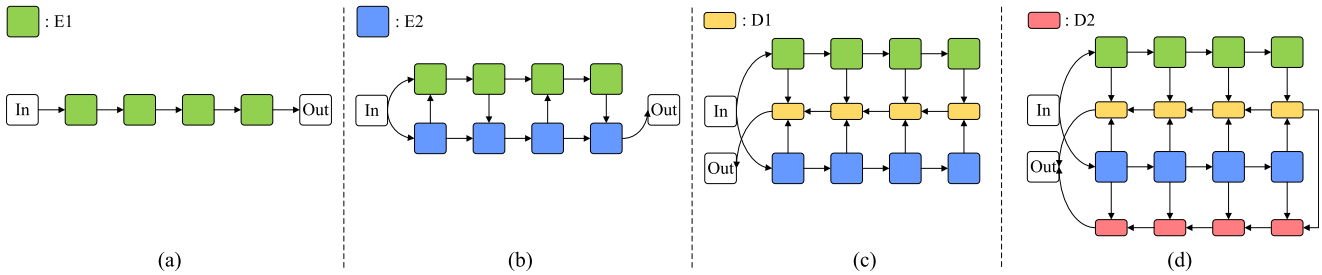

Fig. 8. Illustration of the architecture variants. (a) Single encoder. (b) Dual encoder. (c) Dual encoder with single decoder. (d) Dual encoder–decoder.

TABLE IV
COMPARISON OF UPSAMPLING SKIP CONNECTIONS IN DECODER2

Method	PA(%)	MPA(%)	MIoU(%)
None	94.45	94.21	89.52
Addition	**94.78**	**94.40**	**90.52**
Concatenation	94.37	94.27	90.07

The bold entities indicate the optimal.

TABLE V
ABLATION FOR DIFFERENT MODULES IN THE MODEL

E1	E2	D1	D2	CLF	MIoU(%)
Swin-T	-	-	-	-	85.11
Swin-T	ResNet50	-	-	-	87.15(2.04↑)
Swin-T	ResNet50	CF	-	-	88.74(1.59↑)
Swin-T	ResNet50	CF	ASPP	-	89.26
Swin-T	ResNet50	CF	PPM	-	89.91
Swin-T	ResNet50	CF	MFE	-	90.18(1.44↑)
Swin-T	ResNet50	CF	MFE	FSPP	90.52(0.34↑)

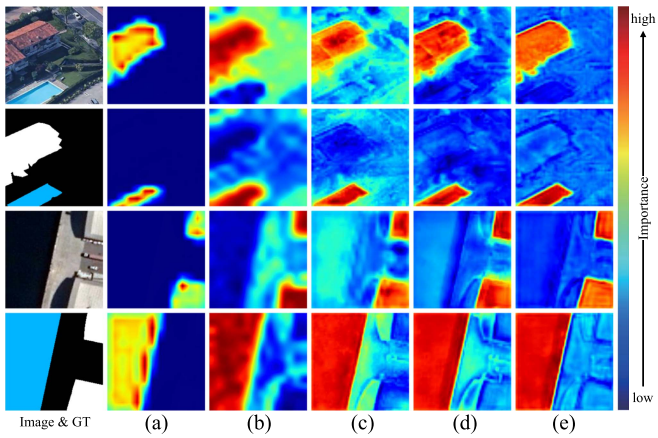

Fig. 9. Ablation heatmaps of the modules. The first line of each sample is the attention to the buildings, and the second line is the attention to the water. (a) ResNet50 encoder. (b) Swin-T encoder. (c) Model without CF and MFE. (d) Model without MFE. (e) Complete model.

skip-connection similar to the decoder of UNet. The experimental results are shown in Table IV. The addition-based skip-connection achieves the best performance during the upsampling stage.

D. Ablation Experiments for Modules

In this section, we conduct an ablation study on BWD to evaluate the efficacy of the modules used in the encoding and decoding stages. Table V shows the results. Fig. 8 shows each ablation variant. In addition, Fig. 9 intuitively shows the impact of our proposed modules on the model.

1) Ablation for Dual Encoder: In Fig. 9(a), the ResNet50 encoder ignores the attention to all backgrounds, but the attention intensity to the target area is weak. The situation in Fig. 9(b) is quite different. Part of the attention of the Swin-T encoder is attracted by the background, but the attention intensity to the target area is very high. These phenomena show that although the CNN-based ResNet50 has limited ability to extract key features, it is also difficult to be interfered. Although the Swin-T based on the self-attention mechanism has a high degree of attention to the target, it has poor anti-interference. Therefore, we upgrade the single-path structure of Fig. 8(a) to the double-path structure of Fig. 8(b), and connect the features of each stage of the two encoders through a simple parallel connection. The experimental results show that the MIoU value is increased to 87.15 %, which is limited compared to the performance improvement of a single encoder. From the unsatisfactory visualization effect of Fig. 9(c), it can be seen that this structure inherits the defects of the two types of encoders.

