



# MCAT-UNet: Convolutional and Cross-Shaped Window Attention Enhanced UNet for Efficient High-Resolution Remote Sensing Image Segmentation

Tao Wang , Chao Xu , *Member, IEEE*, Bin Liu , *Member, IEEE*, Guang Yang , Erlei Zhang , *Member, IEEE*, Dangdang Niu , and Hongming Zhang , *Member, IEEE*

**Abstract**—Semantic segmentation is a crucial step in the intelligent interpretation of high-resolution remote sensing images (HRSIs). Convolutional neural networks and transformers are widely used for semantic feature extraction in remote sensing images, but the former inevitably has limitations in modeling long-range spatial dependency information, while the latter lacks the ability to learn local semantic features. Existing remote sensing image segmentation methods are optimized and modified based on the backbone networks used in natural image processing. Despite achieving relatively good results, the complexity of their network structures leads to high computational costs and limited improvements in accuracy. These methods have limited boundary distinction for ground objects in complex environments, especially for small targets. In this article, we propose an efficient semantic segmentation architecture for HRSIs called MCAT-UNet, which utilizes multiscale convolutional attention (MSCA) and the cross-shaped window transformer (CSWT) to reconstruct UNet. The encoder stacks a sequence of MSCA to exploit the advantages of convolution attention to encode context information more effectively and enhance hierarchical multiscale representation learning. The proposed U-shaped decoder integrates three skip connections using the CSWT block to further capture long-range spatial dependency and gradually restore the size of the feature map. We benchmark MCAT-UNet on three common datasets, Potsdam, Vaihingen, and LoveDA. Comprehensive experiments and extensive ablation studies show that our proposed MCAT-UNet outperforms previous state-of-the-art methods with remarkable performance.

**Index Terms**—Convolutional attention, cross-shaped self-attention, remote sensing image, semantic segmentation, transformer.

Manuscript received 15 March 2024; revised 11 April 2024; accepted 25 April 2024. Date of publication 7 May 2024; date of current version 14 May 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 423377341 and in part by the Key Research and Development Program of Shanxi Province under Grant 2023-ZDLNY-69. (*Corresponding author: Hongming Zhang.*)

Tao Wang is with the College of Information Engineering, Northwest A&F University, Yangling 712100, China, also with the College of Information Engineering, Tarim University, Alaer 843300, China, and also with the Key Laboratory of Tarim Oasis Agriculture (Tarim University), Ministry of Education, Alaer 843300, China (e-mail: tao\_wang@nwafu.edu.cn).

Chao Xu, Bin Liu, Guang Yang, Erlei Zhang, Dangdang Niu, and Hongming Zhang are with the College of Information Engineering, Northwest A&F University, Yangling 712100, China (e-mail: cxu@nwafu.edu.cn; liubin0929@nwsuaf.edu.cn; cie\_yg@nwafu.edu.cn; erlei.zhang@nwafu.edu.cn; niudd@nwafu.edu.cn; zhm@nwsuaf.edu.cn).

The code and models used are available at <https://github.com/wujiang0156/MCAT-UNet>.

Digital Object Identifier 10.1109/JSTARS.2024.3397488

## I. INTRODUCTION

IN THE geomatics community, remote sensing images have become increasingly important for earth observation since they are easy to access, can be obtained in real time, and contain abundant spatial details and rich potential semantic content. It is highly important in many remote sensing applications such as agricultural planning [1], land cover [2], land change [3], climate change [4], disaster monitoring [5], and deforested regions [6]. This is a challenging task because some land covers that characterize a given class may have large variability, objects from the same category normally present different shape layouts, and class distributions at different locations in the HRSIs [7]. Moreover, due to the complex background environment, objects belonging to different categories can have similar appearances, making them difficult to identify, especially for small targets, which severely impacts the performance of semantic segmentation networks [8], [9], [10]. To cope with those limitations, it is essential to obtain strong semantic representations at both the global context level and the local level.

As an effective method for extracting hierarchical multiscale visual features from images, convolutional neural networks (CNNs) are the most common neural networks [8], [11], [12], [13]. For semantic segmentation, fully convolutional networks (FCNs) [14] achieve impressive results on challenging segmentation benchmarks. FCN and its variants employ a convolutional encoder-decoder architecture. The typical U-shaped network, the UNet architecture [15], introduces a symmetric encoder-decoder structure with skip connections to enhance detail retention. Many variants of UNet have been proposed in subsequent studies to further improve segmentation performance [16], [17], [18]. However, long-range spatial dependency is limited by the locality property of CNN-based methods. Many approaches attempt to enlarge the receptive field of CNNs. DeepLab [19] and Dilation [20] introduce atrous convolution. Chen et al. [21] adopted an atrous spatial pyramid pooling module, which probes convolutional features at multiple scales to capture long-range context. PSPNet [22] designs a pyramid pooling module to capture global context information that contains multiscale information. DANet [23] introduces dual attention modules to cap-

ture global dependencies in the spatial and channel dimensions. CFNet [24] and OCNet [25] consider the relations between the pixels and aggregate the representations of the contextual pixels. CCNet [26] uses two consecutive criss-cross attention modules to aggregate contextual information in the horizontal and vertical directions. K-Net [27] splits an image into different groups with learned static kernels and then iteratively improves these kernels and their partitioning of the image by the features assembled from their split groups. Although these methods have improved feature representation, the limited receptive fields of convolution kernels are unable to learn global context information, which is essential for dense prediction tasks. Moreover, these methods are proposed based on natural image processing, and whether they can be applied to remote sensing image processing needs further verification.

Recently, vision transformers have shown great potential in various computer vision tasks because of their ability to model long-range dependencies using self-attention mechanisms. Transformers can outperform standard CNNs by a significant margin, such as classification, segmentation, and object detection [28], [29], [30], [31]. Motivated by this, many researchers in the remote sensing field have applied transformers for remote sensing image semantic segmentation. DC-Swin [32] combines the Swin transformer and densely connected feature aggregation module to extract multiscale relation-enhanced semantic features for precise segmentation. Wang et al. [33] proposed a new rotated varied-size attention mechanism to extract rich context from generated diverse windows. ST-Unet [34] constitutes a novel dual encoder structure of the Swin transformer and CNN in parallel to obtain more discriminative features. CG-Swin [35] introduces the Swin transformer as the encoder and designs a class-guided Transformer block to construct the decoder. Zhang et al. [36] used Swin as the backbone to extract features, obtained multiscale context information from depth-separable products, and used a U-decoder to gradually restore the size of feature maps. Despite the success of these approaches, they enlarge the receptive field quite slowly, and a great number of blocks need to be stacked to achieve global self-attention. Moreover, they are limited in modeling local visual structures and scale-invariant representations, and the completeness of object boundaries and the precise identification of small objects are still insufficient.

Inspired by Guo et al. [37] and Dong et al. [30], we apply convolutional attention and cross-shaped window self-attention to reconstruct the U-shaped encoder–decoder network, which can model long-range spatial dependency and extract local representations effectively, and capture multiscale targets more accurately, greatly reducing time and memory complexity. We conduct extensive experiments on various challenging remote sensing semantic segmentation datasets, i.e., LoveDA, Potsdam, and Vaihingen. Our proposed MCAT-UNet is superior to multi-scale context schemes such as PSPNet [22] and UPerNet [38]; recent relational context schemes such as CCNet [26], DANet [23], and OCRNet [39]; and recent remote sensing image semantic segmentation methods such as EMRT [40], UNetFormer [41], ST-Unet [34], and DC-Swin [32]; moreover, its efficiency has improved.

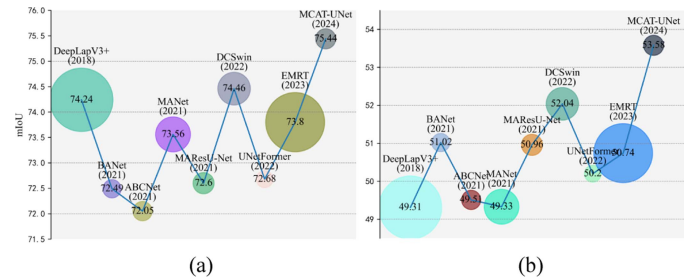


Fig. 1. Comparison with previous methods in terms of performance and efficiency on Potsdam and LovDA dataset. The size of each bubble corresponding to the FLOPs of the variant segmentation methods, the larger the bubble, the higher the computational complexity. MCAT-UNet achieves state-of-the-art performance with a 75.44% and 53.58% mIoU on the Potsdam and LoveDA dataset, significantly reducing computational overhead while maintaining a competitive performance. (a) On potsdam. (b) On LoveDA.

Our main contributions are summarized as follows.

- 1) By utilizing multiscale convolutional attention (MSCA) and the cross-shaped window transformer (CSWT), we have reconstructed a more efficient and accurate U-shaped encoder and decoder architecture. It can extract local representations and capture long-range spatial dependencies, enabling more effective segmentation of the geo-objects in complex scenes.
- 2) The proposed U-shaped decoder performs three skip connections between MSCA and CSWT, further capturing long-range spatial dependencies, segmenting small targets, significantly improving segmentation results, and without the need for complex and computationally demanding modules.
- 3) We comprehensively compared the effects of several of the most popular semantic segmentation methods of remote sensing images and natural images. As shown in Fig. 1, the proposed MCAT-UNet significantly outperforms the existing alternative approaches and achieves great performance on three challenging datasets: the LoveDA, Potsdam, and Vaihingen datasets.

