




Enhancing Efficient Global Understanding Network With CSWin Transformer for Urban Scene Images Segmentation

Jie Zhang , Mingwen Shao , *Member, IEEE*, Yuanjian Qiao , and Xiangyong Cao , *Member, IEEE*

Abstract—The global context is crucial to the semantic segmentation task of remote sensing (RS) urban scene imagery since objects have large size variations, high similarity, and mutual occlusion. However, the existing methods for extracting global context information have limitations when directly applied to very high-resolution RS images, mainly in high complexity of computation and memory consumption. To alleviate this limitation, we propose a novel Efficient Global Understanding semantic segmentation Network (EGUNet) to extract global context information efficiently for applicability to RS images. Specifically, EGUNet is a hybrid U-shaped architecture of convolutional neural networks (CNNs) and Transformer in which the encoder uses the CSWin Transformer to capture global semantic information, and the decoder uses the CNNs structure to recover local detail information. Thus, the proposed EGUNet has a powerful global extraction capability and local position information recovery capability. In addition, three effective modules are proposed to improve the segmentation accuracy to make EGUNet more applicable for urban scene image segmentation tasks. First, a feature adaptive fusion module is introduced in the decoder to improve the fusion of the deep semantics and the location detail features. Second, an adaptive atrous-spatial pyramid pooling is designed at the skip connections to enhance the multiscale understanding of high-level semantic context. Finally, we introduce a lightweight enhanced segmentation head to utilize the information from each decoder stage for segmentation. Extensive experimental results on ISPRS Vaihingen and Potsdam datasets demonstrate the exceptional segmentation accuracy of EGUNet, outperforming the state-of-the-art methods.

Index Terms—CSWin Transformer, global information extraction, remote sensing (RS) urban scene imagery, semantic segmentation.

Manuscript received 15 September 2023; revised 10 October 2023; accepted 24 October 2023. Date of publication 30 October 2023; date of current version 23 November 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2021 YFA1000102, in part by the National Natural Science Foundation of China under Grant 61673396, Grant 62272375, and Grant 62376285, and in part by the Natural Science Foundation of Shandong Province under Grant ZR2022MF260. (*Corresponding author: Mingwen Shao.*)

Jie Zhang is with the College of Computer Science and Technology, China University of Petroleum, Qingdao 266580, China, and also with the School of Humanities and Foreign Languages, Qingdao University of Technology, Qingdao 266520, China (e-mail: zjedu1225@126.com).

Mingwen Shao and Yuanjian Qiao are with the College of Computer Science and Technology, China University of Petroleum, Qingdao 266580, China (e-mail: smw278@126.com; yjqiao@s.upc.edu.cn).

Xiangyong Cao is with the School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China, and also with the Ministry of Education Key Laboratory for Intelligent Networks and Network Security, Xi'an 710049, China (e-mail: caoxiangyong@mail.xjtu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3328559

I. INTRODUCTION

SEMANTIC segmentation of remote sensing (RS) urban scene images is a practical computer vision (CV) task with broad applications in urban construction [1], land cover mapping [2], environment development management [3], and road extraction [4]. The purpose of RS urban scene image segmentation is to categorize each pixel within an image into a target semantic or background category. Notably, in recent years, due to the advances in aerial technology and modern satellite sensing technology, RS images captured by UAVs and satellites now cover richer spatial detail information and texture features [5], which makes RS images more complex. To cope with the complexity, powerful and sophisticated algorithms are required to effectively capture the intricate urban features, making RS image segmentation challenging.

As shown in Fig. 1, the challenges of the semantic segmentation of RS urban scene images are mainly in three aspects. First, ground objects are varied in size, so different scale receptive fields are needed to obtain multiscale feature information [6]. For example, “Buildings” in Fig. 1(a) have varied sizes, and the semantic categories of buildings cannot be accurately understood using only single-scale feature information [7]. Second, objects of different semantic categories may have high interclass similarity with similar size, material, and spectral features, and it is challenging to distinguish them using only local information. Observing Fig. 1(b), we can note the similarity between the Building’s roof and the “Impervious Surface” regarding their material and appearance. Similarly, from Fig. 1(c), it can be noticed that the skylight of the “Building” and the small “Car” are similar in appearance. Finally, since the RS images are taken from an overhead perspective [8], there is mutual occlusion between objects, leading to incomplete feature extraction, loss of context information, and semantic ambiguity. From the example in Fig. 1(d), the “Car” is obscured by the “Low Vegetation” from the top view. Analytically, the difficulty in solving the above three challenges (objects with multiple scales, high intraclass similarity, and mutual occlusion) is attributed to insufficient contextual information extraction. In other words, it is difficult to accurately segment RS images relying solely on local information, emphasizing the criticality of incorporating global information into the model, which is the challenge of the segmentation task [9]. Based on the above analysis, how to efficiently extract global information to apply to RS urban scene

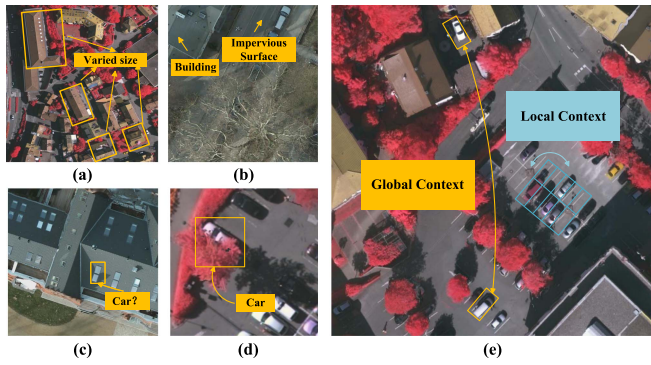


Fig. 1. Illustration of the dilemma in RS imagery segmentation and a schematic diagram of the relationship between local and global context information. (a) “Building” with varied size. (b) “Building” and “Impervious Surface” have similar materials. (c) Roof skylight and “Car” have a similar appearance. (d) “Car” with “Low Vegetation” covering the body. (e) For “Car,” the orange line shows the need for global information for the long-range dependency, and the blue line shows the role of local information for semantic understanding.

images motivates us to propose Efficient Global Understanding semantic segmentation Network (EGUNet).

To alleviate the aforementioned concerns, several deep-learning approaches have been explored to enhance the efficiency of semantic segmentation. Representatively, convolutional neural networks (CNNs)-based methods [10], [11], [12], [13], [14], [15] have achieved remarkable results due to their powerful ability to capture detailed low-level information and flexible hierarchical feature representations. Subsequently, some works focus on increasing the network’s magnitude to improve networks’ fitting ability, enlarging the receptive field of semantic understanding, and multiscale information representation [16]. Nevertheless, the convolution operation’s fixed receptive field still limits the ability of CNNs to model global context and establish long-range dependencies [17]. To enlarge the range of CNNs for context modeling, some current work [18], [19] uses attention mechanisms to address this limitation, while it brings huge computation and memory consumption. Alternatively, a multiscale feature fusion strategy [7], [20], [21] is also an effective solution for enlarging the receptive field by using atrous convolutions and feature pyramids, which can obtain multiscale information. The above two ways can enlarge the receptive field to some extent, but inherently, it is still not independent of the convolutional operation and does not model global information.

Recently, Transformer has been breaking new ground in CV research field due to its powerful global modeling capabilities. Unlike CNNs that process 2-D images directly, the Transformer converts the image into a sequence and models that global sequence, achieving excellent results in object detection [22] and image classification [23]. Driven by this, researchers have tried applying the Transformer to the semantic segmentation task [24]. More specifically, referring to CNNs, Transformer-based methods adopt the encoder–decoder framework with the skip connection to assist information recovery, which can effectively model global information and recover spatial location information. Nevertheless, when practically applying the

Transformer-based model to RS images, the computational complexity (quadratic to the imagery resolution [25]) is significantly high, seriously affecting the feasibility of applying Transformer-based methods in very high-resolution (VHR) RS urban scene imagery.