2) Ablation for CF: Comparing Fig. 8(b) and (c), it can be seen that CF is used to fuse the output features of the two encoders and serve as a decoder. It can enhance the local feature extraction ability of the transformer encoder while improving the global attention of the convolutional encoder, fully realizing the complementary advantages of the two. Experimental results show that CF can increase the MIoU value by 1.59% compared to simply parallelizing two encoders, but the surrounding attention area of the target edge in Fig. 9(d) is scattered, which causes the edge between classes to be blurred.

3) Ablation for MFE: Comparing Fig. 8(c) and (d), it can be seen that we use MFE as the second decoder strategy, responsible for multiscale extraction of deep features, to improve edge segmentation ability and reduce interclass ambiguity of the model. This is consistent with the role of the ASPP module in DeepLabV3plus and the PPM module in PSPNet. Therefore, we replace MFE with ASPP and PPM for comparison. Experimental results show that MFE performs the best among these modules, increasing MIoU of DEDNet by 1.44%. As shown Fig. 9(e), the attention of the target edge is optimized to reduce interclass ambiguity.

4) Ablation for FSPP: As a classifier, this module can fuse the outputs of two decoder strategies at multiple scales and

TABLE VI
EVALUATION RESULTS OF DIFFERENT MODELS ON THE BWD

Frameworks	Models	Building			Water			Overall results		
		P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	PA(%)	MPA(%)	MIoU(%)
ViTs	DeiT	69.09	85.08	76.26	93.29	89.22	91.21	88.58	86.20	79.82
	SETR	76.02	85.40	80.44	93.52	90.19	91.82	90.47	88.70	82.76
	PVT	80.55	89.75	84.90	96.09	91.02	93.49	90.64	90.00	83.32
	Swin-Unet	80.09	85.98	82.93	95.87	92.64	94.23	91.42	92.14	85.38
CNNs	BiseNetV2	78.86	82.21	80.50	93.95	94.28	94.11	91.67	90.24	84.30
	PAN	80.89	85.62	83.19	96.88	93.39	95.10	92.51	92.11	86.59
	DeepLabV3plus	83.94	85.41	84.67	96.76	92.97	94.83	94.03	92.67	87.29
	HRNet	85.36	83.84	84.59	98.07	94.95	96.48	93.93	93.15	87.96
	PSPNet	86.33	91.19	88.69	98.51	94.77	96.60	94.05	93.79	88.32
	UNet	86.15	92.59	89.25	98.70	95.54	97.09	93.97	93.41	88.49
	OCRNet	84.23	92.07	87.89	98.32	**98.16**	98.20	93.86	93.54	88.83
	ACFNet	85.31	**93.51**	89.22	98.72	97.85	98.28	94.21	93.56	89.51
Hybrid structures (ViT+CNN)	LeViT	75.33	85.25	79.98	95.49	91.88	93.65	89.40	86.86	80.46
	CVT	83.34	84.12	83.73	95.76	95.01	95.38	88.83	89.17	83.61
	DBNet	86.57	92.01	89.21	98.91	96.44	97.66	94.15	93.55	89.37
	DBPNet	86.33	92.19	89.16	98.87	97.12	97.99	94.46	93.82	89.48
	DEDNet(Ours)	**86.75**	92.94	**89.74**	**98.95**	97.64	**98.29**	**94.78**	**94.40**	**90.52**

The bold entities indicate the optimal.

effectively avoid overfitting. Experimental results show that FSPP can increase MIoU by 0.34%.

E. Comparative Experiments on Different Datasets

In this section, we compare our proposed model with state-of-the-art models on three different datasets in order to demonstrate the feasibility of our algorithm. In the table, CNNs represent models based on convolution, including DeepLabV3plus [16], UNet [14], etc. ViTs represent models based on version transformer, including SETR [22], DeiT [50], etc. Hybrid structures represent combination models based on CNN and transformer, including LeViT [28], CVT [29], etc. These methods all have their own characteristics. DBNet [37] and DBPNet [38] adopt a dual branch architecture to extract spatial and semantic information. PAN [51] use pyramid structures for multiscale feature fusion. HRNet [52] and OCRNet [53] have efficient parallel computing capabilities and use high-resolution feature maps to improve the spatial resolution of the model. ACFNet uses a category-based receptive field adjustment and attention mechanism. Swin-Unet [54], UNet [14] use different backbones to compose encoder–decoder structures and skip connections. DeiT [50] uses knowledge distillation and data augmentation techniques to achieve performance comparable to large-scale pretrained models when using limited amounts of labeled data.

1) Comparative Experiments on BWD: In Table VI, P, R, and F1 scores are used to evaluate the segmentation performance for two targets. For building detection, our model outperforms other methods in terms of P and F1, reaching 86.75% and 89.74%, respectively, with R slightly lower than ACFNet. For water detection, our model achieves the best P and F1 scores, reaching 98.95% and 98.29%, respectively, with R slightly lower than OCRNet. Second, PA, MPA, and MIoU are selected to evaluate the comprehensive segmentation capability of the model. Our model achieve the best scores, reaching 94.78%, 94.40%, and 90.52%, respectively. It can be seen that the ViTs struggle to leverage their advantages in pixel-level semantic segmentation scenarios, with overall scores generally lower than CNNs. However, hybrid structures have great potential.