## II. RELATED WORK

Semantic segmentation aims to partition an image into several visually meaningful or interesting regions for visual understanding according to semantic information. Since FCNs [14] have made great progress in image semantic segmentation, CNNs [15], [21], [22], [42], [43], [44] have achieved great success and become the mainstream framework in computer vision. Recently, transformer-based methods [45], [46], [47], [48], [49] have shown the great potential of attention-based models and achieved significantly better performance than CNN-based methods on different visual tasks. We start this section by reviewing CNN-based methods and Transformer-based methods for computer vision tasks. Then, we turn our focus to a review of remote sensing image segmentation using deep neural networks.

### A. CNN-Based Methods in Computer Vision Tasks

Over the past decade, CNNs have dominated vision architectures in the computer vision field since AlexNet was first

proposed [50]. CNNs are naturally equipped with the intrinsic inductive bias of scale invariance and locality to greatly improve the effectiveness of neural networks [22], [43], [51], [52]. Deeper and more effective convolutional neural architectures, e.g., ResNet [51], DenseNet [53], EfficientNet [54], and MobileNetV3 [55], have been proposed to further improve computer vision performance.

Encoder–decoder CNN-based methods have achieved encouraging performance, but CNN-based segmentation networks with limited receptive fields can extract only local semantic features and lack the capability to model global information from whole images. Recently, CNN-based methods [37], [56], [57], [58], such as PoolFormer, ConvNeXt, VAN, and SegNeXt, have been shown to perform comparably to transformer-based methods with proper design while retaining simplicity and efficiency. ConvNeXt [57] is constructed entirely from pure ConvNet modules and competes favorably with Transformers in terms of accuracy and scalability. PoolFormer [56] replaces the attention module in Transformers with a simple spatial pooling operator to conduct only basic token mixing. VAN [58] leverages the large-kernel attention mechanism, which absorbs the advantages of convolution and self-attention to build both channel and spatial attention. By using a cheaper and larger multiscale convolutional attention module to evoke spatial attention, SegNeXt [37] showed that convolutional attention is a more efficient and effective way to encode multiscale contexts from local to global levels than both standard convolutions and self-attention in spatial information encoding.

### B. Transformer-Based Methods in Computer Vision Tasks

Transformers have been applied with notable success in fundamental computer vision tasks such as image recognition [29], [59], object detection [31], [60], image segmentation [61], [62], and video captioning [63], [64]. Vision transformer [29] represents the first attempt to apply the transformer structure to image tasks. In semantic segmentation, SETR [61] adopts ViT as a backbone to demonstrate the feasibility of using transformers. Segmenter [46] adopts ViT [29] as a backbone and incorporates a transformer-based decoder generating class masks to improve the performance. Transformer architectures are good at establishing global relations but are less robust at extracting local information and handling tiny objects. The quadratic complexity of full-attention is too expensive for high-resolution images, which seriously affects its potential and feasibility for remote sensing image-related real-time applications.

To address these limitations, some efficient approaches have been proposed. For example, SegFormer [45] proposed a novel hierarchically structured transformer encoder and a simple multilayer perceptron (MLP) decoder to render powerful representations. ResT [65] designed a multihead self-attention module to reduce the computational cost. Swin transformer [28] proposed shifted windows to increase efficiency by limiting self-attention computations to nonoverlapping local windows while also allowing for cross window connections. Wang et al. [66] proposed a pyramid vision transformer (PVT), which achieved considerable improvements over its ResNet [51] counterpart in semantic segmentation. Beit [67] proposed a masked image modelling task to pretrain vision Transformers in a self-supervised

manner. Twins [68] presented two powerful vision transformer backbones called Twins-PCPVT and Twins-SVT; the former explored the applicability of conditional positional encodings in PVTs, and the latter revisited the current attention design to propose a more efficient attention paradigm. Moreover, CSWin [30] developed the cross-shaped window self-attention mechanism for computing self-attention in the horizontal and vertical stripes in parallel, which introduces no extra computational cost while enlarging the receptive field for computing self-attention.

### C. Remote Sensing Image Segmentation

Modern satellite imagery provides higher resolution than traditional satellite images, which allows collected images to more closely match the actual scene and increases the number of multiscale and multiclass objects. Consequently, semantic segmentation of remote sensing images faces the challenges of large intraclass and small interclass variances in the pixel values of objects of interest. To this end, CNN-based models and Transformer architecture-based models are improved upon to capture important edge, shape and textural features.

FCNs [14] and their variants have become popular solutions for remote sensing image segmentation and have performed well on numerous datasets. BANet [69] and ABCNet [70] capture abundant details and global context information through two branches, where the spatial path is a simple convolution stack and can only obtain limited spatial information. MANet [71] integrates different levels of semantic information through a multiscale strategy and combines the self-attention module to hierarchically aggregate relevant contextual features. MAREsUNet [17] reconstructed the skip connections in the raw UNet based on ResNet and the proposed linear attention mechanism and improved the classification accuracy and computational efficiency. UNetFormer [41] constructed a UNet-like transformer and developed an efficient global-local attention mechanism to model both global and local contextual information in a decoder for real-time urban scene segmentation. These methods have made significant progress, but CNN-based backbones are unable to model long-distance dependencies due to their limited receptive fields.

Recently, researchers have begun to actively explore the application of transformers to improve remote sensing image semantic segmentation. EMRT [40] adopts the deformable self-attention mechanism in transformer to achieve the context on multiscale feature maps. MMT [72] proposes a mixed-mask attention mechanism to learn more explicit intraclass and interclass correlations by capturing long-range interdependent representations and solves the problem of large-scale-varied targets in remote sensing images. AerialFormer [73] created a hybrid model that incorporates a transformer encoder with a multidilated CNN decoder to effectively capture the global context and local features simultaneously. Despite their relatively favorable accuracy, these methods are commonly associated with excessively intricate network structures. The excessive calculation of transformer results in the need for additional memory space and computing resources to train the model, which severely hinders the processing of high-resolution remote sensing images.

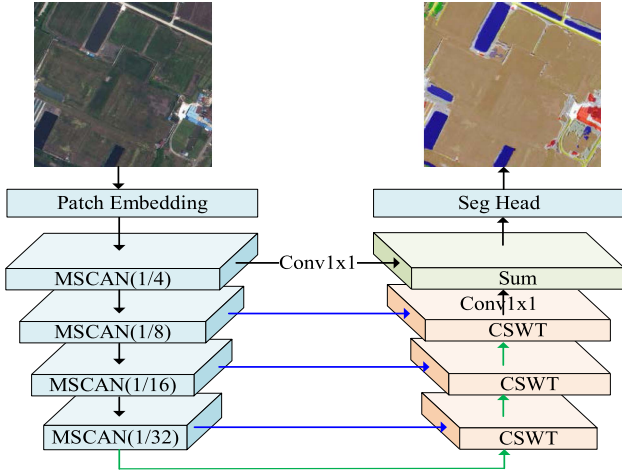


Fig. 2. Pipeline of the proposed MCAT-UNet framework for semantic segmentation of remote sensing image.

Unlike the models proposed in previous works, MCAT-UNet employs a novel multiscale convolutional attention mechanism to evoke spatial attention, which is more effective than standard convolutions and recent self-attention in spatial information encoding. The MSCA sequences are stacked to generate a convolutional encoder for capturing multiscale feature targets more accurately. Additionally, cross-shaped self-attention is embedded in a U-shaped decoder to further capture long-range spatial dependency. Extensive experimental results evaluate the efficiency of our proposed method.

### III. PROPOSED METHOD

In this section, we introduce the MCAT-UNet model in detail. Section III-A describes the overall design of our pipeline. Section III-B describes the principle and network structure of the MSCAN. Section III-C describes the structure of the cross-shaped window transformer decoder. The details are described in the following sections.

#### A. Overall Network Architecture

An overview of our MCAT-UNet framework is presented in Fig. 2. This framework proposes an enhanced UNet by utilizing convolutional attention and cross-shaped window self-attention, which can model long-range spatial dependencies, extract local representations, and capture multiscale targets more accurately.

- 1) The UNet encoder utilizes an MSCAN CNN-based encoder for multilevel feature extraction. For the building block in the encoder, a novel multiscale convolutional attention module is designed to aggregate local information, capture multiscale context, and model relationships between different channels.
- 2) We use CSWT blocks to construct the decoder of U-shaped structure. Three skip connections are built between the encoder and decoder feature maps of identical spatial resolution to preserve the global and local details and facilitate the communication of multiscale features.

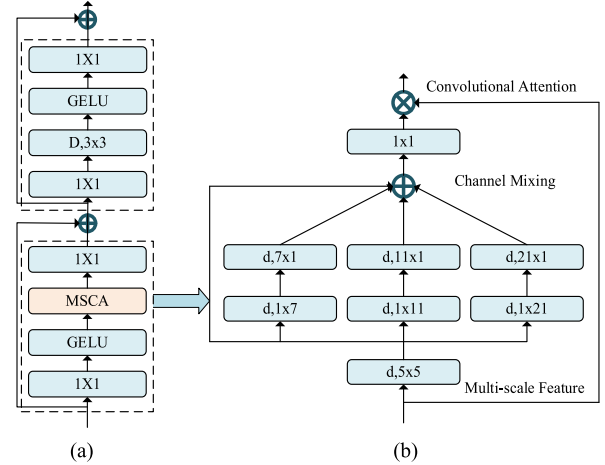


Fig. 3. Illustration of the proposed MSCAN and MSCA. Here,  $d, k_1 \times k_2$  means a depth-wise convolution ( $d$ ) using a kernel size of  $k_1 \times k_2$ . (a) Stage of MACAN. (b) MSCA.

