CMLFormer: CNN and Multiscale Local-Context Transformer Network for Remote Sensing Images Semantic Segmentation

Honglin Wu, *Member, IEEE*, Min Zhang, Peng Huang, and Wenlong Tang

Abstract—**The characteristics of remote sensing images, such as complex ground objects, rich feature details, large intraclass variance and small interclass variance, usually require deep learning semantic segmentation methods to have strong feature learning representation ability. Due to the limitation of convolutional operation, convolutional neural networks (CNNs) are good at capturing local details, but perform poorly at modeling long-range dependencies. Transformers rely on multihead self-attention mechanisms to extract global contextual information, but it usually leads to high complexity. Therefore, this article proposes CNN and multiscale local-context transformer network (CMLFormer), a novel encoder-decoder structured network for remote sensing image semantic segmentation. Specifically, for the features extracted by the lightweight ResNet18 encoder, we design a transformer decoder based on multiscale local-context transform block (MLTB) to enhance the ability of feature learning. By using a self-attention mechanism with nonoverlapping windows and with the help of multiscale horizontal and vertical interactive stripe convolution, MLTB is able to capture both local feature information and global feature information at different scales with low complexity. In addition, the feature enhanced module is introduced into the decoder to further facilitate the learning of global and local information. Experimental results show that our proposed CMLFormer exhibits excellent performance on the Vaihingen and Potsdam datasets.**

Index Terms—**Convolutional neural network (CNN), multiscale, remote sensing images, semantic segmentation, transformer.**

I. Introduction

With the rapid development of sensor technology and remote sensing platforms, high-resolution remote sensing images have been widely used in the fields of land cover [1], [2], city planning [3], change detection [4], [5], and scene classification [6], [7], [8]. However, high-resolution remote sensing images generally suffer from large intraclass variance and small interclass variance, which makes remote sensing image segmentation and detection tasks extremely challenging.

In recent years, with the development of deep learning, researchers began to apply convolutional neural networks (CNNs) to semantic segmentation. In this context, Long et al. [9] proposed the fully convolutional network (FCN), which replaces the traditional fully connected layers with convolutional layers. This innovative work has had a far-reaching impact on the field of semantic segmentation of remote sensing images. Subsequently, Ronneberger et al. [10] proposed a network model with an encoder-decoder structure, widely known as UNet. The structure introduces a skip connection, which further improves the segmentation accuracy. The encoder-decoder architecture shows great potential in image semantic segmentation tasks and is gradually becoming the dominant architecture of the field. Due to the influence of complex targets and rich features in remote sensing images, the pixel-level fine-grained classification tasks in remote sensing semantic segmentation usually require more comprehensive semantic information. Researchers have noticed that multiscale and attention mechanisms play a very important role in enhancing semantic representation and improving semantic segmentation performance [11], [12], [13], [14]. Chen et al. [11] proposed DeepLabv3+ to enhance multiscale feature representation by using atrous spatial pyramid pooling module with atrous convolution. In addition, Cheng et al. [13] proposed context aggregation network, which combines the attention mechanism with multiscale features to achieve higher positioning accuracy. Zhao et al. [14] proposed an end-to-end attention-based semantic segmentation network (SSAtNet), which introduces an attention mechanism into a multiscale module to refine the features, improving the accuracy of semantic segmentation. The introduction of multiscale and attention mechanisms greatly enhances feature learning and enables the model to comprehensively deal with scale changes and complex objects of the image, thereby improving the accuracy of semantic segmentation.

Despite the outstanding performance of CNNs in image processing, their limited receptive fields restrict the capability to model long-range contextual dependencies. Long-range contextual dependencies are particularly crucial for dense classification tasks, such as semantic segmentation. Recognizing the importance of context, researchers have begun to investigate the transformer of computer vision [15], [16], [17]. In 2021, Dosovitskiy et al. [15] proposed vision transformer (ViT), which shows potential application prospects in image classification tasks. Zheng et al. [16] proposed segmentation transformer (SETR), which uses transformers as the backbone pushing the

Manuscript received 13 October 2023; revised 24 December 2023 and 21 January 2024; accepted 6 March 2024. Date of publication 14 March 2024; date of current version 2 April 2024. This work was supported in part by the Scientific Research Fund of Hunan Provincial Education Department under Grant 21B0329, in part by Changsha Municipal Natural Science Foundation under Grant kq2208236, and in part by Postgraduate Scientific Research Innovation Project of Hunan Province under Grant CX20220939. *(Corresponding author: Honglin Wu.)*

The authors are with the School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410114, China (e-mail: honglinwu@csust.edu.cn; 21208051645@stu.csust.edu.cn).

Code is available online at https://github.com/DrWuHonglin/CMLFormer.

Digital Object Identifier 10.1109/JSTARS.2024.3375313

further development of transformers in the field of segmentation. Transformers rely on multihead self-attention mechanisms to model long-range dependencies. However, its computational complexity grows quadratically with the input data size, which usually imposes heavy computational overhead and limits its application. To address this issue, Liu et al. [18] proposed swin transformer based on shifted windows, which restricts attention computations to local windows and effectively reduces the computational complexity. Wang et al. [19] proposed pyramid vision transformer (PVT), which reduces the required key-value pairs in the traditional multihead self-attention mechanism and adopts pyramid structure to extract image features from multiple scales, thereby reducing the computational complexity associated with traditional transformer models.

In semantic segmentation tasks, both global and local information are crucial for accurately understanding the semantic structure of images. Researchers have begun to integrate CNN and transformer to fully leverage their respective advantages. Chen et al. [20] proposed TransUNet, a UNet-like architecture combining CNN and transformer, which showed excellent performance in medical image classification. He et al. [21] proposed ST-UNet, which uses a parallel dual encoder structure of CNN and swin transformer to extract local and global features, and integrates them through the relational aggregation module. Zhang et al. [22] proposed a segmentation network that combined the shift window-based swin transformer with a CNN-based decoder to establish a deep learning framework for semantic segmentation of remote sensing images.