Fig. 10 shows the segmentation results of different models on object boundaries. The first and second rows show images of a residential house and a swimming pool photographed from a low angle. Due to the low angle, the buildings and vegetation in the images have complex occlusion relationships, which result in extremely irregular shapes for each category. There are many protrusions and cavities at category boundaries, such as the chimney of the house in the first row. PSPNet, UNet, and OCRNet cannot recognize the shape of the chimney, while DBNet, DBPNet, and ACFNet only roughly segment the chimney into a jagged shape. DEDNet can restore the rectangular shape of the chimney completely. The third and fourth rows show images of a factory photographed from a high angle. Due to the higher angle, the shapes of the objects are relatively regular, but sometimes objects and shadows or road colors are similar, such as the narrow street, shadow, and house colors in the third row. CNN-based models cannot fully identify the background between houses. Due to the lack of multiscale extraction of deep features, DBNet and DBPNet are difficult to locate sparse edge features, so they can only roughly divide the chimney into jagged shapes. The MFE module uses multiscale strip convolution and coordinate attention to make DEDNet better understand the geometric structure of the target edge, which improves the anti-interference ability of the model when dealing with sparse and irregular edges.

Fig. 11 shows the attention of different models to small objects. The first and second rows show images of amusement facilities and swimming pools photographed from a close distance. Due to the short distance, small interfering objects at the water boundary in the image, like the handrail at the edge of the swimming pool in the first row, are prominently marked in the label. PSPNet, OCRNet, and DBNet are unable to identify the shape of the handrails. UNet and ACFNet could only extract

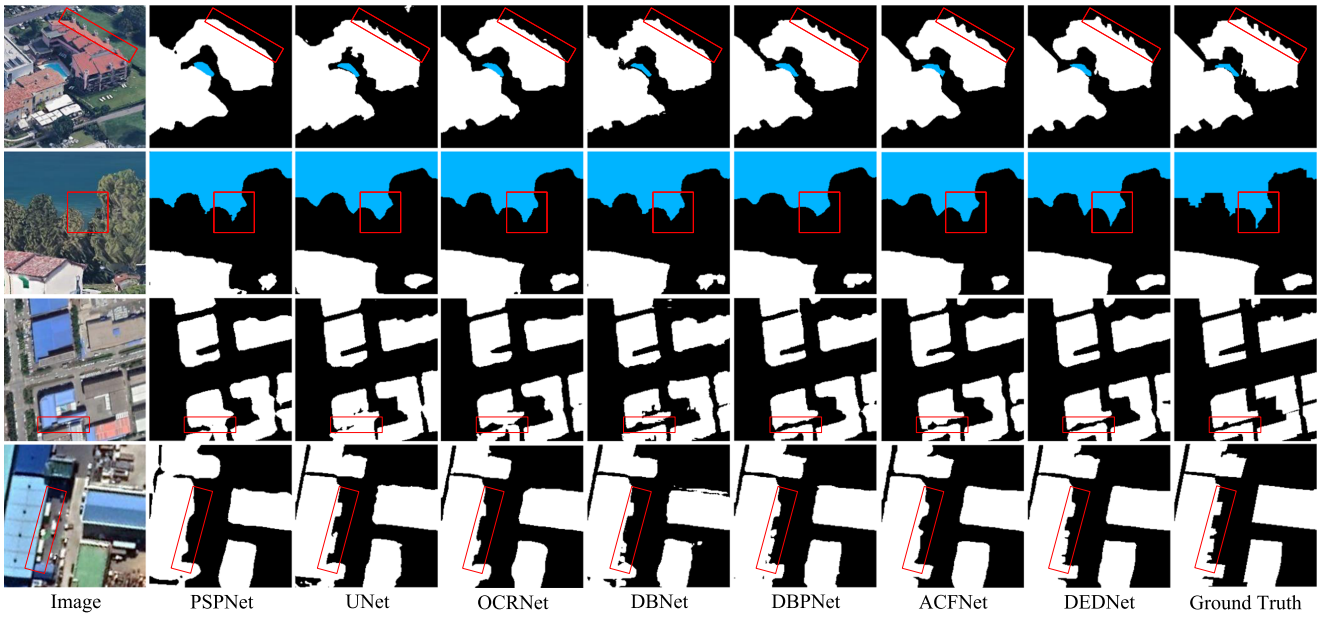

Fig. 10. Comparison of edge segmentation results among different models on BWD.