To tackle the aforesaid challenges, we design an innovative EGUNet to extract global context information efficiently while keeping the computational complexity within a reasonable limit. Specifically, we employ the CSWin Transformer [26] as the encoder backbone for extracting global context and CNNs to build a decoder for local spatial details recovery. Thus, the proposed EGUNet has a powerful global extraction capability and local position information recovery capability for RS imagery. Besides, three effective modules are proposed to enhance the segmentation accuracy of RS imagery. In each decoder stage, the feature adaptive fusion module (FAFM) is proposed to enhance the fusion of low-level detail features from the encoder stage and semantic features from the deep network. At the skip connection, an adaptive atrous-spatial pyramid pooling (AASPP) is designed to enhance the multiscale understanding of high-level semantic information. At the output stage, a lightweight enhanced segmentation head (ESegH) is introduced to better utilize the information from each decoder stage for segmentation. By adopting these modules, our EGUNet can attain superior effectiveness and precision in the semantic segmentation of urban scene imagery.

In conclusion, this article offers the following significant insights and contributions.

- 1) To enhance global modeling and reduce complexity, we construct a novel architecture, called EGUNet, which combines CSWin Transformer, as an encoder backbone network, with a CNNs decoder.
- 2) At each decoder stage, the FAFM is designed to adaptively enhance low-level detail features or high-level semantic features.
- 3) An AASPP is proposed to tackle object target and receptive field mismatch, while the ESegH enhances output for high-quality segmentation maps.

II. RELATED WORK

A. Semantic Segmentation on the Base of CNNs

Along with the emergence of deep learning, the semantic segmentation model based on CNNs is an attempt by researchers to explore end-to-end semantic segmentation methods using machine learning. To be noted, fully convolutional network (FCN) [12] improves on VGGNet [27] and is the pioneering end-to-end processing network for semantic segmentation implemented by full convolution. Since then, researchers have explored the effectiveness of using CNNs-based models in the semantic segmentation of RS and accomplished remarkable results [19], [28], [29], [30]. However, since the FCN architecture is too simple, the segmentation accuracy is unsatisfactory for some semantically complex VHR RS images. To enhance the model’s efficacy, based on the FCN, Ronneberger et al. [13] proposed a U-shaped symmetric structure model, called Unet, which has become a standard framework segmentation with its simple structure and accurate segmentation [31]. With Unet

network architecture, the following efforts focus on three main areas: improving the standard encoder standard backbone [27], [32], [33], [34], designing more efficient decoders [14], [32], and developing multiscale semantic understanding over skip connections [35], [36], [37].

Despite the commendable performance in capturing local features, CNNs-based approaches' ability to model global information remains a significant challenge due to the limited receptive fields [38]. This drawback becomes particularly evident when applying CNNs-based approaches to RS urban scene images characterized by high-resolution nature [38], intricate categories [39], and high object similarity [40], leading to inadequate semantic understanding and ambiguity.

B. Semantic Segmentation on the Base of Self-Attention Mechanism

Extensive empirical evidence has consistently validated the attention mechanism as a highly effective methodology for enlarging receptive fields and establishing extensive long-range dependencies. Unlike the strategy of increasing model scales, the self-attention mechanism simplifies the context problem by explicitly establishing relationships with relevant locations to build long-range dependencies. Specifically, self-attention allows the model to emphasize the essential features better while suppressing some interfering features. In the context of application examples, LANet [9] and AFNet [18] try to adopt the attention mechanism to integrate low-level details and high-level semantics. Moreover, CCNet [41] considers the limitation of hardware conditions and designs recurrent criss cross attention module to optimize self-attention computation, thereby enhancing the network's overall efficiency. Nevertheless, the abovementioned methods exclusively account for the attention relations of a single dimension and ignore the dependencies of other dimensions in the calculation process. To obtain multidimensional dependency information, SCAttNet [42] and DPA-Net [43] consider two dimensions of attention, channel and spatial, to refine features adaptively using a lightweight attention mechanism. Furthermore, HMANet [39] considers the attention of the category dimension for calibrating the category information. Unfortunately, the above attention mechanism still relies on convolutional operations and does not directly model global information.

C. Semantic Segmentation Based on Transformer

Recently, Transformer [44] has gained significant prominence in CV applications attributed to its strong global modeling and parallel processing powers, far superior to CNNs-based and self-attention models. Significantly, the general vision framework Visual Transformer (ViT) [23] first introduced the Transformer to image vision tasks, and its transformation of images into sequences makes global modeling of images feasible. Nonetheless, ViT introduces a computationally intensive self-attention calculation for all image sequences, resulting in significant computational complexity and prolonged training duration. Consequently, enhancing the training efficiency and optimizing the

training method of ViT have become primary research focuses in the CV domain. To illustrate, Chen et al. [45] design a general pretraining method that can be directly applied to construction, denoising, and rain removal tasks after fine-tuning. Meanwhile, T2T-ViT [46], TNT [47], and Twins [48] redesign the Transformer architecture to enhance Transformer performance through local self-attention mechanisms. Following that, TinyViT [49], DearKD [50], and DeiT [51] use distillation strategies to improve pretraining methods through enhancing the performance of smaller pretraining models by pretrained models, which saves memory cost and computational overhead. Nevertheless, the computational complexity of the ViT has been a challenge since it grows quadratically with increasing image resolution. To overcome this challenge, Swin Transformer [52], with its improved version [53], comes through the shifted window self-attention strategy, effectively reducing the computational complexity.

Driven by the achievements of the Transformer in CV, researchers attempt to employ Transformer on RS imaging tasks, such as hyperspectral image classification [54], building extraction [55], change detection [56], and, notably, semantic segmentation [57]. The current RS image segmentation methods follow Unet architecture, which could be mainly categorized into pure Transformer architecture and CNNs-Transformer hybrid architecture. Moreover, combining CNNs and Transformer, the latter hybrid architecture merges both strengths, which are effective in RS semantic segmentation tasks. Along this line, CCTNet [58], NT-Net [59], and WiCoNet [57] employ hybrid networks to tackle specific practical applications in RS, such as crop image segmentation, lake water extraction, and land object segmentation. Furthermore, researchers [8], [16] have also tried to use Swin Transformer as a backbone network in combination with CNNs for RS imaging segmentation tasks. In this regard, related research works [60], [61], [62], [63] have indicated that Swin Transformer as an encoder can be combined with different decoder architectures (e.g., Unet [13], PSP [20], and FPN [64]) for diverse tasks to achieve the optimal segmentation outcomes. Besides, UnetFormer [65], the current state-of-the-art (SoTA) network for semantic segmentation in RS, proposes a hybrid Transformer and CNN lightweight network for real-time urban scene segmentation, which looks similar to ours but is quite different. UnetFormer models local and global information with efficient local-global attention in the decoding stage, but the global information is not sufficiently extracted in the encoding stage. However, our EGUNet efficiently extracts global information in the encoding stage and efficiently recovers local information in the decoding stage, which is more in line with the semantic segmentation of RS images.

In order to make Transformer-based segmentation methods more feasible for VHR RS urban scene imagery segmentation, we present a novel EGUNet, which can not only extract global and local information effectively but also reduce the computational complexity to an acceptable level. In addition, using a lightweight self-attention mechanism, we develop the FAFM and AASPP to enhance semantic understanding, and propose an ESegH to enhance segmentation effects.