Inspired by the aforementioned literature, this article proposes a new semantic segmentation architecture for remote sensing images called CMLFormer. CMLFormer adopts a hybrid architecture that combines CNN and transformer components. Specifically, we employ the lightweight ResNet-18 as the encoder and propose a multiscale local-context transformer block (MLTB). In contrast to traditional transformers, MLTB integrates attention mechanisms with multiscale strategies to enhance feature learning. To be precise, the self-attention mechanism is constrained within nonoverlapping windows to efficiently capture local contextual information with low complexity. In order to overcome the limitation of window for long-range modeling, we use different scales of depth-wise separable large convolutions instead of the feed forward network of the traditional transformer to obtain multiscale contextual information. Due to the high computational complexity of the large convolutions, we split the large convolutions into two depth-wise separable strip convolutions, taking into account that strip convolution has certain advantages for the recognition of strip targets such as cars or rivers in remote sensing images. Furthermore, a Feature Enhancement Module (FEM) is proposed to efficiently integrate features from global and local information to achieve more comprehensive information fusion in the channel and spatial dimensions. The main contributions of this article are as follows.

1) This article proposes a CMLFormer network architecture that utilizes a lightweight CNN and a multiscale local-context transformer in the encoder-decoder structure, effectively exploiting both global and local information in remote sensing image segmentation with less computational overhead.
2) An efficient and flexible MLTB is designed to capture the global context in remote sensing images with low computational complexity by combining the nonoverlapping block self-attention and multiscale strategies to establish long-range dependencies between pixels.
3) We introduce a FEM to mitigate the omission of local context details during long-range dependencies modeling in remote sensing image segmentation.

The rest of this article is organized as follows. Section II analyzes the work related to this article. In Section III, we describe the research methodology of this article. Section IV shows the robustness of the proposed model through experimental comparisons and ablation experiments. Finally, Section V concludes this article.

II. Related Work

A. Semantic Segmentation Methods Based On CNN

Semantic segmentation of remote sensing images aims to categorize individual pixels into their respective semantic classes, thus facilitating the automatic identification and segmentation of different objects, features or regions. In contrast to conventional pixel-based classification methods, CNN-based approaches are able to capture the spatial details and contextual relations within the images, thus increasing the accuracy and robustness of the semantic segmentation of remotely sensed imagery.

FCN [9] represents a significant milestone in the realm of semantic segmentation and lays the foundation for further model development. For example, UNet [10] introduced skip connections to recover detailed information at different scales. Various UNet-based variants followed. Wu et al. [23] proposed DenseUnet to combine dense connections into the encoder-decoder structure of UNet, which enables the model to better capture features. Diakogiannis et al. [24] proposed ResUnet to add residual connections into the encoder-decoder structure of UNet, which solves the problems of model training difficulty and gradient vanishing in order to better capture detailed information. To address the challenges posed by objects and scenes at different scales, ACNet [25] introduced an adaptive context module that enhances the segmentation of multiscale objects in remote sensing images by capturing context information at different scales at different levels. To overcome the limitation of receptive fields in CNNs, researchers employed attention mechanisms in the structure to model long-range dependencies. Fan et al. [26] proposed MA-Net, which adaptively combines local and global dependencies via positional attention block and multiscale fusion attention block to capture rich contextual and channel dependencies. Sun et al. [27] proposed a SPANet, which uses dual-branching to extract global and local information and fuses multiscale features through a successive pooling attention module to effectively alleviate the blurring of object boundary segmentation in remote sensing images. Huang et al. [28] proposed a CCNet based on a cross-attention mechanism to capture long-range correlations between pixels to more accurately identify the semantic information in remote sensing images.

B. Semantic Segmentation Methods Based On Transformer

Transformer is initially used in the field of natural language processing (NLP) and has achieved remarkable success in NLP [29]. Subsequently, transformers have been applied to the field of computer vision. ViT [15] is the first transformer-based model for image recognition. It divides the input image into fixed-sized blocks and then captures contextual information within the image using self-attention mechanism, ultimately producing feature representations for tasks, such as classification. ViT has achieved significant success in the field of computer vision and has become one of the crucial architectures in computer vision. With continuous efforts from researchers, the application of transformers in image processing has gradually matured. Inspired by ViT, the transformer-based model for semantic segmentation called SETR [16] was proposed. SETR utilizes transformers as the encoder in the semantic segmentation model and can be combined with other decoders to achieve more complex segmentation models. However, the core of the transformer is the self-attention mechanism, which requires computing similarity between each position and all other positions, leading to a quadratic increase in computational complexity with image resolution. Consequently, transformer-based models struggle to handle high-resolution images.

To address the above issues, extensive research has been conducted. Wang et al. [19] proposed PVT, which combines the traditional transformer model with a pyramid structure for dense prediction. Liu et al. [18] proposed swin transformer, which employs a hierarchical design with shifted windows and local self-attention confined within windows. Xie et al. [30] proposed a novel design of positionless coded hierarchical transformer encoder and lightweight all-MLP decoder named SegFormer. Dong et al. [31] proposed a Cswin transformer, which parallelly computes self-attention within horizontally and vertically striped cross-shaped windows, improving the efficiency of capturing global receptive fields.

C. Approach of Combining CNN and Transformer

CNN and transformer possess different advantages in the fields of image processing and NLP. CNN excels at extracting local features from images and adapts to objects of different scales through its translation invariance and feature reuse capabilities. On the other hand, transformer is particularly good at modeling global dependencies to capture long-range semantic dependencies in images when dealing with sequential data. Although CNNs and transformers each have advantages in different areas, recent research has shown that combining CNNs and transformers in image segmentation can fully exploit their strengths and improve the performance of the model. Wang et al. [32] proposed dual-branch hybrid CNN-transformer network (DBCT-Net), which fully exploits the advantages of CNN in local specific feature extraction, and achieves the modeling of global dependencies through transformer. Gao et al. [33] designed a dual-encoder model that uses independent CNN and transformer branches to extract features and adaptively fuse them to enhance the feature presentation of the model. Guo et al. [34] proposed CMT, a novel hierarchical transformer architecture that combines convolutional operations to capture local and global features, thereby improving performance and reducing computational cost. We propose a novel network architecture combining CNN and transformer, which uses a lightweight CNN and a multiscale local-context transformer in an encoder-decoder structure. The CMLFormer facilitates the comprehensive use of the global and local information in the image and enables the accurate segmentation of high-resolution remote sensing images at low computational cost.