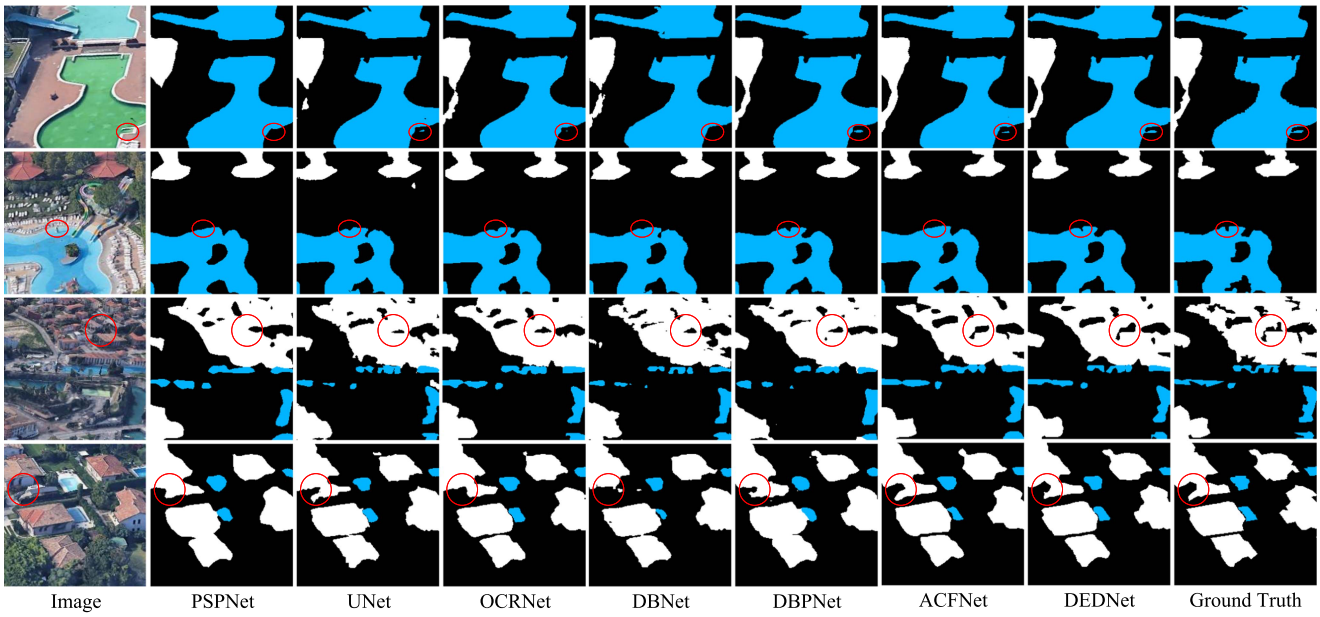

Fig. 11. Comparison of small target results among different models on BWD.

very little information, resulting in a disconnect between the steps leading into the water and the pool. However, our model can accurately recognize the handrails and steps at the edge of the pool, ensuring the continuity of the water. The third and fourth rows show images of multiple buildings and water photographed from a long distance. Due to the long distance, the resolution of the targets is low and the attention of the model to the details of the targets will decrease significantly, such as the buildings in the distance and the vegetation in the background in the third row image. None of the CNN-based models fully capture the outline of the hut. DBNet and DBPNet utilize deconvolution layers and channel attention modules as decoders, respectively. However, these decoders can only correlate a portion of the encoder's information through skip connections, lacking the global information combined with the dual encoder. This global information integration is crucial for capturing subtle features. In contrast, the CF module in our model allows for a more comprehensive integration of the global information from the dual encoder, enabling DEDNet to better focus on small targets.

2) Comparative Experiments on GID: In Table VII, we use the class pixel accuracy (CPA) score to evaluate segmentation performance on five target objects. Our proposed method

TABLE VII
EVALUATION RESULTS ON THE GID

Frameworks	Models	CPA(%)					Overall results		
		Building	Farmland	Forest	Meadow	Water	PA(%)	MPA(%)	MIoU(%)
ViTs	DeiT	86.85	73.23	95.48	73.78	89.61	85.24	84.08	74.46
	SETR	90.24	76.94	94.46	73.48	90.14	85.93	85.80	75.99
	PVT	89.94	77.38	92.35	74.81	91.61	86.11	85.52	76.15
	Swin-Unet	93.86	94.14	97.41	87.90	92.80	90.68	91.95	82.04
CNNs	BiseNetV2	90.72	83.59	96.64	90.08	94.48	90.61	91.69	83.46
	DeepLabV3plus	91.55	89.87	97.30	90.02	94.52	92.20	91.85	85.30
	PAN	91.75	92.87	96.90	90.36	96.13	93.40	93.47	87.78
	HRNet	95.30	94.78	97.42	89.11	96.78	93.69	94.10	88.27
	UNet	90.89	89.65	97.24	80.49	95.08	93.92	92.85	88.30
	ACFNet	92.50	93.36	97.09	87.70	95.70	93.94	93.68	88.94
	PSPNet	91.31	93.74	95.74	91.79	96.91	94.23	93.86	89.55
	OCRNet	94.22	94.40	97.33	90.03	97.50	94.89	94.60	90.61
Hybrid structures (ViT+CNN)	LeViT	89.46	88.28	95.11	76.15	91.95	86.79	85.66	77.82
	CVT	91.29	78.40	96.68	82.38	91.32	88.18	86.91	78.53
	DBNet	**96.13**	95.95	97.26	89.96	97.13	95.09	94.99	90.68
	DBPNet	95.36	94.73	**97.44**	91.79	96.22	95.21	95.02	90.99
	DEDNet(Ours)	94.28	**96.61**	97.40	**91.92**	**97.71**	**95.78**	**95.57**	**92.07**

The bold entities indicate the optimal.