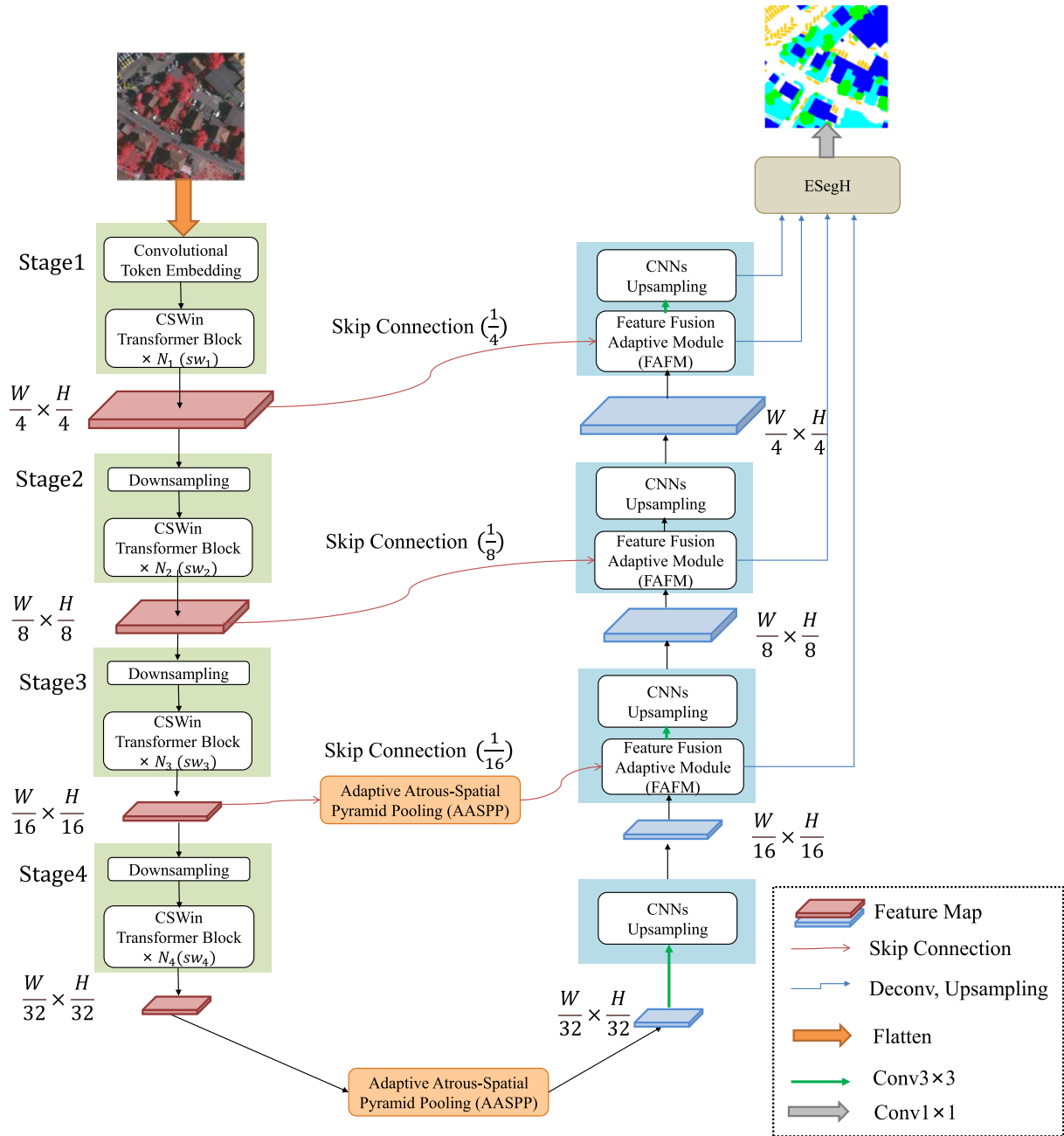


Fig. 2. Proposed EGUNet's general framework. The model presents a novel hybrid CNNs-Transformer architecture comprising the CSwin backbone network and three crucial modules (FAFM, AASPP, and ESegH). The backbone network effectively captures global information, while the latter modules focus on enhancing segmentation accuracy.

III. METHOD

This section begins by outlining the general organization of our proposed EGUNet and describing the relevant CSWin Transformer. After that, we introduce three crucial modules in EGUNet, namely the FAFM, the AASPP, and an ESegH.

A. Architecture

Fig. 2 demonstrates the overall structure of EGUNet, and it can be seen that EGUNet adopts an encoder-decoder architecture and adds skip connections to assist in locating information

recovery. Specifically, in the encoder stage, the EGUNet is divided into four stages for multiscale and hierarchical feature representation. In particular, for the i th stage, the size of the feature map is $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}$, and the number of channels is $\frac{C}{2^{i-1}}$, which is consistent with other common CNNs backbone network structures. During the first stage, a given image $X \in \mathbb{R}^{H \times W \times 3}$ first enters the token embedding layer (consisting of a convolution of size 7×7 with stride 4), where the image is divided into patch tokens of size $\frac{W}{4} \times \frac{H}{4}$ and the channels C . In each of the latter three stages, the downsampling module employs a 3×3 convolution with a stride of 2, reducing the dimensions

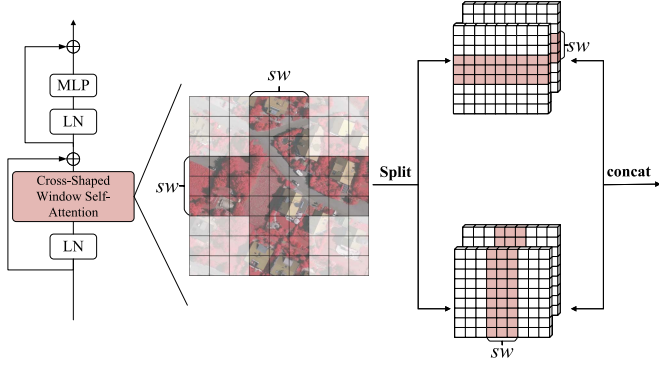


Fig. 3. Overview of the cross-shaped window self-attention mechanism of CSWin Transformer. This mechanism divides the multiheads into two groups, performing horizontal and vertical strip self-attention with strip width sw separately. The computations of the two self-attention groups are performed in parallel and finally concatenated together.

of the feature map by half while simultaneously increasing the channel count twofold.

After the encoder stage, the model obtains a feature map of size $\frac{W}{32} \times \frac{H}{32} \times 8C$ and sends this feature map to the decoder stage. The decoder exhibits a symmetrical structure to the encoder, consisting of four stages, mainly of the CNNs upsampling module and the FAFM. In detail, the upsampling module employs a 2×2 deconvolution to increase the feature map size twofold and reduce the channels by half. Besides, the FAFM uses 1×1 convolution to design a lightweight attention mechanism that facilitates the fusion of detailed and semantic features in an adaptive weight manner. For a more detailed explanation, refer to Section III-C.

In the four stages corresponding to the encoder and the decoder, we follow the classical Unet design and incorporate four skip connections to recover detailed information, such as location. Since the feature in deep stages has large receptive fields and rich deep semantic information, it could enhance its comprehension of object semantics if the model can understand deep semantic information at multiple scales. Therefore, in the skip connection of stage 3 and stage 4, we design the AASPP based on the attention mechanism described in Section III-D.

Finally, the model upsamples the output from each of the four decoder stages to a uniform size, feeds it into the ESegH, and outputs a segmentation map with the original input image's resolution.

B. CSWin Transformer Block

The CSWin Transformer presents an innovative cross-shaped window self-attention mechanism, enabling the effective modeling of global context information while minimizing computational overhead. This mechanism employs horizontal and vertical striped window blocks, forming a distinctive cross-shaped window, as depicted in Fig. 3.

For the horizontal direction, the CSWin Transformer block divides the input $X \in \mathbf{R}^{(H \times W) \times C}$ into sw horizontal strips that do not overlap and have the same width, $[X^1, X^2, \dots, X^M]$, where each strip consists of $sw \times W$ tokens. In particular, the

width of sw , which is not a fixed size, can be adjusted according to the computational complexity and the stage of the model. Thus, assuming that the dimensions of queries (Q), keys (K), and values (V) in the CSWin transformer are d_k , and the number of multiheaded attention heads is k , there then the attention result in the horizontal direction $H\text{-Attention}_k(X)$ is defined as follows:

$$X^i = [X^1, X^2, \dots, X^M], \text{ where } M = H/sw$$

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

where $i = 1, \dots, M$

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

where $X^i \in \mathbf{R}^{(sw \times W) \times C}$, $W_k^Q \in \mathbf{R}^{C \times d_k}$, $W_k^K \in \mathbf{R}^{C \times d_k}$, and $W_k^V \in \mathbf{R}^{C \times d_k}$ represent the projection matrix of the k th attention head Q, K, and V, respectively, and d_k is set to C/K . Correspondingly, the attention result in the vertical direction is similar to the definition in the horizontal direction, denoted as $V\text{-Attention}_k(X)$. Finally, the attention of the two directions is concatenated to form the self-attention result CSWin-Attention(X)

$$\begin{aligned} \text{CSWin-Attention}(X) &= \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_k) W^o \\ \text{head}_k &= \begin{cases} H\text{-Attention}(X), & k = 1, 2, \dots, \frac{K}{2} \\ V\text{-Attention}(X), & k = \frac{K}{2} + 1, \dots, K \end{cases} \end{aligned} \quad (2)$$

where $W^o \in \mathbf{R}^{C \times C}$ is the projection matrix that projects the self-attention results to the target dimension C , from this, the CSWin Transformer block in the encoder is calculated as

$$\begin{aligned} \hat{X}^l &= \text{CSWin-Attention}(\text{LN}(X^{l-1})) + X^{l-1} \\ X^l &= \text{MLP}(\text{LN}(\hat{X}^l)) + \hat{X}^l \end{aligned} \quad (3)$$

where LN denotes layer normalization, MLP means multilayer perception, and \hat{X}^l and X^l denote the output of CSWin-Attention, and the output of MLP, respectively.