III. Methodology

A. Overall Architecture

The CMLFormer is shown in Fig. 1(a), which follows encoder-decoder architecture. The encoder uses a lightweight ResNet-18 network architecture to extract image features with low complexity. ResNet-18 consists of four ResBlock stages, each of which double downsamples the feature map. The decoder consists of MLTB and FEM. The output of ResBlock4 of the encoder is fed into the MLTB. ResBlock3, ResBlock2, and ResBlock1 outputs are fed into the FEM, respectively. The MLTB is exploited to capture long-range dependencies with low complexity by adopting local-context multihead self-attention (LMSA) in nonoverlapping windows and multiscale stripe convolution (MSC). The outputs of the MLTB and encoder stages are simultaneously fed into the FEM, which is utilized to efficiently fuse global and local information in channel and spatial dimensions. Finally, the output features of the MLTB from the three stages are subjected to a summing operation and concatenated with the final output of the decoder in order to make full use of the feature information at each level and improve the accuracy of the semantic segmentation. The concatenated features are sent to the segHead for prediction. Specifically, through 3×3 convolution, 1×1 convolution and bilinear interpolation upsampling to obtain the final feature map.

B. Multiscale Local-Context Transformer Block

Global information is essential for accurate recognition of complicated categories in remote sensing images, such as buildings of different sizes, long stretches of rivers, and cars. Although transformers have great potential for integrating global information, traditional transformers may lead to degradation of model performance due to their high complexity. Therefore, we propose a low-complexity MLTB to efficiently extract global context information from image regions of different scales, as shown in Fig. 1(b). The MLTB consists of two key components, LMSA and MSC. The LMSA restricts attention mechanism to the local context through window partitioning and computes the attention matrix from the query, key and value, thus effectively capturing the relevance of local details in an image. In order to make up for the potential information loss, we introduce residual connection and combine 3 × 3 depth-wise separable convolution to enhance the output features. The MSC module effectively expands the multiscale perceptual capability

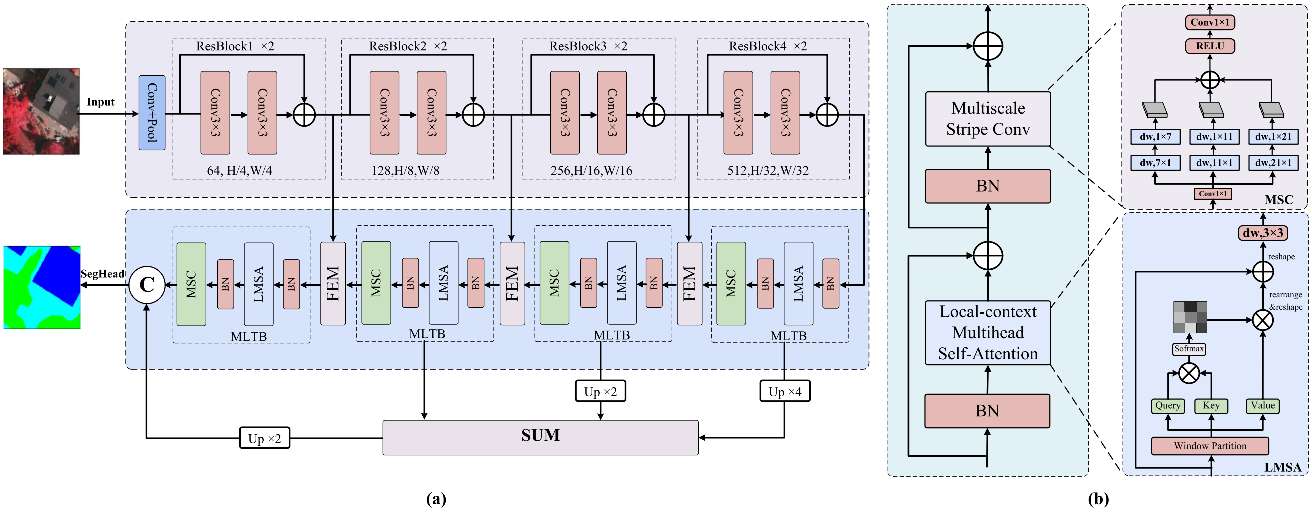

Fig. 1. Architecture of CMLFormer and MLTB. (a) Framework of CMLFormer proposed in this article consists of a ResNet-18 encoder and a decoder consisting of MLTB and FEM. ResNet-18 consists of four stages Resblock. (b) MLTB consists of LMSA and MSC. The LMSA contains window operations and multiple self-attention mechanisms. The MSC adopts strip depth convolution at three different scales.

of the model through multiscale horizontal and vertical depthwise separable stripe convolution operations to make up for the potential global information loss in MLTB. This strategy plays a key role in improving the sensitivity of the model to multiscale features and helps to improve the performance of semantic segmentation.