#### B. Multiscale Convolutional Attention Encoders

Recently, CNN-based methods have been compared to transformer-based methods [37], [57], [58]. We apply the CNN-style backbone of MSCAN as the encoder to capture multilevel and multiscale features. As depicted in Fig. 3(a), the encoder adopts a novel multiscale convolutional attention module that contains three parts: a depthwise convolution to aggregate the local context, multibranch depthwise strip convolutions to capture multiscale information, and a  $1 \times 1$  convolution to simulate the relationships between different channels. The output of the  $1 \times 1$  convolution is used for the attention weights and the input of the MSCA is directly reweighted. Mathematically, the MSCA can be formulated as follows:

$$\text{Att} = \text{Conv}_{1 \times 1} \left( \sum_{i=0}^3 \text{Scale}_i (\text{DW} - \text{Conv}(F)) \right) \quad (1)$$

$$\text{Out} = \text{Att} \otimes F \quad (2)$$

where  $F$  is the input feature,  $\text{Att}$  and  $\text{Out}$  are the attention map and output, respectively,  $\otimes$  is the elementwise matrix multiplication operation,  $\text{DW-Conv}$  represents depthwise convolution and  $\text{Scale}_i, i \in \{0, 1, 2, 3\}$ , represents the  $i$ th branch in Fig. 3(b). In each branch, two depthwise strip convolutions are used to approximate standard depthwise convolutions with large kernels. Here, the kernel sizes of the branches are set to 7, 11, and 21.

Stacking a sequence of MSCA yields the proposed convolutional encoder MSCAN; we adopt a hierarchical structure similar to traditional CNNs [50], [51] and recent hierarchical transformer variants [28], [65], which contains four stages with decreasing output spatial resolutions  $H/4 \times W/4$ ,  $H/8 \times W/8$ ,  $H/16 \times W/16$ ,  $H/32 \times W/32$ . Here,  $H$  and  $W$  represent the width and height of the input image, respectively. With decreasing resolution, the number of output channels is  $C$ ,  $2C$ ,  $4C$ , and  $8C$ , where  $C$  is set to 64 in our experiments for fair comparison. The encoder produces multilevel and multiscale features given an input image. These features provide both low-resolution

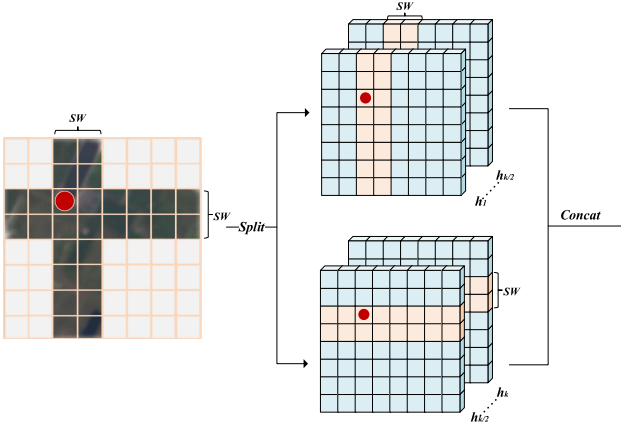


Fig. 4. Illustration of cross-shaped window self-attention.

fine-grained features and high-resolution coarse features that boost the performance of semantic segmentation.

### C. Cross-Shaped Window Transformer-Based Decoder

We adopt a U-shaped decoder to gradually restore the size of the feature map and predict the semantic segmentation result. Skip connections are established between the encoder and decoder feature maps of the same size to preserve the global and local details and facilitate the communication of multiscale features. The cross-shaped self-attention window transformer is used to further capture the long-range spatial dependency.

1) *Cross-Shaped Window Self-Attention*: To capture long-range context, mainstream solutions focus on attaching a single attention block at the end of the network or introducing transformer as the encoder. The former approach cannot capture multiscale global features, whereas the latter significantly increases the computational complexity of the network and causes spatial details to be lost. To address this issue, we present the cross-shaped window self-attention mechanism, which can enlarge the attention area and achieve global self-attention more efficiently, as shown in Fig. 4. First, we split multiple heads ( $\{h_1, \dots, h_k\}$ ) into two groups and perform self-attention simultaneously on the horizontal and vertical stripes. Second, we adjust the stripe to balance the learning capacity and computational complexity. According to the multihead self-attention mechanism, the input feature  $X \in R^{(H \times W) \times C}$  is first linearly projected to  $K$  parallel attention heads, after which each head performs local self-attention within either the horizontal or vertical stripes. For horizontal stripe self-attention,  $X$  is evenly divided into nonoverlapping horizontal stripes  $[X^1, \dots, X^M]$  of equal width  $sw$ , and each of them contains  $sw \times W$  tokens.

Formally, the projection queries, keys and values of the  $k_{th}$  head all have dimensions  $d_k$ ; then, the output of the horizontal self-attention for the  $k^{th}$  head is defined as follows:

$$X = [X^1, X^2, \dots, X^M] \quad (3)$$

$$Y_k^i = \text{Attention} \left( \left[ X^i W_k^Q, X^i W_k^K, X^i W_k^V \right] \right) \quad (4)$$

$$H - \text{Attention}_k(X) = [Y_k^1, Y_k^2, \dots, Y_k^M] \quad (5)$$

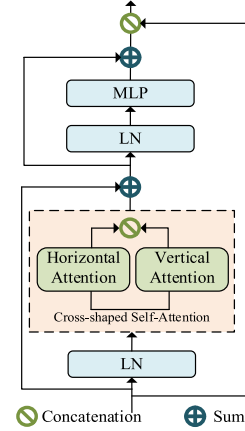


Fig. 5. Illustration of the cross-shaped window transformer block.

where  $X^i \in R^{(sw \times W) \times C}$  and  $M = H/sw$ ,  $i = 1, \dots, M$ .  $W_k^Q \in R^{C \times d_k}$ ,  $W_k^K \in R^{C \times d_k}$ , and  $W_k^V \in R^{C \times d_k}$ , represent the projection matrices of queries, keys and values for the  $k^{th}$  head, respectively, and  $d_k$  is set as  $C/K$ . Similarly, the vertical stripe self-attention can be derived, and its output for the  $k^{th}$  head is denoted as  $V - \text{Attention}_k(X)$ .

The  $K$  heads are equally split into two parallel groups (each has  $K/2$  heads, where  $K$  is often an even value). One group performs horizontal stripe self-attention, and the other group performs vertical stripe self-attention. Finally, the outputs of these two parallel groups are concatenated

$$\text{CSAttention}(X) = \text{Concat}(\text{head}_1, \dots, \text{head}_k) W^O \quad (6)$$

$$\text{head}_k = \begin{cases} H - \text{Attention}_k(X) & k=1, \dots, K/2 \\ V - \text{Attention}_k(X) & k=K/2+1, \dots, K \end{cases} \quad (7)$$

where  $W^O \in R^{C \times C}$  is the projection matrix that projects the self-attention results into the target output dimension. As described above, one key insight in cross-shaped self-attention mechanism design is splitting the multiple heads into different groups and using different self-attention operations. In contrast, existing self-attention mechanisms apply the same self-attention operations across multiple heads. In the experimental section, we illustrate that this design results in better performance.

2) *Cross-Shaped Window Transformer (CSWT) Block*: Cross-shaped self-attention is crucial for semantic segmentation of long-range information and is embedded into the Transformer architecture, as shown in Fig. 5. The Transformer block deploys cross-shaped self-attention, which is achieved by conducting self-attention on horizontal and vertical stripes in parallel to form a cross-shaped window to capture the global context. The stripe width  $sw$  is set to 1, 4, and 4 for the three stages by default. The details of the cross-shaped window transformer can be found in the CSwin transformer [30]. Finally, we employ layer normalization and a common MLP layer to characterize the fine-grained global context. The Transformer block is formally defined as follows:

$$X^l = \text{MLP} \left( \text{LN} \left( \text{CSAttention} \left( \text{LN} \left( X^{l-1} \right) \right) + X^{l-1} \right) \right) \quad (8)$$

where  $X^l$  represents the output of the  $l$ th transformer block.

## IV. EXPERIMENTS

In this section, to evaluate the rationality of our proposed model, we first conduct a series of ablation experiments on the validation sets of LoveDA, Potsdam, and Vaihingen with different settings and frameworks. Next, we compare our methods with other approaches on public benchmarks and show the superiority of our proposed model in terms of computational complexity.

### A. Dataset Description

1) *LoveDA*: The LoveDA<sup>1</sup> dataset contains 5987 fine-resolution optical remote sensing images (GSD 0.3 m) with a size of 1024×1024 pixels and includes seven land cover categories, i.e., building, road, water, barren, forest, agriculture, and background. Specifically, 2522 images were used for training, 1669 images were used for validation, and 1796 images were officially provided for testing. The dataset encompasses two scenes (urban and rural) collected from three cities (Nanjing, Changzhou, and Wuhan) in China. Consequently, the dataset presents a significant research challenge due to the presence of multiscale objects, complex backgrounds, and inconsistent class distributions.

2) *Potsdam*: The Potsdam<sup>2</sup> dataset was collected by aerial cameras with a resolution of 6000×6000 pixels over Potsdam city, and the ground sampling distance was 5 cm. The dataset has 38 samples, with 24 for training and 14 for testing. Each sample contains three images with a true orthophoto (TOP), a digital surface model (DSM), and ground truth. The dataset was manually classified into the six most common land cover categories, and the ground sampling distance between the TOP and the DSM was 5 cm. In this article, we follow the approach used in and use 23 images (excluding image 7\_10 with error annotations) for training and 14 images for testing.