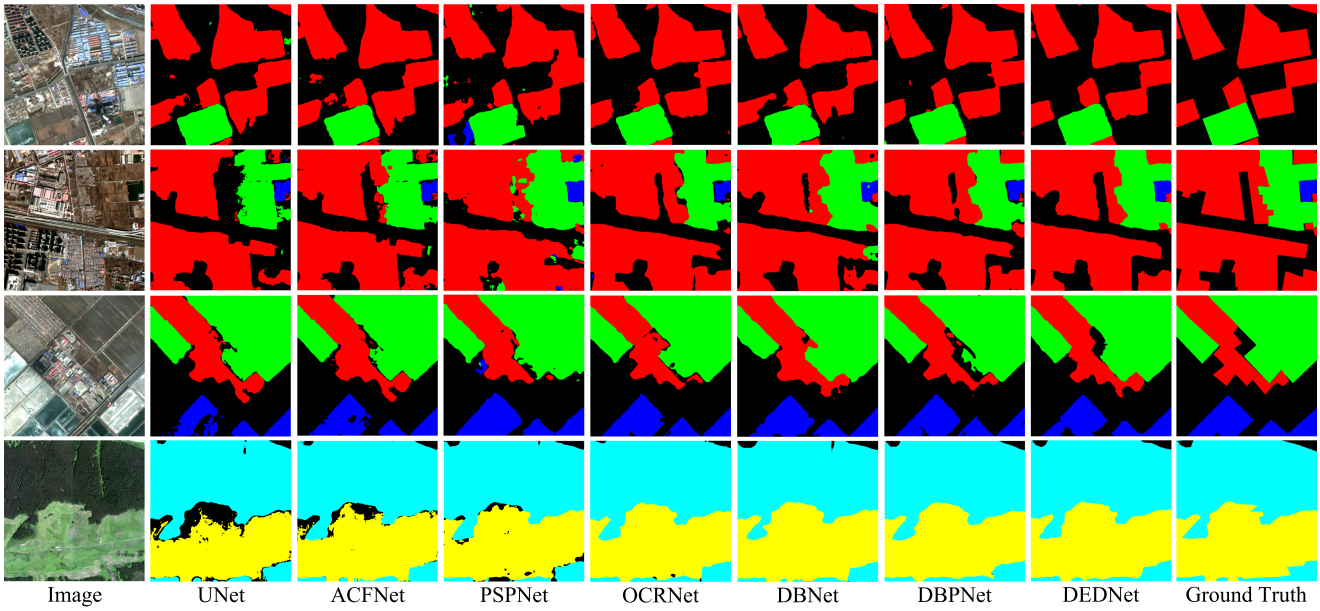

Fig. 12. Comparison of segmentation results among different models on GID.

achieves the highest accuracy in segmenting farmland, meadow, and water, with scores of 96.61%, 91.92%, and 97.71%, respectively, while slightly lower scores than DBNet and DBPNet in building and forest. Then, we use PA, MPA, and MIoU to evaluate the overall segmentation ability. Our proposed model achieve the best scores of 95.78%, 95.57%, and 92.07%, respectively.

In Fig. 12, we find that multiclass land cover segmentation faces some challenges and difficulties during the segmentation process. First, the main difficulty lies in interclass ambiguity. Different types of land cover may appear visually similar, for example, the boundary between buildings and farmland may be blurred, and water and farmland may have similar colors and textures. This interclass ambiguity makes it difficult for the model to accurately distinguish between these similar categories, resulting in misclassification and confusion. Another difficulty is intraclass inconsistency. The same type of land cover may have different appearances and features in different images, such as different types of buildings and different stages of farmland, which increases the difficulty of handling intraclass variability. UNet, PSPNet, and OCRNet have rough boundaries for building groups. ACFNet tends to mistake dark backgrounds for crops. UNet, ACFNet, and PSPNet have obvious intraclass inconsistencies. DBNet and DBPNet mistakenly identify containers with similar colors and shapes to buildings. However, our model can effectively avoid interclass ambiguity and intraclass inconsistency.