Specifically, in MLTB, the channel dimension of the input feature image $X \epsilon R^{(B\times C\times H\times W)}$ is first expanded tripled in size using 1×1 convolution, which maps the C dimension features into $3 \times C$ dimensions. Then, we separate the 1D sequence $\epsilon R^{(3\times B\times \frac{H}{ws}\times \frac{W}{ws}\times heads)\times(ws\times ws)\times \frac{C}{heads}}$ into the Query (Q), Key (K) and Value (V) vectors using window partition operations, where *heads* denote the number of attentional heads. We set both the number of attention heads and the window size *ws* to 8. The attention weights are computed based on Q, K, and V. Specifically, Q and K are dot-produced to obtain the raw attention scores, which are then normalized using the softmax function. Finally, the attention weights are applied to the V to obtain the attention output for each position, as shown in (1). In order to restore the features to their original size for effectively extracting multiscale information, we perform a rearrangement operation on the window-level feature representation followed by residual and reshape operations. The feature representation is further enhanced using 3×3 convolution before MLTB output. In MSC, multiscale information is obtained using stripe convolution with three different scale sizes (7,11,21) and a sum operation is performed to capture the detailed information at different scales. Finally, the multiscale information is nonlinearly augmented and dimensionally converted by RELU activation function and 1×1 convolution. The flow of MLTB is as (2), (3):

$$\text{Attention} = \text{soft max}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

$$X_1^t = \text{LMSA}\left(BN\left(X^{t-1}\right)\right) + X^{t-1} \tag{2}$$

$$X^t = \text{MSC}\left(BN\left(X_1^t\right)\right) + X_1^t \tag{3}$$

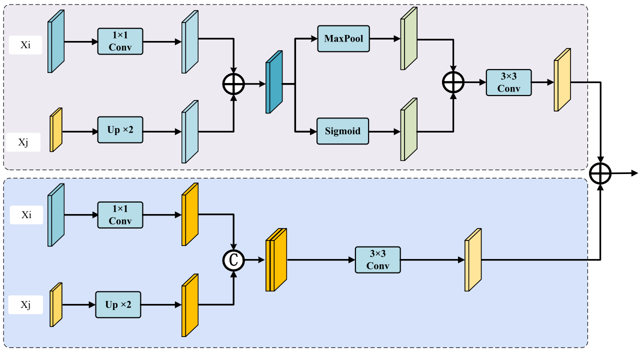

Fig. 2. Architecture of FEM. FEM contains two branches, the top branch obtains the spatial features by sum operation, and the bottom branch obtains the channel features by concatenation operation.

where d_k is the channel dimension of K, X^{t-1} represents the input features of LMSA, X_1^t represents the output features of LMSA, and X^t represents the output features of MSC.

C. Feature Enhanced Module

Local information and global information complement each other in semantic segmentation tasks. Local information helps the model capture local features, such as details, boundaries, and textures. The global information provides contextual semantic information and guides the model to consider overall consistency during the segmentation process. The effective use of local and global information can improve the segmentation performance of the model. However, most methods simply concatenate local information and global information, which can lead to information imbalance and impact the performance of model. To effectively integrate local information and global information, we construct FEM, as shown in Fig. 2. The FEM consists of two branches: the first branch obtains the spatial features by sum operation, and the second branch obtains the channel features by concatenation operation. Then, the channel features and spatial features are summed to effectively fuse global information and local information. The fused information considers the details information of the local area and the semantic information of the global scene to enhance the understanding and inference ability of the model for the target. By integrating information from both channel and spatial dimensions, the model can better incorporate contextual information.

Specifically, in one branch, the local context information from the encoder stage is added in the spatial dimension to the global information from the MLTB stage. Then, the sigmoid activation features are added to the Max pooling features at the pixel level. Immediately, a 3 × 3 convolution is used to enhance important contextual information and discard irrelevant features. In the other branch, the local contextual information from the encoder stage is concatenated with the global information from the MLTB stage in the channel dimension, and a 3 × 3 convolution is used to enhance the feature representation. Then, the enhanced spatial features and the enhanced channel features are aggregated using an addition operation. The FEM process is as follows:

$$X' = \mathrm{Conv}_{1\times 1}\left(X_i\right) + Up\left(X_j\right) \tag{4}$$

$$X_S = \mathrm{Conv}_{3\times 3}\left(\mathrm{MaxPool}\left(X'\right) + \mathrm{Sigmoid}\left(X'\right)\right) \tag{5}$$

$$X_C = \mathrm{Conv}_{3\times 3}\left(\mathrm{Concat}\left(\mathrm{Conv}_{1\times 1}\left(X_i\right),\ Up\left(X_j\right)\right)\right) \tag{6}$$

$$X = X_S + X_C \tag{7}$$

where X_i is the feature output from the encoder, X_j is the feature output from MLTB, X_S denotes the output of the spatial branch, and X_C denotes the output of the channel branch.

IV. Experimental Results

A. Datasets

In this section, the superiority of CMLFormer is evaluated on the Vaihingen and Potsdam datasets.

1) Vaihingen: The Vaihingen dataset, named after the Vaihingen region of Stuttgart, Germany, is a publicly available dataset of high-resolution aerial images. The dataset consists of high-resolution color aerial images captured by drones and ground truth labels for each image. Ground truth labels are usually used to indicate different categories in an image, such as buildings, roads, and trees. The dataset contains a total of 33 ortho-corrected images. In this article, we selected 16 specified images for training, while the remaining images were retained for testing. Each image is cut into small blocks of 256×256 to meet the experimental requirements.

2) Potsdam: The Potsdam dataset is constructed from aerial images located in Potsdam, Germany. It is a public resource widely used in computer vision and remote sensing image processing research. The dataset includes high-resolution aerial images covering different landscapes in urban areas, such as buildings, roads, and lawns. Each image is accompanied by a detailed ground truth label, covering different types of ground object information. In total, the Potsdam dataset contains 38 images that have been orthorectified. In this article, we selected 16 specific images for training, while the remaining 17 images were retained for testing. Each image is cropped to form a small image block with a size of 256×256 to meet the experimental needs.