C. Feature Adaptive Fusion Module

In each decoder stage, the role of FAFM is to better integrate low-level details and high-level semantics. To enhance the fusion process, the FAFM incorporates a lightweight attention mechanism, enabling adaptive selection between low-level and high-level features. In detail, Fig. 4 provides a visual depiction of the FAFM's structure.

Specifically, FAFM utilizes input from both the encoder, which provides detailed low-level information, and the decoder, which contributes high-level semantic information. For the i th stage, first, the low-level detail information S_{Li} passes through the 1×1 convolution and batch normalization (BN) layer to obtain the output result \hat{S}_{Li} . Then, two branches from the decoder use the high-level semantic information G_{Hi} as input, with one branch producing the semantic weight \hat{G}_{H1i} by 1×1 convolution, BN, and sigmoid activation, the other branch producing \hat{G}_{H2i} by 1×1 convolution and BN. Finally, \hat{S}_{Li} and the semantic weight \hat{G}_{H1i} are multiplied and then added

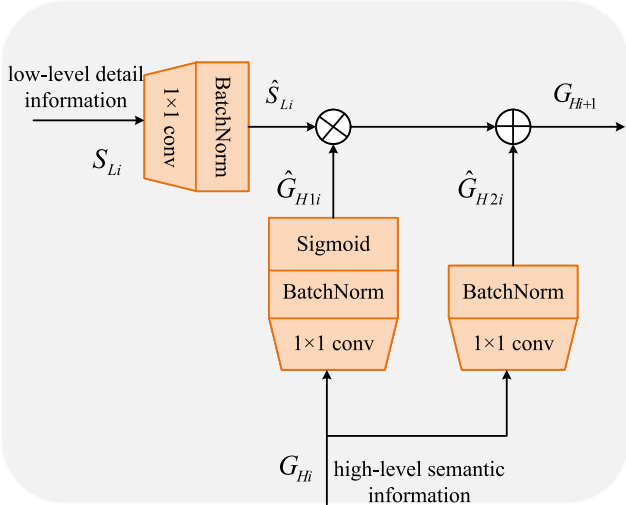


Fig. 4. Structural design of FAFM. FAFM achieves the fusion of low-level detailed features and high-level semantics through a lightweight self-attention mechanism.

with \hat{G}_{H2i} to obtain the final output G_{Hi+1} . It is worth noting that converting the high-dimensional semantic information into a weight matrix \hat{G}_{H1i} can adaptively adjust the weight values according to the importance of the semantic information. In turn, the final output G_{Hi+1} will adaptively adjust the proportion of the high-dimensional and low-dimensional information according to the dynamically changing weight matrix. The specific formula is as follows:

$$\begin{aligned} \hat{S}_{Li} &= \text{Conv}(\text{BN}(S_{Li})) \\ \hat{G}_{H1i} &= \sigma(\text{Conv}(\text{BN}(G_{Hi}))) \\ \hat{G}_{H2i} &= \text{Conv}(\text{BN}(G_{Hi})) \\ G_{Hi+1} &= \hat{S}_{Li} \times \hat{G}_{H1i} \oplus \hat{G}_{H2i} \end{aligned} \quad (4)$$

where σ stands for sigmoid activation function, BN stands for batch normalization, and Conv stands for 1×1 convolution operation.

D. Adaptive ASPP

In the encoder–decoder architecture, skip connections facilitate the delivery of vital information from the encoder to the decoder, enabling efficient feature recovery. Depending on the characteristics of the Unet model, for deep stages of the network containing rich semantic information, atrous-spatial pyramid pooling (ASPP) is an effective strategy to enlarge the receptive fields and enhance the understanding of semantic information. Complementarily, ASPP typically has five branches to obtain feature maps with different receptive field sizes: a 1×1 convolution branch, three 3×3 dilation convolution branches with different dilated rates, and a global average pooling branch. However, each of the five branches of ASPP has a fixed-size receptive field, which causes mismatches between the target objects and the receptive fields of the feature map when extracting context information. To alleviate this limitation, we

redesign the AASPP using the self-attention mechanism, which can adaptively use the attention map to weigh the multiscale feature map. Specifically, through the attention fusion module (AFM), MASPP adaptively enhances the branches that match receptive fields with the target object while suppressing other branches. Fig. 5 shows the specific structure of AASPP.

First, the input feature map F_{in} has to go through the traditional five branches of ASPP and output five feature maps F'_{in} with the same resolution but different receptive fields, where the dilated rates of the dilation convolution branches are [6, 8, 12]. Then, each branch's feature map passes through the AFM to generate the attention weight map F_{ω} . F'_{in} is multiplied by F_{ω} , then added to the original input F'_{in} to obtain the output feature map F'_{out}

$$F'_{out} = F_{\omega} \cdot F'_{in} + F'_{in} \quad (5)$$

where F'_{in} represents the output of the five branches of ASPP, and F'_{out} represents the feature map output after the AFM. F_{ω} represents the attention weight map, which can make the pixel point pay more attention to its related pixel point. Depending on the size of the receptive fields adapted to objects of different scales, F_{ω} will adaptively adjust the weights of the five branches, that is, enhance the weights of the branches adapted to the receptive fields of the target objects while suppressing the other branches. F_{ω} is defined as:

$$F_{\omega} = \text{Sigmoid}(\text{BN}(\text{Conv} \otimes F)) \quad (6)$$

where Conv denotes the 1×1 convolution operation, BN stands for batch normalization, and the formula illustrates that F_{ω} is differentiable.

Finally, the module concatenates the F'_{out} of the five branches and adjusts the channel quantity through 1×1 convolution, yielding the ultimate output feature map F_{out} .

E. ESegH Module

After FAFM and AASPP, feature maps for each stage in the decoder contain different levels of spatial location and semantic information, which are crucial for RS urban scene images. In existing methods, only a simple structure (i.e., utilizing the final output) is employed to generate the final segmentation map, which overlooks the decoder's crucial semantic information of different scales. To further elevate the segmentation performance, we introduce an ESegH module. Specifically, ESegH first upsamples the feature maps of the decoder's four stages to the same resolution and performs element summation, then adjusts the number of channels by two-layer convolution and eventually generates a semantic segmentation map. Fig. 6 illustrates the structure of ESegH.