TABLE VIII
EVALUATION RESULTS ON THE L8SPARCS

Frameworks	Models	CPA(%)					Overall Results		
		Cloud	Cloud shadow	Snow/Ice	Water	Land	PA(%)	MPA(%)	MIoU(%)
ViTs	DeiT	89.68	78.25	91.62	94.03	94.60	92.27	89.63	80.37
	SETR	91.74	82.29	91.25	92.61	93.96	92.50	90.37	80.73
	PVT	91.95	82.25	93.75	90.95	94.62	92.97	90.70	81.86
	Swin-Unet	91.78	82.53	94.26	92.10	95.31	93.47	91.20	83.07
CNNs	UNet	87.57	72.62	92.58	92.63	92.96	90.60	87.67	78.62
	BiseNetV2	92.20	84.51	93.98	93.67	94.44	93.30	91.76	82.52
	PSPNet	93.56	85.35	94.24	93.48	95.71	94.34	92.47	84.98
	PAN	92.52	85.79	93.90	94.46	96.03	94.40	92.54	85.12
	DeepLabV3plus	93.40	85.48	94.37	95.92	96.03	94.65	93.04	85.52
	HRNet	93.12	86.83	96.18	93.48	96.77	95.11	93.28	86.88
	ACFNet	93.25	86.78	95.38	96.31	96.66	95.19	93.70	87.11
	OCRNet	84.91	89.32	96.46	93.15	93.68	94.27	93.69	88.17
Hybrid structures (ViT+CNN)	LeViT	94.15	85.52	93.91	93.47	92.13	88.36	85.31	82.48
	CVT	93.89	82.52	95.05	89.41	97.09	94.05	90.38	84.43
	DBNet	**95.40**	88.15	96.10	95.91	96.87	95.72	94.36	88.25
	DBPNet	95.13	**89.38**	96.32	96.15	97.03	95.95	94.68	88.86
	DEDNet(Ours)	94.76	88.53	**98.73**	**96.40**	**97.51**	**96.34**	**94.98**	**89.90**

The bold entities indicate the optimal.

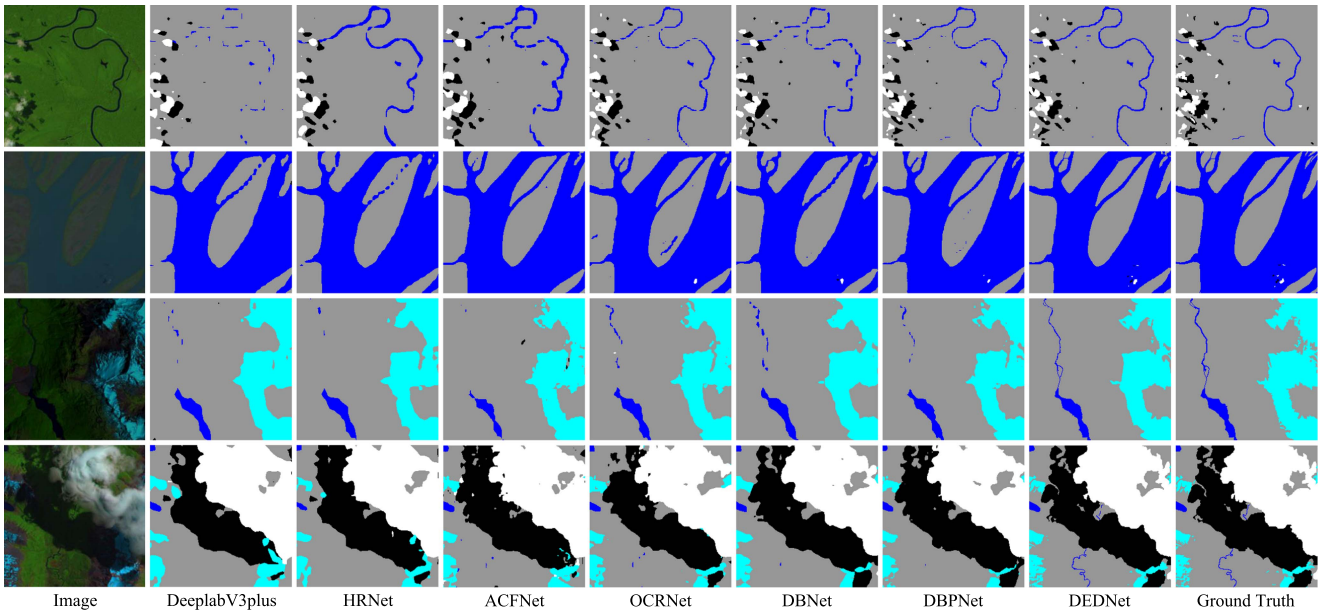

Fig. 13. Comparison of segmentation results among different models on L8SPARCS.

3) Comparative Experiments on L8SPARCS: In Table VIII, we use the CPA to evaluate segmentation performance on five target objects. Our proposed method achieves the highest accuracy in segmenting snow/ice, water, and land, with scores of 98.73%, 96.40%, and 97.51%, respectively, while slightly lower scores than DBNet and DBPNet in cloud and cloud shadow. Then, we use PA, MPA, and MIoU to evaluate the overall segmentation ability. Our proposed model achieve the best scores of 96.34%, 94.98%, and 89.90%, respectively.