B. Experimental Setup

Each experiment is performed on a single NVIDIA Tesla V100S GPU. We used the SGD optimizer for training. The weight decay is set to 0.0001. The initial learning rate is set to 0.01, and the learning rate is updated using the "poly" learning strategy, which is a polynomially decaying learning rate method commonly used to train deep learning models. Its update rule can be expressed as follows:

$$\mathrm{lr} = \mathrm{base\_lr} \times \left(1 - \frac{\mathrm{cur\_iter}}{\mathrm{max\_iter}}\right)^{\mathrm{power}} \tag{8}$$

where base_lr denotes the initial learning rate, cur_iter denotes the current number of iterations, max_iter denotes the maximum number of iterations, and power is a power exponent controlling the decay of the polynomial. In this article, power is set to 0.9. The training images are randomly flipped and cropped for data augmentation.

As shown in formula (9), the construction of the loss function follows the weighted combination of cross-entropy and dice loss [41]. We set α to 0.4 in this article. To evaluate our results, we rely on two main evaluation metrics: mean intersection over union (mIoU) and mean F1 score (mF1). The specific formula can be seen in the following:

$$\mathrm{Loss} = \alpha \cdot \mathrm{CrossEntropyLoss} + (1 - \alpha) \cdot \mathrm{DiceLoss} \tag{9}$$

$$\mathrm{Precision} = \frac{1}{M} \sum_{j=1}^{M} \frac{\mathrm{TP}_j}{\mathrm{TP}_j + \mathrm{FP}_j} \tag{10}$$

$$\mathrm{Recall} = \frac{1}{M} \sum_{j=1}^{M} \frac{\mathrm{TP}_j}{\mathrm{TP}_j + \mathrm{FN}_j} \tag{11}$$

$$\mathrm{F1} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{12}$$

$$\mathrm{mIoU} = \frac{1}{M} \sum_{j=1}^{M} \frac{\mathrm{TP}_j}{\mathrm{TP}_j + \mathrm{FP}_j + \mathrm{FN}_j} \tag{13}$$

where FN_j, FP_j, and TP_j represent false negatives, false positives, and true positives, respectively, for object indexed as class j.

C. Semantic Segmentation Results and Analysis

1) Results on the Vaihingen Dataset: The effectiveness of the CMLFormer on the Vaihingen dataset is illustrated by the quantitative metrics presented in Table I. In particular, the CMLFormer scores 84.18% for mF1 and 73.07% for mIoU. The methods we compared include CNN-based approaches, such as FCN, DeeplabV3+, DANet, PSPNet, and MANet. The transformer-based method we compared is SwinUnet. In addition, to thoroughly validate the effectiveness of the CMLFormer, we also

TABLE I
QUANTITATIVE COMPARISON RESULTS ON THE VAIHINGEN DATASET

Method	F1%					mIoU%	mF1%	Param./M	FLOPS/G
	Imp.surf.	Building	Low.veg.	Tree	Car				
FCN [9]	85.86	89.34	75.62	83.72	75.33	69.84	81.79	19.17	25.54
DeeplabV3+ [11]	87.74	91.55	76.80	85.01	73.99	71.52	83.02	59.34	65.27
DANet [35]	85.92	90.79	75.50	83.57	60.48	66.85	79.25	47.44	13.85
BANet [36]	86.99	91.01	76.05	84.09	72.05	70.14	82.04	**3.25**	**12.73**
PSPNet [37]	84.33	88.03	74.50	83.33	58.38	64.71	77.71	52.52	50.51
MANet [38]	86.19	89.67	75.68	84.42	76.44	70.56	82.48	35.86	77.75
SwinUnet [39]	86.27	89.01	76.22	84.30	71.26	69.17	81.41	41.34	68.34
TransUnet [20]	87.75	91.74	77.21	84.56	76.86	72.29	83.62	100.44	35.84
UnetFormer [40]	87.54	91.34	76.43	84.77	77.42	72.10	83.50	11.48	16.99
CMLFormer(ours)	**88.01**	**91.90**	**77.26**	**84.87**	**78.82**	**73.07**	**84.18**	11.81	21.25

Bold indicates the best results. The category metrics are F1 scores.

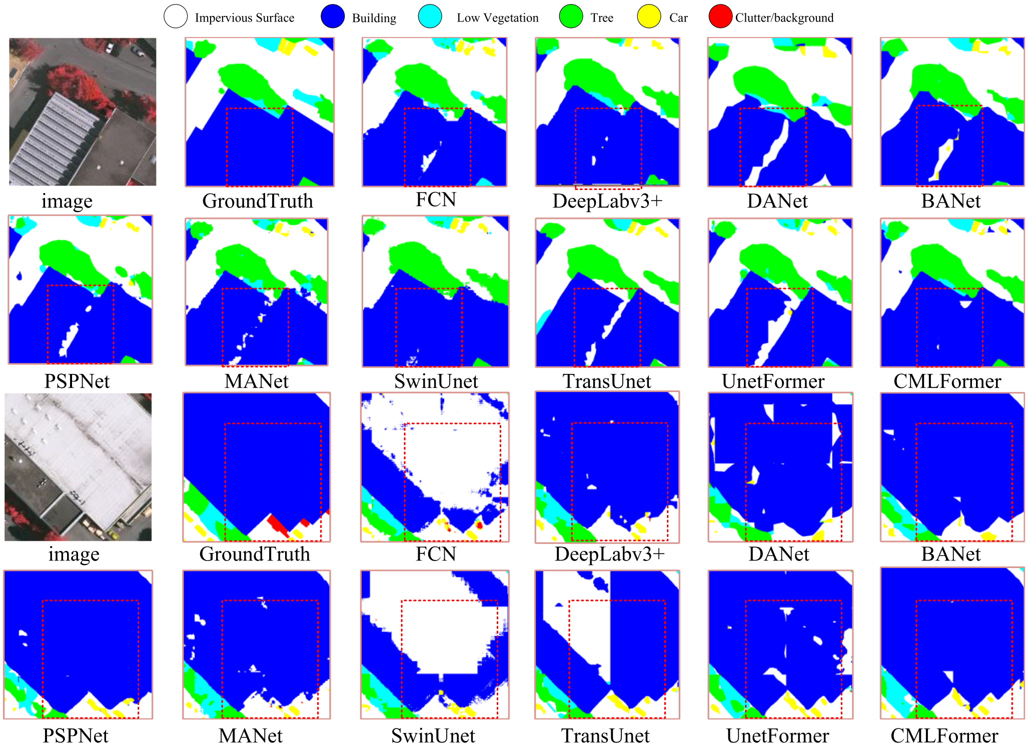

Fig. 3. Visualization results of the Vaihingen dataset. The main differences are highlighted by the red boxes in the figure.