IV. EXPERIMENTS

A. Datasets

To substantiate the efficacy and generalization performance of EGUNet, we conducted comprehensive experiments by comparing its performance with several existing methods on the

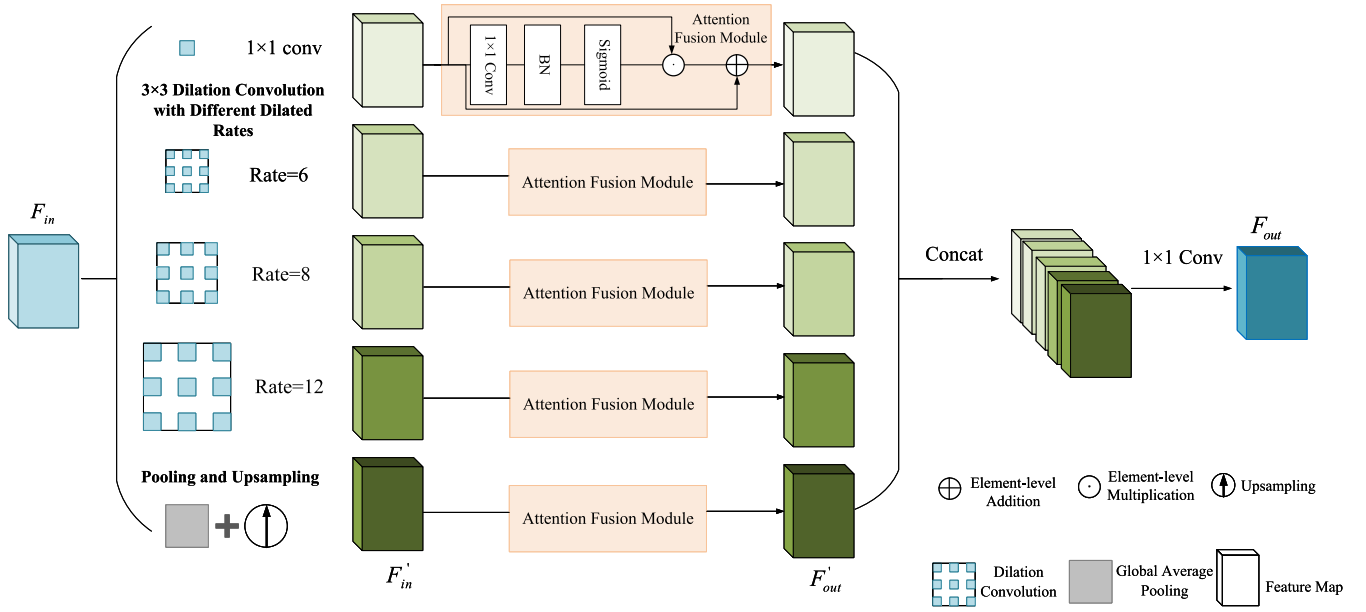


Fig. 5. Structural design of AASPP. AASPP effectively incorporates a lightweight self-attention module, AFM, to select multiscale features adaptively.

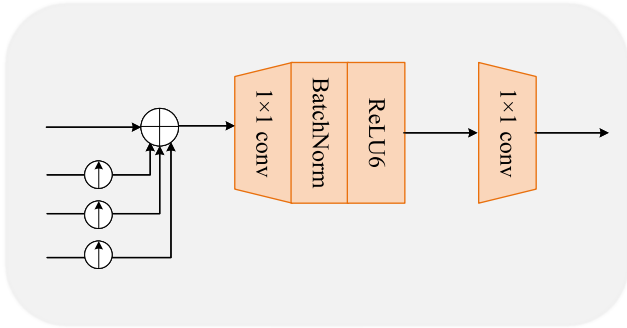


Fig. 6. Structural design of ESegH. ESegH efficiently fuses the features from the decoder's four stages to generate the final segmentation map in a lightweight manner.

ISPRS-provided aerial RS image datasets for the Vaihingen [66] area and Potsdam [67] area in Germany.

Vaihingen: The Vaihingen dataset consists of 33 RS urban scene images of different sizes with an average resolution of 2496×2064 pixels, 16 of which have feature category labels, and the remaining 17 have no label. The images in this dataset have been corrected and processed and are three-channel orthophoto images (Top images) with a ground sampling distance (GSD) of 9 cm. Each Top image in the Vaihingen dataset contains three multispectral bands: near-infrared, red, and green. In addition, it includes a digital surface model (DSM) and a normalized digital surface model (NDSM). However, our experiments solely utilized the Top image tiles, excluding the DSM and NDSM. For the city characterized as a small and scattered village, the dataset for Vaihingen consists of five target semantic categories (Impervious Surface, Tree, Low Vegetation, Building, and Car) and one background category (objects that differ from the other defined categories, called Clutter/Background). Referring to [8]

and [65], we select 17 images (with the same number IDs as in [8]) for testing, while the remaining 16 are assigned to the training set.

Potsdam: The Potsdam dataset contains 38 VHR urban scene images, all uniformly sized at 6000×6000 pixels. Similar to the Vaihingen dataset, the images in this dataset have been corrected and processed and are three-channel orthophoto images with a GSD of 5 cm. Notably, four multispectral bands (red, green, blue, and near-infrared) and DSM and NDSM are available in the dataset. Similarly, only three bands (red, green, and blue) are used in the experiments. Regarding urban characteristics, Potsdam epitomizes a classic historical city characterized by its expansive building complexes, tightly woven streets, and dense architectural structures. Consistent with the Vaihingen dataset, the dataset categories encompass five categories for the foreground and one category for the background. Referring to [8] and [65], we select 14 images as the testing set, while the remaining 24 are assigned to the training set.

Fig. 7 showcases the proportion of each semantic label within the two datasets mentioned. When quantitatively evaluating these datasets, we adopt the approach presented in [8], [65], and [68] and exclude the “Clutter/Background” category.

B. Implementation Details

We conduct experiments based on Python 3.9 and PyTorch 1.13.0 under Windows 10 OS and train the model using a single NVIDIA GTX 3090ti GPU. To accelerate convergence, the model uses AdamW [69] to optimize with a learning rate $6e-4$ (referring to the setting of [65]), adjusting the learning rate with a cosine strategy.

To enhance the training process by augmenting data diversity, we adopt a random cropping strategy, resizing images to 512×512 , and employ various data augmentation techniques,

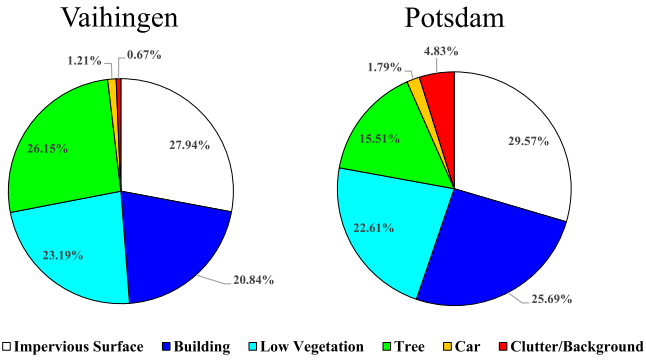


Fig. 7. Proportion of different semantic labels within Vaihingen and Potsdam datasets.

including random scaling (with factors of [0.5, 0.75, 1.0, 1.25, 1.5]), vertical flipping, and random horizontal flipping of the images. During the training, the training epoch is 105, and the batch size is 16. The encoder backbone network of EGUNet uses CSWin-Tiny in CSWin Transformer [26] with four stages in CSWin Transformer blocks of [1, 2, 21, 1], and the strip width sw of [1, 2, 7, 7]. It should be noted that the optimal results are indicated by *bold* values in all tables, whereas the second-best results are represented by underlined values.

Fig. 7 illustrates an imbalance in the distribution of ground objects across various categories within the two datasets. This imbalance can lead the model to prioritize categories with a larger proportion while neglecting those with a smaller proportion during training. To mitigate the impact of this problem, referring to [70] and [71], we use a joint supervised model training with dice loss [72] L_{Dice} and cross-entropy loss L_{CE} , with the total loss L calculated as follows:

$$L = L_{Dice} + L_{CE}. \quad (7)$$

C. Evaluation Metrics

When performing semantic segmentation on urban scene images, each pixel in the input image is assigned a category to achieve pixel-level classification. The resulting predictions can be classified into four types: True positive (TP), false positive (FP), true negative (TN), and false negative (FN). To assess the performance of the segmentation algorithm, two key evaluation metrics are used: precision (P) and recall (R). Precision measures the proportion of accurately predicted positive cases relative to the total number of positive cases, while recall indicates the ratio of predicted positive cases to the total number of positive cases. The calculation methods for precision and recall are as follows:

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}. \quad (8)$$

We employ the average $F1$ score ($F1$), the overall accuracy (OA), and the mean intersection over union (mIoU) as evaluation

metrics, which are calculated as follows.