In Fig. 13, the focus of the first and second rows is on small streams within mountainous and wetland areas. The small size and meandering shapes of these streams, combined with complex terrain and various environmental interferences like vegetation and rocks, make it challenging to distinguish the target from the background. In addition, variations in lighting conditions can affect image brightness and contrast, further adding to the segmentation challenges. In the third and fourth rows, the areas exhibit significant changes in elevation, transitioning from low-altitude regions to high-altitude regions with transformations from forests to grasslands and then to snowy areas. These regions lack distinct boundary features, making it difficult for segmentation algorithms to accurately capture and delineate these edge details. Furthermore, clouds and their shadows can severely disrupt coherent semantic information, leading to the loss of contextual cues. To overcome these challenges, the model needs to combine sensitivity to subtle texture differences with the utilization of contextual information from the surrounding background. DeepLabV3plus and HRNet

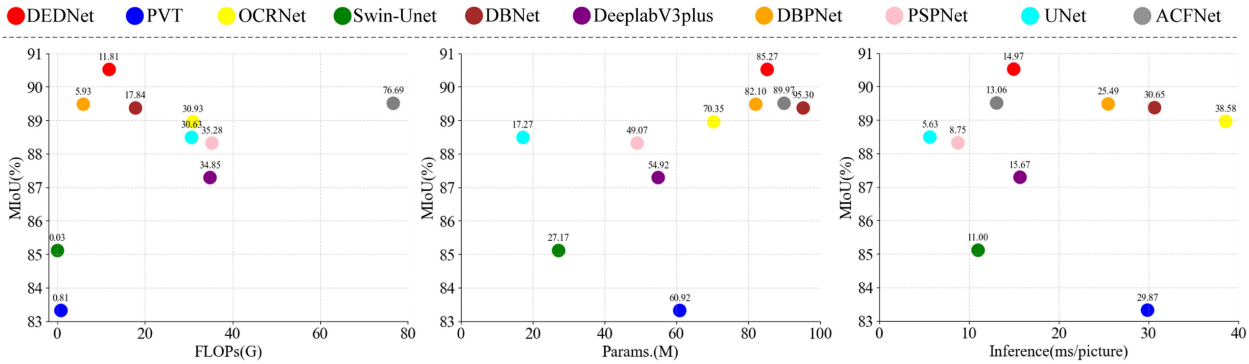

Fig. 14. Floating point comparison diagram of models efficiency metrics. The label corresponds to the abscissa data.

exhibit strong local feature extraction capabilities, but they lack the ability to handle interference effectively, making it difficult for them to capture continuous streams. ACFNet and OCRNet, while excelling at category discrimination, suffer from limited global receptive fields, making it challenging to associate semantic information between distant pixels. This limitation can lead to confusion between ridge shadows and cloud shadows. DBNet and DBPNet, due to their use of transformer encoding at multiple scales, exhibit strong global comprehension abilities, almost entirely avoiding false detections. However, their ability to recover fine details is still not as strong as our proposed model.

4) Efficiency Analysis: Based on performance metrics such as floating-point operations (Flops), parameter count (Params.), and inference time, we compare the efficiency of our proposed model with other models. We randomly select 500 images of size 224 × 224 pixels from the validation set for inference operations, and average all results to evaluate model inference time. As shown in Fig. 14, the Flops of our proposed model is significantly lower than those of CNNs, but due to frequent use of self-attention mechanism and deeper encoder channels, the Params. of our model is higher. ViTs have smaller Params and Flops, but their segmentation results are far inferior to our proposed model. The comprehensive performance of the compared hybrid models is better, but their inference speed is slower than ours.

IV. Conclusion

In this article, we proposes an end-to-end land cover remote sensing image segmentation method using a dual encoder–decoder network. The proposed method first uses CNN and transformer as dual encoder to extract local and global features of the image. Then, CF and MFE are used as dual decoder to perform multiscale feature fusion and deep feature mining. The experiments show that compared with other state-of-the-art methods, our proposed method can accurately locate small targets and restore target boundaries completely, avoiding intraclass inconsistency and interclass ambiguity. In addition, it can handle different types of land cover segmentation tasks with strong generalization performance.

However, there is still room for improvement in our proposed method. First, there is an excessive reliance on a large and diverse set of labeled data, as our model generally employs self-attention mechanism, making the network prone to overfitting on small datasets. Second, the model has a large number of parameters, making it challenging to efficiently deploy the model in edge devices or resource-constrained environments. Future work will focus on optimizing the multiscale fusion of the model by introducing innovative encoding strategies to maintain performance while reducing the number of parameters. In addition, we will fully explore the model's outstanding multiclass semantic segmentation capability and investigate its applicability in other domains, such as healthcare and environmental monitoring.

References

[1] Y. Li, H. Zhang, X. Xue, Y. Jiang, and Q. Shen, "Deep learning for remote sensing image classification: A survey," *Wiley Interdiscipl. Rev.: Data Mining Knowl. Discov.*, vol. 8, no. 6, 2018, Art. no. e1264.

[2] S.-H. Lee, K.-J. Han, K. Lee, K.-J. Lee, K.-Y. Oh, and M.-J. Lee, "Classification of landscape affected by deforestation using high-resolution remote sensing data and deep-learning techniques," *Remote Sens.*, vol. 12, no. 20, 2020, Art. no. 3372.

[3] M. Ortega Adarme, R. Queiroz Feitosa, P. Nigri Happ, C. Aparecido De Almeida, and A. Rodrigues Gomes, "Evaluation of deep learning techniques for deforestation detection in the Brazilian Amazon and cerrado biomes from remote sensing imagery," *Remote Sens.*, vol. 12, no. 6, 2020, Art. no. 910.