3) *Vaihingen*: The village of Vaihingen<sup>3</sup> comprises many individual buildings and small multistory buildings, and similar to the Potsdam dataset, it has been classified into six common land cover categories. The dataset includes 3-band remote sensing TIFF files (near-infrared, red, green) and a single band DSM, with 33 HRS images of varying sizes. For the experiment, we followed [74] to select the remote sensing images with IDs 2, 4, 6, 8, 10, 12, 14, 16, 20, 22, 24, 27, 29, 31, 33, 35, and 38 for testing, while the remaining 16 images were used for training.

### B. Experimental Details

1) *Training Settings*: We use mmsegmentation [74] codebases and follow the commonly used training settings. All the models in the experiments are implemented with the PyTorch framework on a single NVIDIA GTX 3090 GPU. Unless otherwise stated, all of the networks use the following settings from [28]: we use the AdamW optimizer with an initial learning rate of 6e−5, a weight decay of 0.01, and a total number of iterations

of 100K. The parameters of all the models are initialized with officially provided pretraining weights. The learning rate warmups with 1500 iterations at the beginning and decays with a linear decay strategy. We use the default augmentation settings for mmsegmentation, which includes random horizontal flipping, random rescaling (ratio range [0.5, 2.0]) and random photometric distortion. All the models are trained with an input size of 512×512. In regard to testing, we report both the single-scale test results and the multiscale test results ([0.5, 0.75, 1.0, 1.25, 1.5, 1.75] × of that in training). The soft cross-entropy loss function is used.

2) *Evaluation Metric*: We use the floating-point operations per second (FLOPs), the number of parameters (Param.) and frames per second (FPS) to evaluate the computational cost of the model. To evaluate the performance, we adopt three common metrics, namely, the overall accuracy (OA), mean intersection over union (mIoU), and F1-score (mF1), which are chosen as evaluation indices.

### C. Ablation Study

1) *Influence of Difference Encoders*: We first explored the performance of several backbone networks based on CNN and transformer combined with our proposed CSWT decoder on the Potsdam dataset. The CNN-based backbone networks include ResNet [53], PoolFormer [56], ConvNeXt [57], VAN [58], and SegNeXt [37]; the transformer-based backbone networks include SegFormer [45], Swin [28], Beit [67], Twins [68], and CSWin [30]. Most of the models are recently proposed methods designed to extract features. Limited by our GPU memory, and to ensure a fair comparison, we set the input image size to 512×512 and the batch size to 4. As shown in Table I, ConvNeXt, the VAN, and MACAN achieved mIoU results of 74.57%, 75.12%, and 75.48%, respectively, on Potsdam, demonstrating that a well-designed convolution model can compete favorably with state-of-the-art hierarchical vision Transformers with relatively fewer parameters and FLOPs. Even a more classic and simple structure, such as ResNet18, achieves a relatively high score of 72.11%, indicating that the cross-shaped window transformer can capture long-range dependency information more efficiently. MSCAN achieves the best mIoU of 75.48%, mF1 of 85.02%, and OA of 83.40% on Potsdam, which indicate that embedding multiscale convolution attention and cross-window transformers in UNet can significantly improve segmentation performance. PoolFormer achieves a lower performance (73.65%) on the mIoU, as it only uses simple space pooling operations for basic token mixing. Considering the model complexity and hardware resources, we apply MACAN as the backbone in our subsequent experiments.

2) *Each Component of MCAT-UNet*: To better evaluate the performance of each critical design in the proposed MCAT-UNet on the Potsdam dataset. We conduct a series of ablation experiments under a completely fair setting in which we use the same architecture and hyperparameter for the following experiments. For fair comparison, we set the input image size to 512×512 and the batch size to 8. The results are illustrated in Table II.

<sup>1</sup>[Online]. Available: <https://github.com/Junjue-Wang/LoveDA>

<sup>2</sup>[Online]. Available: <https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx>

<sup>3</sup>[Online]. Available: <https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx>

TABLE I  
COMPARISON WITH STATE-OF-THE-ART ENCODERS ON THE LOVEDA AND POTSDAM DATASETS

Model	Backbone	Param.	FLOPs	FPS	Potsdam		
					mIoU(%)	mF1(%)	OA(%)
Transformer-Based	MiT-B1 [45]	22.9	15.6	52.5	74.07	83.75	82.11
	Swin-Tiny [28]	47.6	35.7	39.3	74.65	84.26	82.69
	Beit [67]	138.2	187.4	28.1	74.86	84.33	82.67
	Twins-S [68]	32.5	21.8	32.5	75.13	84.65	83.17
	CSWin-Tiny [30]	30.7	29.0	20.8	74.93	84.44	82.87
CNN-Based	ResNet50 [51]	170	92.5	42.5	73.76	83.48	81.83
	ResNet18 [51]	20.1	15.3	81.1	72.11	82.29	80.48
	PoolFormer-S24 [56]	30.7	22.9	48.6	73.65	83.45	81.70
	ConvNeXt-Tiny [57]	47.8	33.4	63.1	74.57	84.25	82.87
	VAN-Small [58]	23.1	18.2	50.1	75.12	84.63	83.11
	MACAN-Small [37]	23.2	18.5	51.7	<b>75.48</b>	<b>85.02</b>	<b>83.40</b>

TABLE II  
ABLATION STUDY OF EACH COMPONENT OF THE MCAT-UNET ON THE POTSDAM DATASET

Dataset	Method.	mIoU (%)	mF1 (%)	OA(%)
Potsdam	Baseline	74.88	84.50	82.81
	Baseline+CSWT	75.50	84.92	83.27

TABLE III  
DIFFERENT BATCH SIZE ON SINGLE GPU FOR TRAINING MODEL ON THE LOVEDA DATASET

Batch size	mIoU(%)	mF1(%)	OA(%)	Train-time
2	71.26	80.38	80.38	4:20'
4	73.88	82.09	82.09	4:30'
8	73.89	82.11	82.11	4:35'
16	74.06	82.24	82.24	5:30'
32	73.47	81.87	81.87	9:00'

a) *Baseline*: The baseline was constructed by the MSCAN backbone with a U-shaped decoder, which can achieve multi-scale context from local to global, obtain adaptability in spatial and channel dimensions, and aggregate information from low to high levels.

b) *The cross-shaped window transformer (CSWT) block*: The CSWT block is incorporated into the baseline to construct the baseline+CSWT. As shown in Table II, the deployment of the CSWT significantly increases the mIoU, mF1, and OA by 0.62%, 0.42%, and 0.46%, respectively, on the Potsdam dataset. This is because the cross-shaped window transformer deeply models the long-distance dependencies and further extracts global contextual information.

3) *Comparison of Different Batch Sizes*: Batch size is important because it affects both the training time and the generalization of the model. To our knowledge, no one has run mmsegmentation [74] on a single GPU to train their model. In this experiment, we investigate the effect of batch size on training dynamics to determine the appropriate batch size for a single GPU for the following experiments. Limited by our GPU memory, the input image size is  $256 \times 256$ . We can only perform our implementation on a single NVIDIA GTX 3090 GPU by using mmsegmentation codebases. Table III shows the influence

of different batch sizes  $\{2, 4, 8, 16, 32\}$ . When other settings are kept constant, batch sizes of 4, 8, and 16 result in comparable top performances. Furthermore, further increases in the batch size (32) lead to a decrease in performance but also a significant increase in training time. This finding illustrates that increasing the batch size merely to train MCAT-UNet does not necessarily lead to performance improvements. For fair comparison, we set the batch size equal to 8 by default to balance the performance and calculation amount in all the following experiments unless otherwise specified.

4) *Influence of Different Input Sizes*: We conduct experiments on the Potsdam dataset to analyze the influence of different input sizes during training, including square inputs of  $256 \times 256$ ,  $512 \times 512$ , and  $1024 \times 1024$ , as well as rectangular inputs of  $256 \times 512$  and  $512 \times 1024$ . Due to the memory limitations of our GPU, the batch size is 4, and the maximum input size that can be loaded is  $1024 \times 1024$ . As shown in Table IV, square inputs yield relatively higher scores than rectangular inputs, and the maximum deviation in the mIoU between  $256 \times 256$  and  $1024 \times 1024$  is 2.2%. As the input size increases, the mIoU deviation decreases gradually, and the training time also increases. With an input size of  $1024 \times 1024$ , the best mIoU is 76.08%, the mF1 is 85.33%, and the OA is 83.81% on the Potsdam dataset. Moreover, in the segmentation of small objects, such as the ‘‘car’’ class, the segmentation accuracy is significantly improved, indicating its effectiveness in segmenting small objects in HRSIs. A smaller input image produces a coarser predictive representation, and a smaller sequence is produced so that transformer can process it faster. Large-scale images can help improve the segmentation performance of the model, but it is necessary to pay attention to calculations in a longer sequence, which increases the training time and calculation cost. For fair comparison, in terms of experimental efficiency, an input size of  $512 \times 512$  is selected as the default to balance the experimental efficiency and performance in all the following experiments.