compared it against some hybrid CNN and transformer methods, namely, UnetFormer, TransUnet, and BANet. From Table I, it can be observed the CMLFormer outperforms the majority of both CNN-based and transformer-based frameworks. As CMLFormer is an architecture combining CNN and transformer, the comparison results with UnetFormer, TransUnet, and BANet are more meaningful.

We visualize the comparative results in CNN-based methods and show that the CMLFormer is effective in alleviating the problem of high intraclass variance caused by occlusions compared with other models. For example, in the red box areas of the first and second rows in Fig. 3, the occlusion of light and shadows creates a gap between the taller building and the neighboring lower building, leading to inaccurate recognition of the building by other models. In contrast, CMLFormer makes an accurate recognition in these cases. In the red box areas of the third and fourth rows, other models exhibit less significant performance in obtaining global contextual information, while CMLFormer effectively mitigates this issue and performs exceptionally well in extracting large targets.

2) Results on the Potsdam Dataset: CMLFormer is compared with the mainstream segmentation methods, as given in Table II. CMLFormer achieves excellent performance with mIoU of 80.06% and mF1 of 88.79%. This is a significant improvement over the other methods. Among the traditional CNN models, Deeplabv3 + performed well, and the mIoU and mF1 of CMLFormer increased by 0.99% and 0.64%, respectively. It is worth noting that the model parameters of CMLFormer represent only 12% of the TransUnet parameters, but mIoU and mF1 still exceed TransUnet by 0.3% and 0.21%, achieving the most advanced performance.This result is clearly shown in the visual effect in Fig. 4.

TABLE II
QUANTITATIVE COMPARISON RESULTS ON THE POTSDAM DATASET

Method	F1%					mIoU%	mF1%	Param./M	FLOPS/G
	Imp.surf.	Building	Low.veg.	Tree	Car				
FCN [9]	89.76	93.64	82.95	83.12	89.82	78.60	87.86	19.17	25.54
DeeplabV3+ [11]	90.40	**95.01**	83.82	**84.22**	87.32	79.07	88.15	59.34	65.27
DANet [35]	89.35	94.71	82.44	84.18	82.71	76.81	86.68	47.44	13.85
BANet [36]	90.36	94.00	83.39	83.04	89.49	78.92	88.06	**3.25**	**12.73**
PSPNet [37]	89.45	93.38	82.13	80.00	89.36	77.12	86.86	52.52	50.51
MANet [38]	90.10	94.38	83.11	83.42	89.85	79.12	88.17	35.86	77.75
SwinUnet [39]	90.30	94.41	83.05	82.83	88.84	78.68	87.89	41.34	68.34
TransUnet [20]	90.52	94.68	83.87	83.58	90.24	79.76	88.58	100.44	35.84
UnetFormer [40]	90.14	94.44	83.15	83.16	90.07	79.16	88.19	11.48	16.99
CMLFormer(ours)	**90.79**	94.96	**84.01**	83.91	**90.31**	**80.06**	**88.79**	11.81	21.25

Bold indicates the best results. The category metrics are F1 scores.

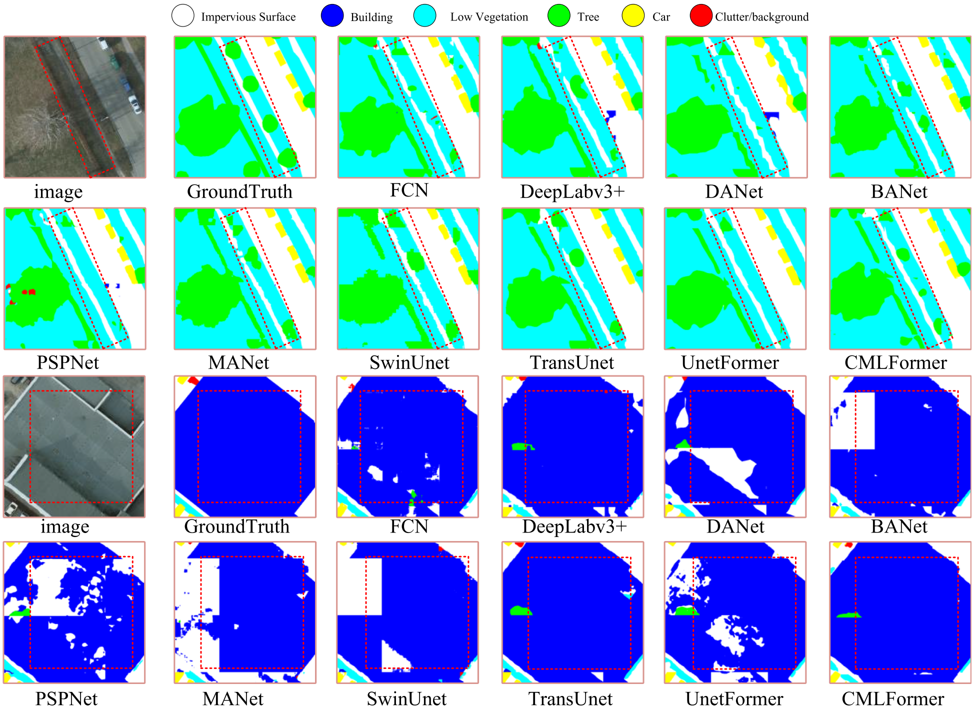

Fig. 4. Visualization results of the Potsdam dataset. The main differences are highlighted by the red boxes in the figure.