[4] 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.

[5] 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, 2021, Art. no. 102597.

[6] Z. Ma, M. Xia, L. Weng, and H. Lin, "Local feature search network for building and water segmentation of remote sensing image," *Sustainability*, vol. 15, no. 4, 2023, Art. no. 3034.

[7] G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review," *ISPRS J. Photogrammetry Remote Sens.*, vol. 66, no. 3, pp. 247–259, 2011.

[8] S. Zheng et al., "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1529–1537.

[9] 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.

[10] 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.

[11] H. Ji, M. Xia, D. Zhang, and H. Lin, "Multi-supervised feature fusion attention network for clouds and shadows detection," *ISPRS Int. J. Geoinf.*, vol. 12, no. 6, 2023, Art. no. 247.

[12] K. Chen, M. Xia, H. Lin, and M. Qian, "Multi-scale attention feature aggregation network for cloud and cloud shadow segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5612216.

[13] 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.

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

[15] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.
[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] S. D. Khan, L. Alarabi, and S. Basalamah, "Deep hybrid network for land cover semantic segmentation in high-spatial resolution satellite images," *Information*, vol. 12, no. 6, 2021, Art. no. 230.
[18] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2017, pp. 4700–4708.
[19] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural. Inf. Process. Syst.*, vol. 30, 2017, pp. 6000–6010.
[20] M. Kaselimi, A. Voulodimos, I. Daskalopoulos, N. Doulamis, and A. Doulamis, "A vision transformer model for convolution-free multilabel classification of satellite imagery in deforestation monitoring," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 7, pp. 3299–3307, Jul. 2023.
[21] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2021, *arXiv:2010.11929*.
[22] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, 2021, pp. 6881–6890.
[23] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 568–578.
[24] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.
[25] 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.
[26] J. Gao, L. Weng, M. Xia, and H. Lin, "Mlnet: Multichannel feature fusion lozenge network for land segmentation," *J. Appl. Remote Sens.*, vol. 16, no. 1, 2022, Art. no. 0 16513.
[27] 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.
[28] B. Graham et al., "Levit: A vision transformer in convnet's clothing for faster inference," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 12259–12269.
[29] H. Wu et al., "Cvt: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 22–31.
[30] Y. Chen et al., "Mobile-former: Bridging mobilenet and transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5270–5279.
[31] 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. 7287–7296.
[32] 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, 2023, Art. no. 4853.
[33] K. Hu, C. Weng, C. Shen, T. Wang, L. Weng, and M. Xia, "A multi-stage underwater image aesthetic enhancement algorithm based on a generative adversarial network," *Eng. Appl. Artif. Intell.*, vol. 123, 2023, Art. no. 106196.
[34] 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.
[35] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding UNET for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jun. 2022, Art. no. 4408715.
[36] 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, Jan. 2023.
[37] 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, Jul. 2022, Art. no. 5410012.
[38] J. Chen, M. Xia, D. Wang, and H. Lin, "Double branch parallel network for segmentation of buildings and waters in remote sensing images," *Remote Sens.*, vol. 15, no. 6, 2023, Art. no. 1536.
[39] S. Mehta and M. Rastegari, "Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer," 2022, *arXiv:2110.02178*.
[40] F. Min, L. Wang, S. Pan, and G. Song, "D ^{2} UNet: Dual decoder U-NET for seismic image super-resolution reconstruction," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5906913.
[41] 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.*, 2018, pp. 325–341.
[42] 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, no. 11, pp. 3051–3068, 2021.
[43] 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.
[44] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13713–13722.
[45] X. Dai, M. Xia, L. Weng, K. Hu, H. Lin, and M. Qian, "Multi-scale location attention network for building and water segmentation of remote sensing image," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5609519.
[46] S. Yadav and S. Shukla, "Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification," in *Proc. IEEE 6th Int. Conf. Adv. Comput.*, 2016, pp. 78–83.
[47] X.-Y. Tong et al., "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, 2020, Art. no. 111322.
[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017, *arXiv:1412.6980*.
[49] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[50] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.
[51] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," 2018, *arXiv:1805.10180*.
[52] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5693–5703.
[53] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 173–190.
[54] H. Cao et al., "Swin-unet: Unet-like pure transformer for medical image segmentation," in *Proc. Comput. Vis. Workshops: Tel*, 2023, pp. 205–218.

Zhongchen Wang received the B.S. degree in automation and the graduation degree in electronic information from the Nanjing Institute of Technology, Nanjing, China, in 2022.

His research interests include deep learning and its applications.

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. He is also the Deputy Director of the Jiangsu Key Laboratory of Big Data Analysis Technology. His research interests include machine learning theory and its application.

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.

Kai Hu received the graduation degree from the China University of Metrology, Hangzhou, China, the bachelor's and masters degrees from the Nanjing University of Information Science and Technology, Nanjing, China, and the Doctoral 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.