#### D. Experimental Results

In this section, we compare the proposed method with several state-of-the-art semantic segmentation methods on three different remote sensing datasets from LoveDA, Potsdam, and Vaihingen, including natural image segmentation methods, such as

TABLE IV  
ABLATION STUDIES OF DIFFERENT INPUT SIZES ON THE POTSMAN DATASET

Input size	Imp.surf.	Building	Lowveg.	Tree	Car	Clutter	mIoU(%)	mF1(%)	OA(%)
256×256	83.27	90.46	73.35	75.16	81.18	39.87	73.88	83.82	82.09
256×512	80.64	89.03	70.70	73.20	81.40	30.03	70.83	81.13	79.45
512×512	84.69	92.05	73.88	75.90	82.79	43.57	75.48	85.02	83.40
512×1024	83.07	91.49	72.47	75.34	83.17	35.25	73.47	83.20	81.6
1024×1024	85.35	92.72	75.01	76.65	84.06	42.69	76.08	85.33	83.81

TABLE V  
ABLATION STUDIES OF DIFFERENT DECODER WITH THE MSCAN ON THE LOVEDA AND POTSDAM DATASET

Model	Params.(Mb)	FLOPs(Gbps)	FPS	LoveDA			Potsdam		
				mIoU(%)	mF1(%)	OA(%)	mIoU(%)	mF1(%)	OA(%)
LightHamHead [37]	<b>13.9</b>	<b>15.6</b>	49.0	51.84	67.41	64.92	74.53	84.27	82.81
UPerNet [38]	43.1	228.4	34.5	51.46	66.87	64.90	75.18	84.67	83.07
DANet [23]	27.9	423.9	20.9	50.60	66.13	65.04	<u>75.43</u>	<b>85.02</b>	<b>83.64</b>
CCNet [26]	27.9	267.3	26.2	52.71	68.15	65.63	75.06	84.61	83.18
OCRNet [39]	19.9	116.7	41.1	<u>52.92</u>	<u>68.27</u>	65.37	75.26	84.72	82.11
PSPNet [22]	29.7	251.9	31.0	<u>52.22</u>	67.38	<u>65.70</u>	74.85	84.42	82.80
ALL-MLP [45]	<u>15.0</u>	31.9	<b>54.7</b>	51.79	67.14	64.13	75.17	<u>84.70</u>	83.12
K-Net [27]	56.3	241.7	28.2	51.91	67.25	65.25	74.81	84.31	82.63
<b>MCAT-UNet(Ours)</b>	23.2	<u>18.5</u>	<u>51.7</u>	<b>53.48</b>	<b>68.83</b>	<b>66.28</b>	<b>75.48</b>	<b>85.02</b>	<b>83.40</b>

UPerNet [38], DANet [23], CCNet [26], OCRNet [39], PSPNet [22], Knet [27], DeeplabV3+ [44], and ALL-MLP [45]; remote sensing image segmentation methods, such as BANet [69], ABCNet [70], MANet [71], MAREs-UNet [17], SRANet [76], DCSwin [32], ST-UNet [34], UNetFormer [41], and EMRT [40]. Among these methods, EMRT, UNetFormer, DCSwin, SRANet, and ST-UNet have recently been proposed. All the models are trained with an input size of  $512 \times 512$ , and the batch size is set to 8. We analyze the segmentation performance on three datasets and benchmark our MCAT-UNet using various metrics, namely, the mIoU, mF1, OA, and IoU per category. The quantitative performance comparisons between our MCAT-UNet and previous state-of-the-art models are presented in Tables V–VIII. The bold and underlined values in each column represent the best and second-best performances, respectively. The comparison details as follows.

1) *Comparison of Difference Decoders*: Many semantic segmentation methods for remote sensing images (e.g., [17], [32], [34], [41], [69], [70], [71], [72], [73], [77], etc.) utilize the backbone of computer vision, but the validity of these methods have not been fully proven. We compare our MCAT-UNet with the state-of-the-art decoders initially designed for natural images on the LoveDA and Potsdam datasets using the same backbone, e.g., MSCAN. Limited by our GPU memory, and to ensure a fair comparison, we set the input image size to  $512 \times 512$  and the batch size to 4. As shown in Table V, MCAT-UNet has the lowest computational complexity (e.g., FLOPs only 15.9% of OCRNet, at least 8.1% of UPerNet, CCNet, PSPNet, and K-Net; and 4.4% of DANet) and achieves the most competitive performance; the mIoU is 53.48% and 75.48%, the mF1 is 68.83% and 85.02%, and the OA is 66.28% and 83.40%, by integrating global and local contextual information respectively. MCAT-UNet is superior to all the previous decoders, which validates the effectiveness of convolutional attention and

cross-shaped self-attention in obtaining context from both local and global perspectives. Additionally, we found that many methods proposed for natural images are equally effective when transferred to the task of remote sensing image segmentation. By comparing Tables V–VII, we observed that many methods in the field of remote sensing do not perform as well as the methods in natural images, but our method outperforms them.

2) *Results on the LoveDA Dataset*: We compare the performances of MCAT-UNet and several state-of-the-art methods on the LoveDA dataset. Different from typical datasets, the LoveDA dataset contains real urban and rural remote sensing images and is recognized as a challenging HSRI dataset for land cover domain adaptive semantic segmentation. This dataset presents three challenges in large-scale remote sensing mapping, namely, multiscale objects, complex background samples, and inconsistent class distributions. Thus, it is difficult to obtain high scores on this dataset. As shown in Table VI, all the models are initialized with officially provided pretraining weights. Compared with other methods, our method demonstrates superior performance, with an mIoU of 53.58%, an mF1 of 68.88%, and an OA of 66.32%, especially for categories with large intraclass variations, such as roads and water. Moreover, the proposed method obtains the highest IoU values for buildings, roads, water, and agriculture, where the road category is typically characterized by narrow and elongated features. Whether in rural or urban scenarios or dense or sparsely distributed, our method can accurately segment objects in complex environments with high confidence. This finding illustrates that the effectiveness of MCAT-UNet is sufficient for learning long-range spatial dependencies and capturing multiscale targets more accurately to enhance the differences among objects.

Furthermore, Fig. 6 provides some examples of visual comparisons, which also demonstrate the effectiveness of our method. As we can see, the visualization results of the proposed



TABLE VI  
QUANTITATIVE COMPARISON RESULTS BETWEEN OUR METHOD AND OTHER PUBLISHED METHODS ON THE LOVEDA DATASET

Model	Backbone	Background	Building	Road	Water	Barren	Forest	Agriculture	mIoU(%)	mF1(%)	OA(%)
DeepLapV3+ [75]	ResNet50	51.73	62.34	55.05	64.99	23.09	40.06	47.88	49.31	64.90	62.22
BANet [69]	ResT	53.55	60.91	53.57	63.74	30.41	40.89	<u>54.09</u>	51.02	66.85	64.08
ABCNet [70]	ResNet18	52.02	60.41	50.71	61.08	<u>32.84</u>	39.02	50.51	49.51	65.66	62.85
MANet [71]	ResNet50	52.96	62.09	54.25	65.12	25.21	38.56	47.15	49.33	65.01	61.17
MAResU-Net [17]	ResNet18	52.85	60.29	52.87	68.58	30.80	40.85	50.47	50.96	66.73	64.34
DCSwin [32]	SwinTiny	<u>53.92</u>	<u>62.89</u>	<u>56.28</u>	66.74	<b>34.24</b>	<u>42.30</u>	47.89	<u>52.04</u>	<u>67.80</u>	<u>65.48</u>
ST-UNet [34]	SwinTiny	50.61	54.59	46.87	59.71	21.13	34.82	44.61	44.62	60.67	64.74
UNetFormer [41]	ResNet18	51.64	59.1	54.99	65.16	30.59	40.07	49.83	50.20	66.13	64.54
EMRT [40]	ResNet50	<b>53.98</b>	61.39	55.37	66.12	29.25	39.52	49.56	50.74	66.46	63.48
<b>MCAT-UNet(Ours)</b>	MSCAN-S	53.77	<b>63.25</b>	<b>56.7</b>	<b>71.26</b>	31.48	<u>41.32</u>	<b>57.26</b>	<b>53.58</b>	<b>68.88</b>	<b>66.32</b>

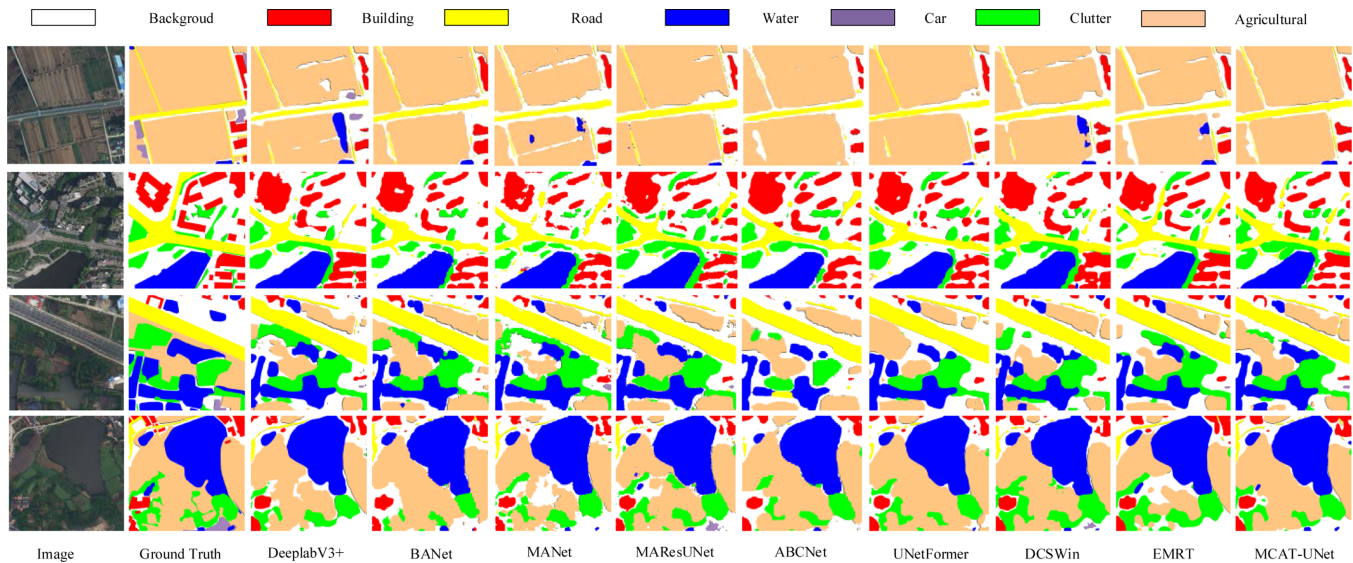


Fig. 6. Qualitative comparison with different methods on the LoveDA dataset.