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

$$IoU = \frac{TP}{TP + FP + FN}$$

$$OA = \frac{TP}{TP + FP + FN + TN}. \quad (9)$$

D. Comparison With Other Methods

The ISPRS Vaihingen and Potsdam datasets have emerged as widely adopted benchmarks for evaluating the performance of RS urban scene analysis methods. Recent research efforts have demonstrated notable advancements in achieving precise segmentation results on these datasets. In this study, we conduct a comprehensive comparative analysis between proposed EGUNet and existing SoTA approaches. Notably, the evaluated methods encompass a range of architectures, including lightweight CNNs-based networks, such as Unet [15], ShelfNet [73], and ABCNet [74], CNNs-based attention networks, such as FANet [75] and MANet [2], and CNNs-based networks tailored for RS image segmentation, such as EaNet [76]. In addition, we assess the performance of encoder-decoder networks exclusively built on Transformer as a comparison network to EGUNet, such as Segmenter [77], as well as hybrid semantic segmentation networks combining CNNs and Transformer architectures, namely BANet [78], BoTNet [79], STUNet [8], DCSwin [80], and UnetFormer [65]. Through comprehensive evaluation, we aim to provide insights into the strengths and limitations of the proposed method EGUNet, compared with SoTAs on the specific segmentation tasks involving the Vaihingen and Potsdam datasets. To mitigate computational expenses, all comparative networks in our experiments adopt lightweight backbone networks, such as ShelfNet, MANet, FANet, EaNet, ABCNet, BoTNet, and UnetFormer employing ResNet18, BANet utilizing ResT-Lite, Segmenter incorporating ViT-Tiny, while STUNet and DCSwin leveraging Swin-Tiny. In our proposed approach, EGUNet, we utilize CSWin-Tiny as the backbone network.

1) *Semantic Segmentation Performance Evaluation on the Vaihingen Dataset:* Table I illustrates the performance of our method, EGUNet, compared with other semantic segmentation networks on the Vaihingen dataset. EGUNet achieves the most exceptional segmentation performance, with Ave. $F1$, OA, and mIoU results reaching 90.7%, 91.2%, and 83.2%, respectively, far outperforming CNNs-based, Transformer-based, and hybrid CNNs-Transformer networks. As observed, the hybrid networks perform better overall than the ones based on CNNs or Transformer alone, owing to their ability to effectively utilize local information and global context. Notably, UnetFormer combines the lightweight transformer and CNNs to propose a new local-global attention mechanism, which surpasses the previous SoTA RS image semantic segmentation model and achieves an Ave. $F1$ score of 90.7%. Furthermore, our EGUNet has a significant performance improvement over UnetFormer, with an Ave. $F1$ improvement of 1.0% and mIoU improvement of

TABLE I
COMPARISON OF SEGMENTATION RESULTS OF EGUNET WITH OTHER EXISTING METHODS ON THE VAIHINGEN DATASET

Methods	F1(%)					Evaluation Metrics		
	Impervious Surface	Building	Low Vegetation	Tree	Car	Ave.F1(%)	OA(%)	mIoU(%)
Unet (2015) [15]	89.5	91.7	80.9	88.2	75.0	85.1	87.6	74.5
ShelfNet (2019) [73]	91.8	94.6	83.8	89.3	77.9	87.5	89.8	78.3
FANet (2020) [75]	90.7	93.8	82.6	88.6	71.6	85.4	88.9	75.6
EaNet (2020) [76]	91.7	94.5	83.1	89.2	80.0	87.7	89.7	78.7
ABCNet (2021) [74]	90.5	92.7	81.4	88.2	76.3	85.8	88.3	75.6
BANet (2021) [78]	90.8	93.7	82.4	88.9	79.3	86.9	88.9	77.2
MANet (2021) [2]	92.2	95.1	83.2	89.3	87.9	89.5	90.2	81.3
Segmenter (2021) [77]	89.8	93.0	81.2	88.9	67.6	84.1	88.1	73.6
BoTNet (2021) [79]	89.9	92.1	81.8	88.7	71.3	84.8	88.0	74.3
STUNet (2022) [8]	91.4	94.1	83.2	89.1	82.6	88.1	88.6	77.2
DCSwin (2022) [80]	90.3	92.0	81.6	88.1	77.7	86.0	88.1	75.8
UnetFormer (2022) [65]	92.4	95.6	83.5	89.5	87.4	89.7	90.5	81.6
Ours (EGUNet)	93.2	96.0	84.9	90.0	89.5	90.7	91.2	83.2

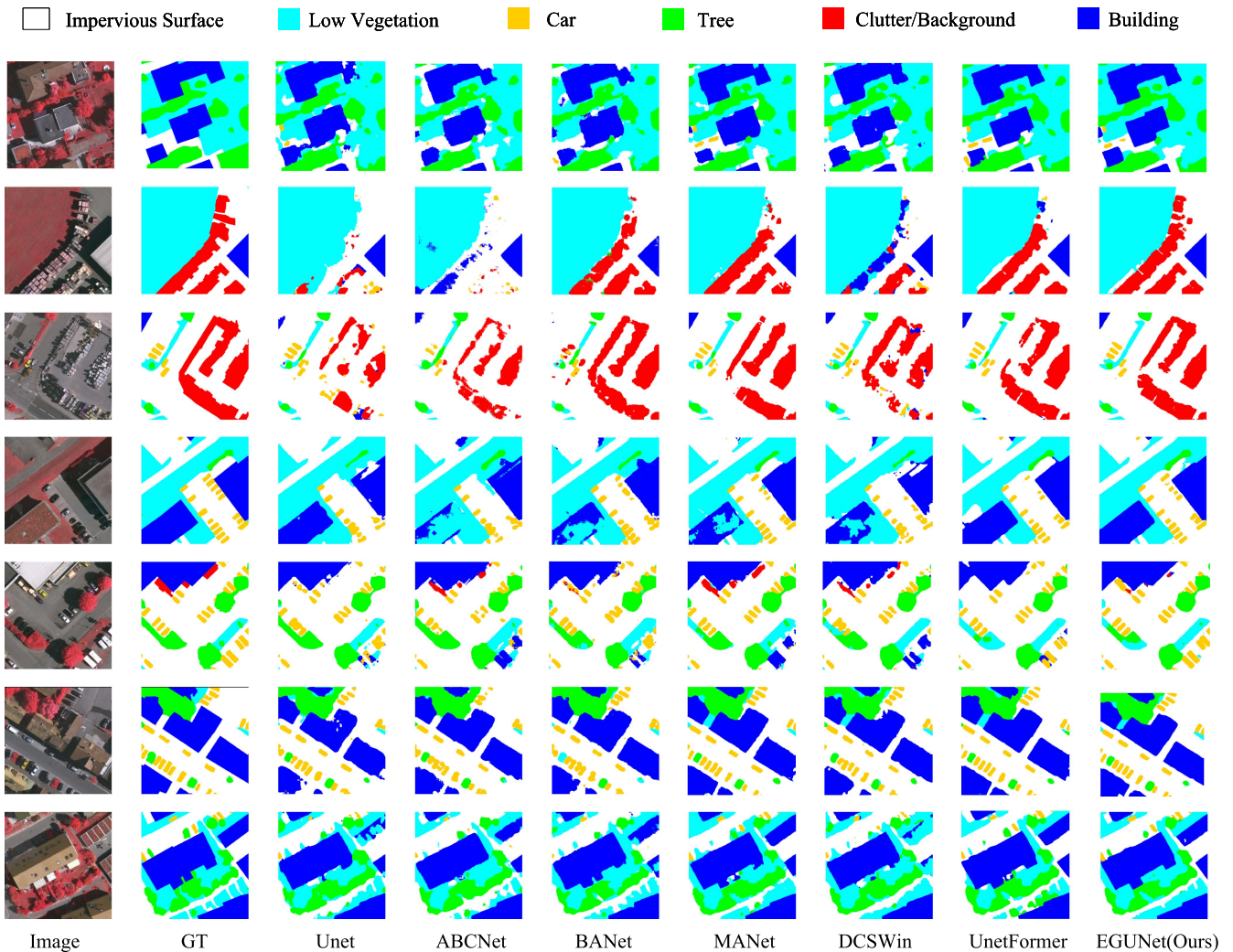


Fig. 8. Visualization comparison of semantic segmentation prediction results for the Vaihingen dataset.