On the Potsdam dataset, the visual comparison results in Fig. 4 show several aspects. In the first two rows, there is some semantic overlap due to the similarity between low vegetation and trees, which poses a challenge for accurate classification by the models. When dealing with such similar and overlapping categories, CMLFormer excels in the clarity of the segmentation boundary. In addition, the size and shape of the buildings in the image vary greatly, increasing the diversity within the building category. As shown in the third and fourth rows in the red box area, CMLFormer performs excellently in extracting buildings, successfully reducing misclassification, boundary blurring, and omission issues, demonstrating outstanding performance.

3) Ablation Study: Comprehensive experiments are conducted on the Vaihingen dataset to verify the effects of MLTB and FEM. Since the CMLFormer uses ResNet-18 as the backbone network, we choose to connect different stages of ResNet-18 as the baseline. The mIoU and mF1 scores are used to assess the performance of each module, and the parameters are used to reflect the degree of lightness. As given in Table III, the results of ResNet-18 + MLTB + FEM verify the effectiveness of each module.

1) *Effectiveness of MLTB:* As given in Table III, when MLTB was incorporated into the CMLFormer framework, the mIoU and mF1 increased by 4.04% and 3.56%, respectively. Notably, the segmentation accuracy of the “car” class shows the most significant improvement, with a remarkable F1 increase of 12.8%. Among the other four “impervious surface” classes, there are F1 improvements of 1.6%. The “building” class also exhibits a notable F1 improvement of 1.45%, while the “low vegetation” class demonstrated a considerable F1 improvement of 1.32%. In addition, the “tree” class shows a substantial F1 increase of 0.62%. From the images in the first row of Fig. 5, it can be seen that the model performed poorly

TABLE III
QUANTITATIVE COMPARISONS AMONG ABLATION STUDIES ON THE VAIHINGEN DATASET

Method	Module	F1%							
	SUM	Imp.surf.	Building	Low.veg.	Tree	Car	mIoU%	mF1%	Param./M
ResNet-18(baseline)	✗	85.92	90.30	74.83	83.60	65.20	67.52	79.97	11.42
baseline + MLTB	✗	87.52	91.75	76.15	84.22	78.01	71.56	83.53	11.65
baseline + MLTB + FEM	✗	87.63	91.48	76.54	84.60	78.41	72.42	83.73	11.77
baseline + MLTB + FEM	✓	**88.01**	**91.90**	**77.26**	**84.87**	**78.82**	**73.07**	**84.18**	11.81

The bold values indicate the best results.

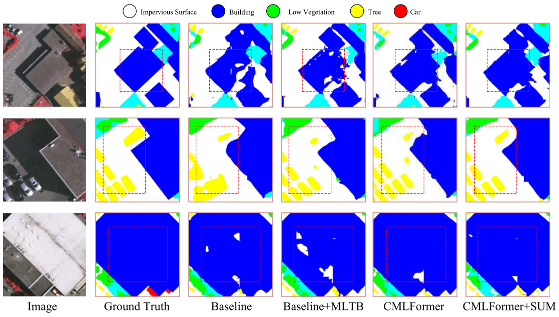

Fig. 5. Ablation experiment on the vaihingen datasets. Main differences are highlighted by the red boxes in the figure.

in recognizing building edges before the introduction of MLTB. With the incorporation of MLTB, the model mitigates the phenomenon of blurry segmentation edges.

2) *Effectiveness of FEM:* Table III reflects the joint effect of studying the two modules under the CMLFormer framework. After incorporating FEM, the mIoU and mF1 increase by 0.86% and 0.2%, respectively. In particular, there was an improvement of 0.4%, 0.11%, 0.4%, and 0.38% in the F1 of "car," "impervious surface," "low vegetation," and "tree." From the images in the second row of Fig. 5, it can be seen that under the influence of sunlight, the shadows cast by the "building" completely obscured the cars, posing a significant challenge to model recognition. In addition, the proximity of cars result in intraclass shadow occlusion. Before introducing FEM, the model struggles to recognize buildings and objects obscured by shadows. The model accurately segments the edges of small objects in dense areas and effectively mitigates the negative impact of shadows on car segmentation after introducing FEM.

3) *Effectiveness of SUM:* The effect of with-SUM-CMLFormer is seen in Table III. After adding SUM, mIoU and mF1 are improved by 0.65% and 0.45%, respectively. Specifically, the F1 of "car," "impervious surface," "building," "low vegetation," and "tree" were improved by 0.41%, 0.38%, 0.42%, 0.72%, and 0.27%, respectively. Fig. 5 shows that the model improves the acquisition of information about building edges. Furthermore, by including SUM, the model successfully reduces the negative effect of shadows on car segmentation.

V. CONCLUSION

In this article, we propose an innovative decoder based on transformer architecture and construct a CMLFormer for remote sensing images semantic segmentation. Considering the importance of global and local information in remote sensing image segmentation, we design a MLTB that integrates attention mechanisms with multiscale strategies to effectively exploit global information with lower computational cost, and develop a FEM to compensate for the omission of local context details in the MLTB. Extensive comparison and ablation experiments on the Vaihingen and Potsdam datasets demonstrate the effectiveness of CMLFormer. Although CMLFormer achieved excellent overall performance on both datasets, on the Potsdom dataset the building class and tree class are not optimal compared with the other compared methods. We are committed to analyzing the causes of this problem in future research, including possible data characteristics, lack of fit in the model structure, and possible overfitting or underfitting. In the meantime, we will continue to work on making CMLFormer more lightweight to improve its segmentation performance on large datasets. This may involve further simplifying the model structure, optimizing parameter settings, and adopting advanced light-weighting techniques.

REFERENCES

[1] D. He, Y. Zhong, and L. Zhang, "Spectral–spatial-temporal map-based sub-pixel mapping for land-cover change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 1696–1717, Mar. 2020.