TABLE VII  
QUANTITATIVE COMPARISON RESULTS BETWEEN OUR METHOD AND OTHER PUBLISHED METHODS ON THE POTSDAM DATASET

Model	Backbone	Imp.surf.	Building	Lowveg	Tree	Car	Clutter	mIoU(%)	mF1(%)	OA(%)
DeepLapV3+ [75]	ResNet50	83.73	91.78	73.02	75.14	83.49	38.27	74.24	83.91	82.39
BANet [69]	ResT	82.66	90.37	72.43	73.88	81.27	34.33	72.49	82.54	80.95
ABCNet [70]	ResNet18	82.31	90.15	71.17	73.92	80.40	34.35	72.05	82.26	80.55
MANet [71]	ResNet50	83.54	91.68	72.47	74.75	82.78	36.13	73.56	83.32	81.73
MAResU-Net [17]	ResNet18	83.10	90.98	71.65	73.81	81.81	34.25	72.60	82.58	80.87
DCSwin [32]	SwinTiny	84.17	91.80	73.72	75.40	81.93	39.75	74.46	84.16	82.69
ST-UNet [34]	/	79.68	86.37	70.08	70.55	77.44	32.50	69.43	85.77	80.48
UNetFormer [41]	ResNet18	82.84	90.29	71.45	74.32	82.51	34.67	72.68	82.67	81.04
SRANet [76]	ResNet50	81.63	89.11	71.69	73.12	81.91	35.62	72.18	82.45	-
EMRT [40]	ResNet50	83.91	91.63	72.85	74.89	83.32	36.19	73.80	83.48	81.83
<b>MCAT-UNet(Ours)</b>	MSCAN-S	<b>84.63</b>	<b>92.46</b>	<b>74.3</b>	<b>76.33</b>	<b>83.75</b>	<b>41.15</b>	<b>75.44</b>	<b>84.84</b>	<b>83.31</b>

method are more complete predictions for large objects (e.g., agriculture, water, and forest) and small multiscale objects (e.g., buildings and roads), where the boundaries remain accurate and smooth without any additional processing. Compared with other methods, MCAT-UNet has higher edge segmentation accuracy for adjacent objects while preserving better spatial details. The results show that MCAT-UNet has higher edge segmentation accuracy for adjacent objects while preserving better spatial

details. The results show that MCAT-UNet can effectively integrate global–local context information and achieve better performance by capturing full image dependencies; this approach is an important design for efficient segmentation.

3) *Results on the Potsdam Dataset*: To further verify the generality of MCAT-UNet, we perform experiments on the Potsdam dataset. As shown in Table VII, compared with other remote sensing image semantic segmentation methods, MCAT-UNet

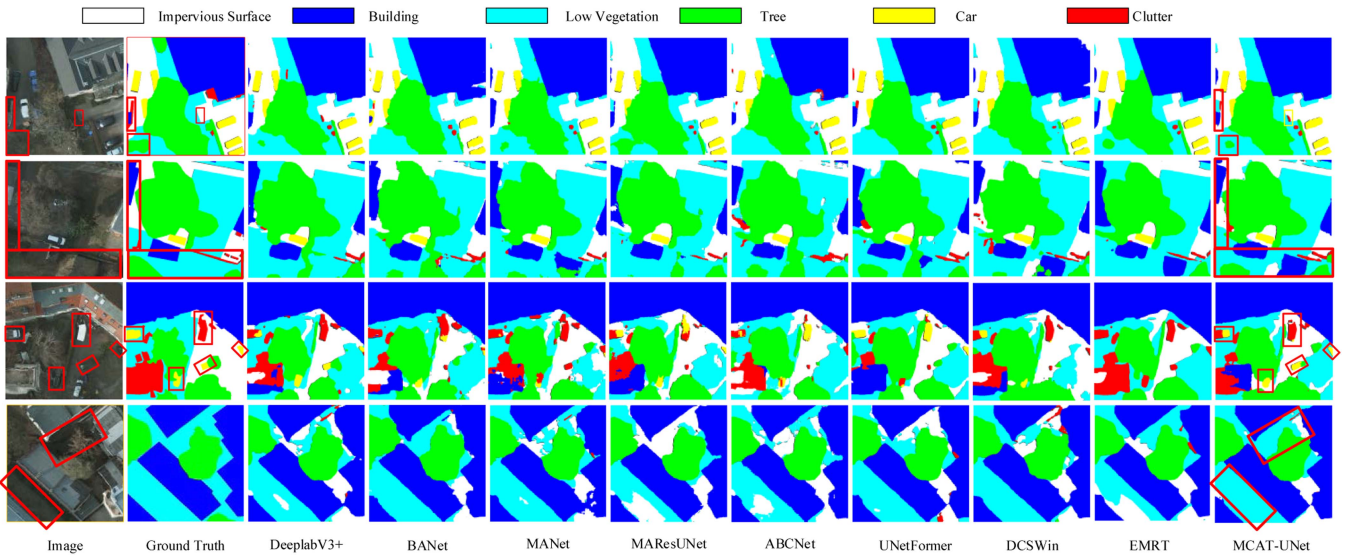


Fig. 7. Qualitative comparison results of different methods on the potsdam dataset.

achieves state-of-the-art performance with an mIoU of 75.44%, an mF1 of 84.84%, and an OA of 83.31% on the Potsdam dataset and achieves the highest accuracy in all six categories. These results indicate the effectiveness of the proposed strategy. Moreover, on the small class (e.g., car), our method achieves an IoU of 83.75%, which is much better than that of the previous state-of-the-art method. In particular, the car category is typically characterized by small features, which require global and local information for segmentation. Extracting contextual information is an effective method for improving model recognition. This result fully indicates that MCAT-UNet has better performance for small target segmentation in HRSIs.

We also provide four representative visualization results to show the preferential performance of our network. As shown in Fig. 7, the small yellow box in the first row is unmarked in the ground truth, and the target object is very small; thus, only our model can make an accurate prediction. The objects in the large red boxes in the second row were marked incorrectly, but our method can correctly identify them. The results of the second row demonstrate that our methods can more precisely identify challenging samples within buildings, and the boundaries of the buildings are smoother and more accurate. The results in the third row show that our methods can identify small samples of cars more accurately. In the first and fourth rows of the results, there are three regions of low vegetation that exhibit visual properties similar to those of trees and impervious surfaces, respectively, with high interclass homogeneity. Most of the methods seriously misidentify low vegetation in this area, but our method is still able to discriminate them accurately. In general, the proposed MCAT-UNet can obtain more accurate segmentation maps, especially for complex irregular targets, and yield finer boundary details.

4) *Results on the Vaihingen Dataset:* We also compare the proposed methods on the ISPRS Vaihingen dataset. This dataset includes a large number of houses obscured by tree branches and multistory small villages, so the networks need to more

accurately identify and segment small targets. As shown in Table VIII, the proposed method achieves the best mIoU of 74.52%, mF1 of 84.01%, and OA of 81.41%. These methods are much better than previous methods. Further analysis of the IoU score of each category showed that the proposed MCAT-UNet achieves the best performance in the four categories of impervious surfaces, building, low vegetation, and car, and is second best in the tree class, following DCSwin. Notably, our method achieves an IoU of 78.81% on the car class, which is more than 2.08% higher than that of other networks. This demonstrates that our method is more capable of modeling complex irregular targets and further reveals the importance of global long-range interaction representations for semantic segmentation of remote sensing images.

A visualization of the results is shown in Fig. 8, which also demonstrates that our proposed models achieve better performance. In regions affected by building shadows, such as the red box marked in the middle of the third row, which represents a tree, and the red box marked in the first and second rows, which represents a car, most of the methods lose spatial information, resulting in incorrect identification of low vegetation or severe missing contours around cars; moreover, the results of our method more closely match the ground truths. Due to the high similarity of trees and low vegetation appearances and because they always appear in adjacent locations, most methods tend to incorrectly predict low vegetations as trees or vice versa; however, our method is still able to accurately distinguish them, such as the red marked boxes in the second and fourth rows. We select images with intraclass variation in the second row, marked with a yellow box, for comparison. Unlike regular buildings, the marked building appears red in the image and has a very confusing appearance that is similar to that of low vegetation. Many methods completely ignore the building and even incorrectly predict areas with low vegetation coverage. Despite the improvements in the Transformer-based methods, the shape of the object is still incomplete. Our method