1.2%. Among the five segmentation categories, the two with the most noticeable performance improvement are “Car” and “Low Vegetation” with an average $F1$ improvement of 2.1% and 1.4%, respectively. Such results show that EGUNet is most effective for categories with semantic occlusion problems (e.g.,

“Low Vegetation”) and categories that require global semantic understanding (e.g., “Car”).

Fig. 8 visualizes the semantic segmentation prediction results of several methods in Table I for a more intuitive comparison. The observations from the figure reveal that CNNs-based and

TABLE II
COMPARISON OF SEGMENTATION RESULTS OF EGUNET WITH OTHER EXISTING METHODS ON THE POTSDAM DATASET

Methods	F1(%)					Evaluation Metrics		
	Impervious Surface	Building	Low Vegetation	Tree	Car	Ave.F1(%)	OA(%)	mIoU(%)
Unet (2015) [15]	87.8	90.1	82.6	81.7	92.5	87.1	84.9	77.4
ShelfNet (2019) [73]	92.5	95.8	86.6	87.1	94.6	91.3	89.9	84.4
FANet (2020) [75]	92.0	96.1	86.0	87.8	94.5	91.3	89.8	84.2
EaNet (2020) [76]	92.0	95.7	84.3	85.7	95.7	90.6	88.7	83.4
ABCNet (2021) [74]	87.7	91.7	81.6	79.9	90.4	86.2	84.5	76.1
BANet (2021) [78]	90.8	94.2	84.7	84.8	93.3	89.6	87.9	81.3
MANet (2021) [2]	92.5	<u>96.4</u>	87.0	88.4	<u>96.0</u>	92.0	90.5	85.5
Segmenter (2021) [77]	91.5	95.3	85.4	85.0	88.5	89.2	88.7	80.7
BoTNet (2021) [79]	92.6	95.3	86.5	87.5	89.7	90.3	88.9	83.8
STUNet (2022) [8]	<u>93.2</u>	<u>96.4</u>	<u>87.6</u>	<u>88.6</u>	95.4	<u>92.2</u>	<u>90.9</u>	<u>85.8</u>
DCSwin (2022) [80]	89.3	92.9	83.6	83.8	92.9	88.5	86.7	79.6
UnetFormer (2022) [65]	91.8	95.4	86.3	87.5	95.4	91.3	89.7	84.2
Ours(EGUNet)	93.3	96.8	87.8	89.4	96.4	92.7	91.2	86.7

CNNs-based attention approaches, such as Unet and ABCNet, exhibit limitations in capturing context information, leading to semantic fragments and misclassifications in their segmentation maps. Similarly, entire Transformer-based methods, such as DCSwin, suffer from blurred object boundaries and imprecise segmentation due to inadequate spatial location and detailed information. Although the hybrid method UnetFormer partially alleviates these issues, challenges of misclassifying objects with multiple scales, high intraclass similarity, and mutual occlusion persist. In contrast, our proposed EGUNet performs better in reducing segmentation prediction errors, particularly for objects with varied sizes, similar appearances, and mutual occlusion. For instance, in the first row, the “Buildings” have different sizes and varied shapes, and the segmentation results of other methods are imprecise. At the same time, our EGUNet showcases a remarkable capability for multiscale building segmentation. Furthermore, in the second row, where there is a high visual similarity and material resemblance, other methods tend to misclassify the “Clutter/Background” as “Building,” whereas EGUNet achieves comparatively accurate segmentation. Then, in the fourth row of observation, the shadow cast by the “Building” covers the “Car,” inducing mutual occlusion that results in poor segmentation outcomes, while EGUNet successfully addresses this challenge with more precise segmentation.

2) *Semantic Segmentation Results for the Potsdam Dataset:* The segmentation results of SoTAs on the Potsdam dataset are presented in Table II, which further proves our proposed EGUNet’s effectiveness. Our approach demonstrates superior performance compared with other methods, achieving an Ave.F1 score of 89.4%, surpassing other methods by 0.8%. However, it is noteworthy that the CNNs–Transformer hybrid method does not substantially enhance segmentation accuracy. This situation can be attributed to the relatively scattered distribution of objects in the Potsdam dataset and the low mutual occlusion, making it less challenging and enabling even simple methods to yield satisfactory segmentation results. One more phenomenon to note is that the segmentation effect of DCSwin is worse than that of the CNNs method because DC-Swin cannot recover the spatial location information well, thus proving that spatial location information is crucial for semantic segmentation.

TABLE III
COMPARISON OF MODEL PARAMETERS AND COMPLEXITY

Methods	Params(M)	FLOPS(G)
Unet (2015) [15]	14.2	33.2
ABCNet (2021) [74]	14.0	31.3
BANet(2021) [78]	12.7	26.1
MANet (2021) [2]	35.9	155.5
TransUnet (2021) [17]	100.4	180.4
UnetFormer (2022) [65]	11.7	23.5
STUNet (2022) [8]	161.0	214.8
DCSwin (2022) [80]	66.9	140.3
EGUNet	37.2	97.0

Fig. 9 visualizes the prediction results of several segmentation methods. In the first two rows, when confronted with a wide range of building sizes and irregular shapes, current SoTA methods display limitations in accurately recognizing the object’s shape and the precise boundary locations, and thus, the segmentation is unsatisfactory. Confronted with this formidable challenge, EGUNet excels in its ability to accurately identify the intricate shapes of “Buildings” and smoothly and precisely delineate boundaries. Next, we visualize and analyze the high intraclass similarity of objects and mutual occlusion. In the third row, “Clutter/Background” is distributed as points in “Low Vegetation,” which is easily misclassified and challenging to locate accurately. Compared with other models, EGUNet has a more accurate segmentation performance. In the fifth row, the “Low Vegetation,” “Impervious Surface,” and “Building” are similar in appearance and can be easily misclassified because they often overlap each other. In this case, EGUNet effectively combines global context and local details, enabling precise segmentation. Similarly occurring in the penultimate row, our model demonstrates improved recognition of the “Low Vegetation” positioned between the “Tree” and “Building” while accurately segmenting the “Car” covered by “Tree.”

3) *Analysis of Efficiency:* For a comprehensive and holistic comparison, we evaluate the efficiency of the model by examining two factors: the floating point operations per second (FLOPS), which measures its complexity, and the number of parameters, which determines its memory requirements. The dual assessment provides a comprehensive understanding of the model’s efficiency. Table III presents the parameters and FLOPS of different methods in the same experimental setting.