[2] M. Zhang, W. Li, X. Zhao, H. Liu, R. Tao, and Q. Du, "Morphological transformation and spatial-logical aggregation for tree species classification using hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5501212.

[3] X. Huang, Q. Li, H. Liu, and J. Li, "Assessing and improving the accuracy of globeland30 data for urban area delineation by combining multisource remote sensing data," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1860–1864, Dec. 2016.

[4] K. Lim, D. Jin, and C.-S. Kim, "Change detection in high resolution satellite images using an ensemble of convolutional neural networks," in *Proc. Signal Inf. Process Assoc. Annu. Summit Conf. APSIPA Asia Pac.*, 2018, pp. 509–515.

[5] Q. Guo, J. Zhang, T. Li, and X. Lu, "Change detection for high-resolution remote sensing imagery based on multi-scale segmentation and fusion," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 1919–1922.

[6] J. Wang, W. Li, M. Zhang, R. Tao, and J. Chanussot, "Remote-sensing scene classification via multistage self-guided separation network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5615312.

[7] J. Wang, W. Li, Y. Wang, R. Tao, and Q. Du, "Representation-enhanced status replay network for multisource remote-sensing image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 28, 2023, doi: 10.1109/TNNLS.2023.3286422.

[8] J. Wang, W. Li, M. Zhang, and J. Chanussot, "Large kernel sparse convnet weighted by multi-frequency attention for remote sensing scene understanding," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Nov. 2023, Art. no. 5626112.

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

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Computer-Assisted Interv.*, 2015, pp. 234–241.

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

[12] W. Song, Y. Dai, Z. Gao, L. Fang, and Y. Zhang, "Hashing-based deep metric learning for the classification of hyperspectral and lidar data," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Oct. 2023, Art. no. 5704513.

[13] W. Cheng, W. Yang, Y. Pan, H. Guo, and Y. Cheng, "Context aggregation network for semantic labeling in aerial images," in *Proc. IEEE Int. Conf. Image Process*, 2019, pp. 4484–4488.

[14] Q. Zhao, J. Liu, Y. Li, and H. Zhang, "Semantic segmentation with attention mechanism for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Aug. 2022, Art. no. 5403913.

[15] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Representations*, 2021.

[16] 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.

[17] Z. Gao et al., "Joint learning of semantic segmentation and height estimation for remote sensing image leveraging contrastive learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jul. 2023, Art. no. 5614015.

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

[19] 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.

[20] J. Chen et al., "Transunet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:abs/2102.04306* .

[21] 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, Jan. 2022, Art. no. 4408715.

[22] C. Zhang, W. Jiang, Y. Zhang, W. Wang, Q. Zhao, and C. Wang, "Transformer and cnn hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4408820.

[23] Y. Wu, J. Wu, S. Jin, L. Cao, and G. Jin, "Dense-u-net: Dense encoder-decoder network for holographic imaging of 3D particle fields," *Opt. Commun.*, vol. 493, 2021, Art. no. 126970.

[24] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 94–114, 2020.

[25] X. Ding, Y. Guo, G. Ding, and J. Han, "Acnet: Strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1911–1920.

[26] T. Fan, G. Wang, Y. Li, and H. Wang, "Ma-net: A multi-scale attention network for liver and tumor segmentation," *IEEE Access*, vol. 8, pp. 179656–179665, 2020.

[27] L. Sun, S. Cheng, Y. Zheng, Z. Wu, and J. Zhang, "Spanet: Successive pooling attention network for semantic segmentation of remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4045–4057, May 2022.

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

[29] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 6000–6010.

[30] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 12077–12090.

[31] X. Dong et al., "Cswin transformer: A general vision transformer backbone with cross-shaped windows," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12124–12134.

[32] Q. Wang, X. Jin, Q. Jiang, L. Wu, Y. Zhang, and W. Zhou, "Dbct-net:a dual branch hybrid cnn-transformer network for remote sensing image fusion," *Expert Syst. with Appl.*, vol. 233, 2023, Art. no. 120829.

[33] L. Gao et al., "Stransfuse: Fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10990–11003, Oct. 2021.

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

[35] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3141–3149.

[36] L. Wang, R. Li, D. Wang, C. Duan, T. Wang, and X. Meng, "Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images," *Remote Sens.*, vol. 13, 2021, Art. no. 3065.

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

[38] R. Li et al., "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2022, Art. no. 5607713.

[39] H. Cao et al., "Swin-unet: Unet-like pure transformer for medical image segmentation," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2023, pp. 205–218.

[40] L. Wang et al., "Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 190, pp. 196–214, 2022.

[41] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc IEEE 4th Int. Conf. 3D Vis. (3DV)*, 2016, pp. 565–571.

Honglin Wu (Member, IEEE) received the B.E. degree in communication engineering and the Ph.D. degree in information and communication engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2004 and 2012, respectively.

He is currently an Associate Professor with the School of Computer and Communication Engineering, University of Science and Technology, Changsha, China. His current research interests include artificial intelligence, computer vision, and remote sensing image processing.

Min Zhang received the bachelor's degree in software engineering from the Qingdao University of Technology, Qingdao, China, in 2021. She is currently toward the M.S. degree in computer technology with the Changsha University of Science and Technology, Changsha, China.

She is proficient in Python programming language. Her research focuses on semantic segmentation of remote sensing imagery.

Peng Huang received the B.E. degree in communication engineering from the Hubei University of Arts and Science, Xiangyang, China, in 2021. He is currently working toward the M.S. degree with the communication and information system, Changsha University of Science and Technology, Changsha, China.

His research interests include deep learning, computer vision, and remote sensing image processing.

Wenlong Tang received the bachelor's degree in computer science and technology from Central South University for Nationalities, Wuhan, China, in 2021. He is currently working toward the M.S. degree in computer technology with the Changsha University of Technology, Changsha, China.

His research interests include semantic segmentation of remote sensing images.