TABLE VIII  
QUANTITATIVE COMPARISON RESULTS BETWEEN OUR METHOD AND OTHER PUBLISHED METHODS ON THE VAIHINGEN

Model	Backbone	Imp.surf.	Building	Lowveg	Tree	Car	Clutter	mIoU(%)	mF1(%)	OA(%)
DeepLapV3+ [75]	ResNet50	83.73	91.78	73.02	75.14	83.49	38.27	74.24	83.91	82.39
BANet [69]	ResT	82.66	90.37	72.43	73.88	81.27	34.33	72.49	82.54	80.95
ABCNet [70]	ResNet18	82.31	90.15	71.17	73.92	80.40	34.35	72.05	82.26	80.55
MANet [71]	ResNet50	83.54	91.68	72.47	74.75	82.78	36.13	73.56	83.32	81.73
MAResU-Net [17]	ResNet18	83.10	90.98	71.65	73.81	81.81	34.25	72.60	82.58	80.87
DCSwin [32]	SwinTiny	84.17	91.80	73.72	75.40	81.93	39.75	74.46	84.16	82.69
ST-UNet [34]	/	79.68	86.37	70.08	70.55	77.44	32.50	69.43	85.77	80.48
UNetFormer [41]	ResNet18	82.84	90.29	71.45	74.32	82.51	34.67	72.68	82.67	81.04
SRANet [76]	ResNet50	81.63	89.11	71.69	73.12	81.91	35.62	72.18	82.45	-
EMRT [40]	ResNet50	83.91	91.63	72.85	74.89	83.32	36.19	73.80	83.48	81.83
<b>MCAT-UNet(Ours)</b>	<b>MSCAN-S</b>	<b>84.63</b>	<b>92.46</b>	<b>74.3</b>	<b>76.33</b>	<b>83.75</b>	<b>41.15</b>	<b>75.44</b>	<b>84.84</b>	<b>83.31</b>

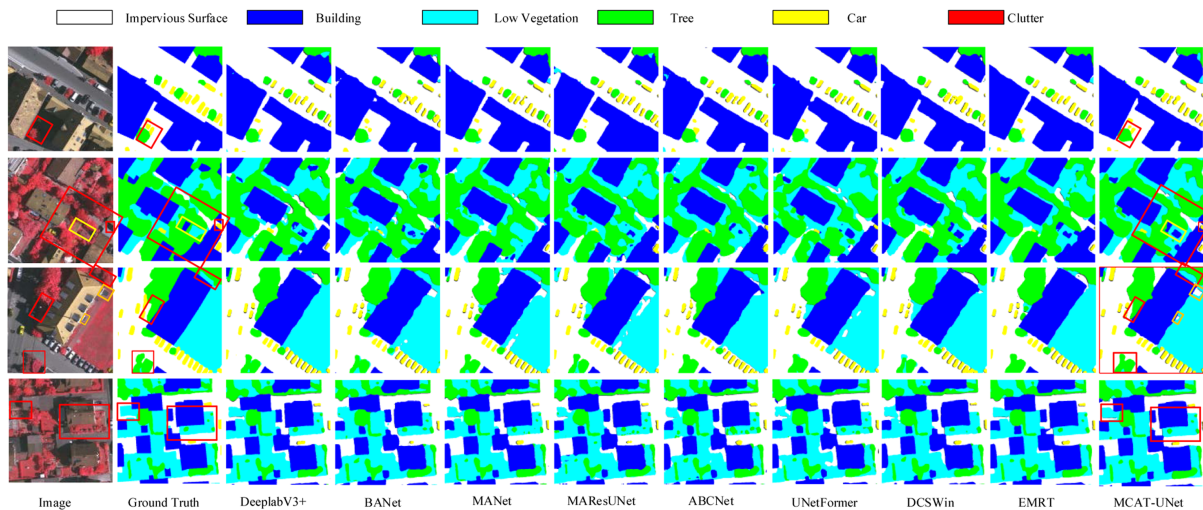


Fig. 8. Qualitative comparison results of different methods on the Vaihingen dataset.

achieves more accurate and consistent segmentation results for buildings with large variations in appearance, illustrating its feature extraction capability.

### E. Feature Visualization

We apply class activation maps to visualize the predicted class on the given image, highlighting the discriminative object parts detected by the MCAT-UNET and its baseline at four stages, to confirm the effectiveness of the model. As shown in Fig. 9, it can be clearly observed that the discriminative regions of the “car” class in the images are highlighted, and as the layers deepen, the class becomes more prominent. Furthermore, compared to the baseline, the MCAT-UNET exhibits stronger attention to the “car” class at each stage. The yellow boxes in the second and third rows indicate the “clutter” class, where we see that MMAT-UNET shows stronger attention. The red boxes in the second and third rows represent the “car” class, and we observe that the MCAT-UNET exhibits stronger attention on the “car” class. Even in complex background environments, MCAT-UNET can effectively identify small targets with similar appearances belonging to different categories, which are difficult to distinguish. This indicates that MCAT-UNET generates local attention similar to convolutions and nonlocal attention similar

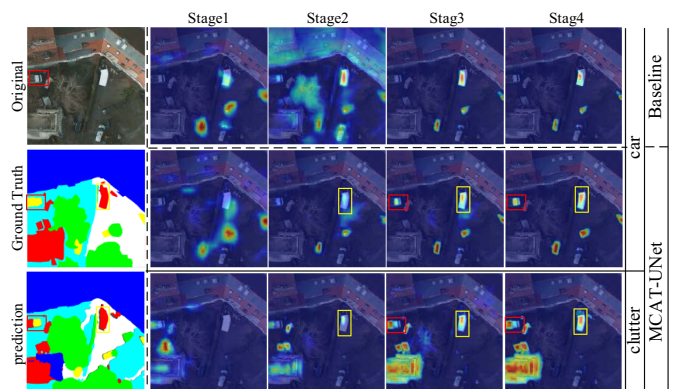


Fig. 9. Visualizations of the output features of the “car” and “clutter” classes by the MCAT-UNET and its baseline at four stages.

to transformers at each stage, which is beneficial for effectively capturing both local and contextual information.

### F. Complexity Analysis

Table IX shows the efficiency comparisons of different models on LoveDA, Potsdam, and Vaihingen. We calculate the overall complexity of the encoder and decoder. The complexity is

TABLE IX  
EFFICIENCY COMPARISON OF DIFFERENT METHODS IN TERMS OF PARAMETERS AND FLOPS

Model	Params.(Mb)	FLOPs(Gbps)	LoveDA		Potsdam		Vaihingen	
			mIoU(%)	mF1(%)	mIoU(%)	mF1(%)	mIoU(%)	mF1(%)
DeepLapV3+ [75]	41.2	181.2	49.31	64.90	74.24	83.91	73.78	83.61
BANet [69]	12.7	15.2	51.02	66.85	72.49	82.54	70.93	81.01
ABCNet [70]	14.0	17.0	49.51	65.66	72.05	82.26	72.37	82.79
MANet [71]	35.9	55.5	49.33	65.01	73.56	83.32	73.44	83.36
MAResU-Net [17]	16.2	20.4	50.96	66.73	72.60	82.58	71.66	81.93
DCSwin [32]	45.6	48.4	52.04	67.80	74.46	84.16	72.54	82.26
UNetFormer [41]	11.7	13.0	50.20	66.13	72.68	82.67	71.80	82.12
EMRT [40]	51.7	158.7	50.74	66.46	73.80	83.48	74.03	83.86
<b>MCAT-UNet(Ours)</b>	23.2	18.5	<b>53.58</b>	<b>68.88</b>	<b>75.44</b>	<b>84.84</b>	<b>74.52</b>	<b>84.01</b>

measured by a  $512 \times 512$  input. We evaluate the computational efficiency against the number of parameters (Param.) measured in million (Mb) and FLOPs. Although the computational complexity of MCAT-UNet is not the lowest, it performs best on the F1, OA, and IoU metrics across the three datasets, with at least increases of 1.54% (mIoU) and 1.08% (mF1) on LoveDA; 0.98% (mIoU) and 0.68% (mF1) on Potsdam; and 0.49% (mIoU) and 0.15% (mF1) on Vaihingen. It is obvious that the performance of the proposed method is significantly improved with relatively few computing resources.

## V. CONCLUSION

In this article, we propose an efficient U-shaped encoder architecture that applies multiscale convolutional attention and a cross-shaped window transformer to reconstruct UNet for efficient HRSI segmentation. The proposed MCAT-UNet can efficiently model long-range spatial dependency with low computational complexity, extract local representations, and enhance hierarchical multiscale targets more accurately. In particular, MCAT-UNet achieves more complete predictions for large-scale varied objects and small discrete multiscale objects, where the boundaries remain accurate and smooth. A comprehensive set of experiments and ablation studies on the LoveDA, ISPRS Vaihingen and Potsdam datasets demonstrate the superiority of the proposed approach compared with other related methods. We expect this enhanced UNet design approach becoming an interesting direction for future research in remote sensing semantic segmentation.

## REFERENCES

- [1] M. Weiss, F. Jacob, and G. Duveiller, "Remote sensing for agricultural applications: A meta-review," *Remote Sens. Environ.*, vol. 236, 2020, Art. no. 111402.
- [2] J. Wang et al., "LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation," 2021, *arXiv:2110.08733*.
- [3] A. Samie et al., "Examining the impacts of future land use/land cover changes on climate in Punjab province, Pakistan: Implications for environmental sustainability and economic growth," *Environ. Sci. Pollut. Res.*, vol. 27, pp. 25415–25433, 2020.
- [4] S. J. O'neill et al., "On the use of imagery for climate change engagement," *Glob. Environ. Change*, vol. 23, no. 2, pp. 413–421, 2013.
- [5] G. J. Schumann et al., "Assisting flood disaster response with earth observation data and products: A critical assessment," *Remote Sens.*, vol. 10, no. 8, 2018, Art. no. 1230.
- [6] R. Andrade et al., "Evaluation of semantic segmentation methods for deforestation detection in the amazon," *Int. Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 43, no. B3, pp. 1497–1505, 2020.
- [7] A. Boguszewski, D. Batorski, N. Ziemia-Jankowska, T. Dziedzic, and A. Zambrzycka, "LandCover. Ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1102–1110.
- [8] Z. Zheng, Y. Zhong, J. Wang, A. Ma, and L. Zhang, "Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4096–4105.
- [9] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2016, pp. 680–688.
- [10] J. Feng et al., "Class-aligned and class-balancing generative domain adaptation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–17, Feb. 2024, Art. no. 5509617.
- [11] K. Xu, H. Huang, P. Deng, and Y. Li, "Deep feature aggregation framework driven by graph convolutional network for scene classification in remote sensing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 10, pp. 5751–5765, Oct. 2022.
- [12] X. Wu et al., "Multi-task multi-objective evolutionary network for hyperspectral image classification and pansharpening," *Inf. Fusion*, vol. 108, 2024, Art. no. 102383.
- [13] J. Feng et al., "MR-selection: A meta-reinforcement learning approach for zero-shot hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–20, Dec. 2023, Art. no. 5500320.
- [14] 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.
- [15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervention-MICCAI*, 2015, vol. 9351, pp. 234–241.
- [16] F. I. Diakogiannis et al., "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 162, pp. 94–114, 2020.
- [17] R. Li, S. Zheng, C. Duan, J. Su, and C. Zhang, "Multistage attention ResUnet for semantic segmentation of fine-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8009205.
- [18] J. Chen, J. Zhu, G. Sun, J. Li, and M. Deng, "SMAF-net: Sharing multiscale adversarial feature for high-resolution remote sensing imagery semantic segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 11, pp. 1921–1925, Nov. 2021.
- [19] L.-C. Chen et al., "Semantic image segmentation with deep convolutional nets and fully connected crfs," 2014, *arXiv:1412.7062*.
- [20] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.
- [21] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
- [22] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.

- [23] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3141–3149.
- [24] H. Zhang, H. Zhang, C. Wang, and J. Xie, "Co-occurrent features in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 548–557.
- [25] Y. Yuan et al., "Ocnnet: Object context network for scene parsing," 2018, *arXiv:1809.00916*.
- [26] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnnet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 603–612.
- [27] W. Zhang et al., "K-net: Towards unified image segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 10326–10338.
- [28] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.
- [29] A. Dosovitskiy et al., "ViT: An image is worth 16x16 words: Transformers for image recognition et al.," 2020, *arXiv:2010.11929*.
- [30] 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. 12114–12124.
- [31] N. Carion et al., "End-to-end object detection with transformers," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.
- [32] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng, and S. Fang, "A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, Jan. 2022, Art. no. 6506105.
- [33] D. Wang et al., "Advancing plain vision transformer toward remote sensing foundation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–20, Nov. 2023, Art. no. 5607315.
- [34] 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, pp. 1–15, Jan. 2022, Art. no. 4408715.
- [35] X. Meng, Y. Yang, L. Wang, T. Wang, R. Li, and C. Zhang, "Class-guided swin transformer for semantic segmentation of remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, Oct. 2022, Art. no. 6517505.
- [36] 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, pp. 1–20, Jan. 2022, Art. no. 4408820.
- [37] M.-H. Guo et al., "Segnext: Rethinking convolutional attention design for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 1140–1156.
- [38] T. Xiao et al., "Unified perceptual parsing for scene understanding," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 418–434.
- [39] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 173–190.
- [40] T. Xiao, Y. Liu, Y. Huang, M. Li, and G. Yang, "Enhancing multiscale representations with transformer for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, Mar. 2023, Art. no. 5605116.
- [41] L. Wang et al., "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 190, pp. 196–214, 2022.
- [42] X. Qin et al., "U2-Net: Going deeper with nested U-structure for salient object detection," *Pattern Recognit.*, vol. 106, 2020, Art. no. 107404.
- [43] L.-C. Chen et al., "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
- [44] H. Zhang et al., "Resnest: Split-attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Recognit. Workshops*, 2022, pp. 2735–2745.
- [45] E. Xie et al., "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 12077–12090.
- [46] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 7242–7252.
- [47] J. Chen et al., "Transunet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.
- [48] H. Cao et al., "Swin-unet: Unet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2022, pp. 205–218.
- [49] B. Zhang et al., "SegViTv2: Exploring efficient and continual semantic segmentation with plain vision transformers," 2023, *arXiv:2306.06289*.
- [50] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.
- [51] 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.
- [52] C. Szegedy et al., "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.
- [53] 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. 2261–2269.
- [54] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
- [55] A. Howard et al., "Searching for mobilenetv3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1314–1324.
- [56] W. Yu et al., "Metaformer is actually what you need for vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10809–10819.
- [57] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11966–11976.
- [58] M.-H. Guo et al., "Visual attention network," *Comput. Vis. Media*, vol. 9, no. 4, pp. 1–20, 2023.
- [59] H. Touvron et al., "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.
- [60] X. Zhu et al., "Deformable detr: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.
- [61] 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.
- [62] L. Ye, M. Rochan, Z. Liu, and Y. Wang, "Cross-modal self-attention network for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10494–10503.
- [63] K. Yamazaki et al., "Vlcap: Vision-language with contrastive learning for coherent video paragraph captioning," in *Proc. IEEE Int. Conf. Image Process.*, 2022, pp. 3656–3661.
- [64] K. Yamazaki et al., "VLTinT: Visual-linguistic transformer-in-transformer for coherent video paragraph captioning," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 3081–3090.
- [65] Q. Zhang and Y.-B. Yang, "Rest: An efficient transformer for visual recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 15475–15485.
- [66] 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.
- [67] H. Bao et al., "Beit: Bert pre-training of image transformers," 2021, *arXiv:2106.08254*.
- [68] X. Chu et al., "Twins: Revisiting the design of spatial attention in vision transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 9355–9366.
- [69] L. Wang et al., "BANet: Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images," *Remote Sens.*, vol. 13, no. 16, 2021, Art. no. 3065.
- [70] R. Li et al., "ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 181, pp. 84–98, 2021.
- [71] R. Li et al., "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5607713.
- [72] Z. Xu, J. Geng, and W. Jiang, "MMT: Mixed-mask transformer for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5613415.
- [73] K. Yamazaki et al., "AerialFormer: Multi-resolution transformer for aerial image segmentation," 2023, *arXiv:2306.06842*.
- [74] M. Contributors, "OpenMMLab semantic segmentation toolbox and benchmark," 2023. Accessed: Jun. 10, 2020. [Online]. Available: <https://github.com/open-mmlab/mmssegmentation>
- [75] L.-C. Chen et al., "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.
- [76] L. Gao et al., "SRANet: Semantic relation aware network for semantic segmentation of remote sensing images," *J. Appl. Remote Sens.*, vol. 16, no. 1, 2022, Art. no. 014515.

- [77] H. AlMarzouqi and L. S. Saoud, "Semantic labeling of high resolution images using EfficientUNets and transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4402913.



**Tao Wang** received the B.S. degree in information and computational science from Henan Normal University, Xinxiang, China, in 2005, and the M.S. degree in mathematics from Xidian University, Xi'an, China, in 2008. He is currently pursuing a Ph.D. degree in Agricultural Information Engineering from the Northwest A&F University, Yangling, China

He is currently an Associate Professor with the College of Information Engineering, Tarim University, Alaer, China, and a Member of the Key Laboratory of Tarim Oasis Agriculture (Tarim University), Ministry

of Education, Alaer. His research interests include deeplearning, computer vision, and remote sensing image processing.



**Chao Xu** (Member, IEEE) received the B.S. degree in electronic information engineering and the Ph.D. degree in information and communication engineering from Xidian University, Xi'an, China, in 2009 and 2015, respectively.

From 2015 to 2017, he was a Postdoctoral Researcher with the School of Telecommunications Engineering, Xidian University. He is currently a Professor with the College of Information Engineering, Northwest A&F University, Yangling, China, and employed as a Researcher with the Key Laboratory

of Agricultural Internet of Things, Ministry of Agriculture and Rural Affairs, Yangling, China. His research interests include artificial intelligence, computer vision, and pattern recognition.



**Bin Liu** (Member, IEEE) received the B.S. degree in computer science and technology from the Shaanxi University of Science and Technology, China, in 2004, the M.S. degree in technology with a major in parallel computing and cloud computing from Yunnan University, China, in 2010, and the Ph.D. degree in electronic and information engineering from Xi'an Jiaotong University, China, in 2014.

He is currently an Associate Professor in the College of Information Engineering, Northwest A&F University, Yangling, China. His research interests

include deep learning and computer vision.



**Guang Yang** received the B.S. degree in Computer Science and Technology from Northwest A&F University, Yangling, China, in 2022. He is currently pursuing the M.S. degree in Computer Science and Technology at Northwest A&F University, Yangling, China.

His research interests include computer vision and remote sensing image semantic segmentation.



**Erlei Zhang** (Member, IEEE) received the Ph.D. degree in electronic engineering from Xidian University, Xi'an, China, in 2015.

From 2018 to 2020, he was a Postdoctoral Fellow with the UT Southwestern Medical Center, USA and Northwest University, China. He is currently an Associate Professor with the College of Information Engineering, Northwest A&F University, Yangling, China. His research interests include pattern recognition, machine learning, and remote sensing image analysis and understanding.



**Dangdang Niu** received the M.S. and Ph.D. degrees in computer software and theory from the College of Computer Science and Technology, Jilin University, Changchun, China, in 2015 and 2018, respectively.

He is currently an Associate Professor with the College of Information Engineering, Northwest A&F University, Yangling, China. His research interests include artificial intelligence, machine learning, combinatorial optimization, and automated reasoning.



**Hongming Zhang** (Member, IEEE) received the B.S. degree in computer science and technology, the M.S. degree in cartography and geographic information systems, and the Ph.D. degree in land resources and spatial information technology from Northwest A&F University, Yangling, China, in 2003, 2008, and 2012, respectively.

He is currently a Professor with the College of Information Engineering, Northwest A&F University. His research interests include computer vision and remote sensing image analysis and understanding.