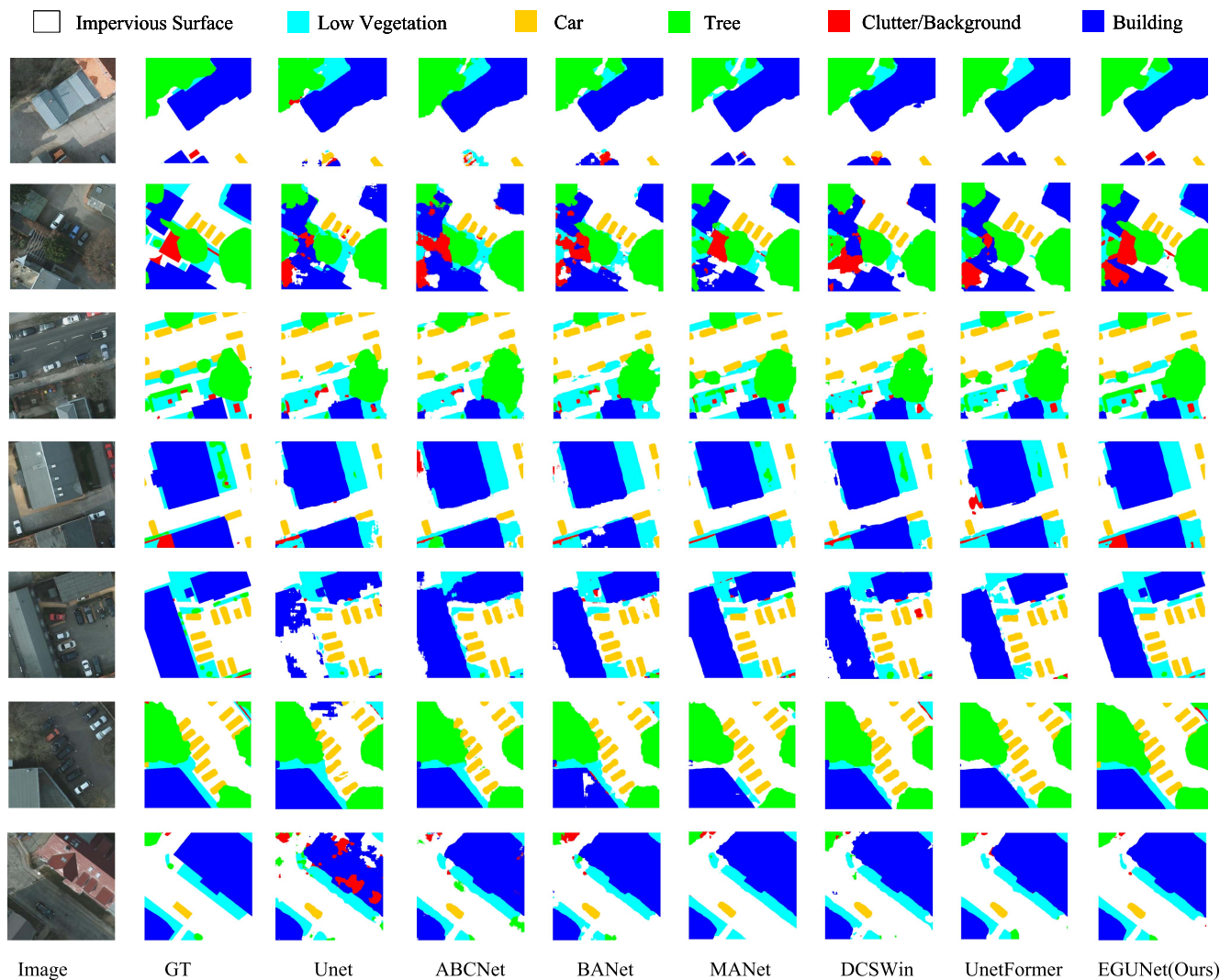


Fig. 9. Visualization comparison of semantic segmentation prediction results for the Potsdam dataset.

Overall, the CNNs-based method has fewer parameters and lower FLOPs than the Transformer-based method due to the typically higher model size and computational complexity of the Transformer. Also consistent with the above rule, our EGUNet features more parameters than the CNNs-based method while holding a notable advantage over the Transformer-based model. As the table data indicate, the number of parameters of EGUNet is 37.2 M, which is 63.2 M lower than TransUnet, 123.8 M lower than STUnet, and 29.7 M lower than DCSwin.

In contrast to UnetFormer, which employs a lightweight encoder, EGUNet exhibits a higher parameter count of 25.5 M due to the adoption of the large-scale CSwin Transformer as the encoder. Although EGUNet's use of a larger backbone network increases computational complexity, it ensures a robust framework for high-performance semantic segmentation tasks. A similar situation also appears in the comparison of FLOPs. The memory footprint of EGUNet is generally higher than CNNs-based methods but lower than most Transformer-based methods.

From the summary, the model scale of EGUNet is slightly larger than a few lightweight networks, which may limit the application on some mobile devices. Nevertheless, the enhanced performance of EGUNet in semantic segmentation of RS images, combined with its smaller model scale relative to most Transformer-based and SWin Transformer-based methods, highlights its continued value for research purposes.

E. Ablation Study

To validate the effectiveness of our proposed model and its three constituent modules (FAFM, MASPP, and ESegH), we conducted extensive ablation experiments on the Vaihingen dataset. In these experiments, we employed the widely adopted Unet as the baseline network to explore the impact of the backbone architecture on the overall performance. Furthermore, to examine three crucial modules, the baseline backbone network employ CSwin-Tiny, with the number of CSwin transformer blocks in four stages [1, 2, 21, 1], the strip width sw [1, 2, 7, 7],

TABLE IV
ABLATION EXPERIMENTS OF CSWIN TRANSFORMERS AS A BACKBONE NETWORK

Network Structure	Evaluation Metrics		
	mIoU(%)	Params(M)	FLOPS(G)
Baseline Unet	72.6	14.2	33.2
TransUnet	74.7	100.4	180.4
SwinUnet	75.8	32.4	113.5
CSWin+Unet	78.5	24.8	56.8

and the number of attention heads [2, 4, 8, 16]. These meticulous ablation experiments scrutinize the impact and significance of FAFM, MASPP, and ESegH within our EGUNet framework.

1) *Ablation Experiment of CSWin Transformer as Backbone Network*: To evaluate the effectiveness of CSWin Transformer as a backbone network, we use CSWin Transformer as an encoder. As a comparison, we use Unet as the encoder baseline network while TransUnet [17] and SwinUnet [81] as the comparison network. Specifically, TransUnet employs ViT as the encoder, SwinUnet adopts Swin Transformer as the encoder, and all networks use the ResNet18 network to build the decoder to ensure the experiment’s validity.

Table IV demonstrates that the segmentation performance of the encoders with ViT and Swin Transformer is significantly improved because both ViT and Swin Transformer model global context information, which is effective for segmenting VHR RS images. Notably, the encoder with the CSWin Transformer has the best performance with mIoU of 78.5%, which exceeds that of the baseline Unet by 5.9%, ViT by 3.8%, and Swin Transformer by 2.7%. Regarding efficiency, CSWin Transformer as an encoder has a network parameter count of 24.8 M and a computational complexity FLOPS of 56.8 G, much lower than the network parameter count and computational complexity of TransUnet and SwinUnet. Among them, the computational complexity reduction is the most obvious, which is 123.6 G lower than TransUnet and 56.7 G lower than SwinUnet. The ablation experiments conclusively establish CSWin Transformer as a superior choice for global encoding as an encoder backbone network.

2) *Effects of the FAFM*: As depicted in Table V, incorporating the FAFM module yields substantial improvements in segmentation performance, as evidenced by a notable 1.4% increase in mIoU. Notably, the “Car” category exhibits the most significant enhancement, with an impressive 2.1% rise in IoU, closely followed by “Low Vegetation” and “Tree,” both displaying a commendable 1.4% increase in IoU. These findings underscore the practical utility of the FAFM in enhancing the overall network performance, particularly in cases involving semantically ambiguous categories (such as “Low Vegetation” and “Tree”) and categories that require global semantic information understanding and precise localization (such as “Car”).

Intuitively, Fig. 10 compares visual segmentation results, highlighting the impact of incorporating the FAFM. In the first and second rows, adding the FAFM improves the segmentation accuracy of “Low Vegetation” adjacent to buildings. Similarly, in the third row, the “Low Vegetation” hidden by the shadows of the buildings is similar in appearance to the “Tree,” leading to

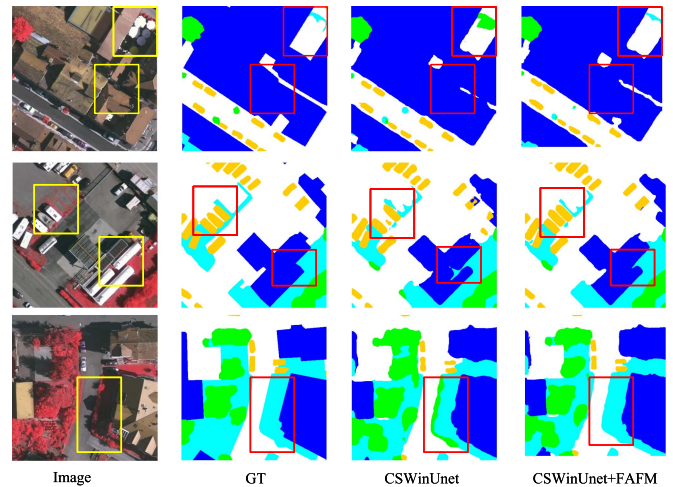


Fig. 10. Visualizing and comparing segmentation results of FAFM ablation experiments.

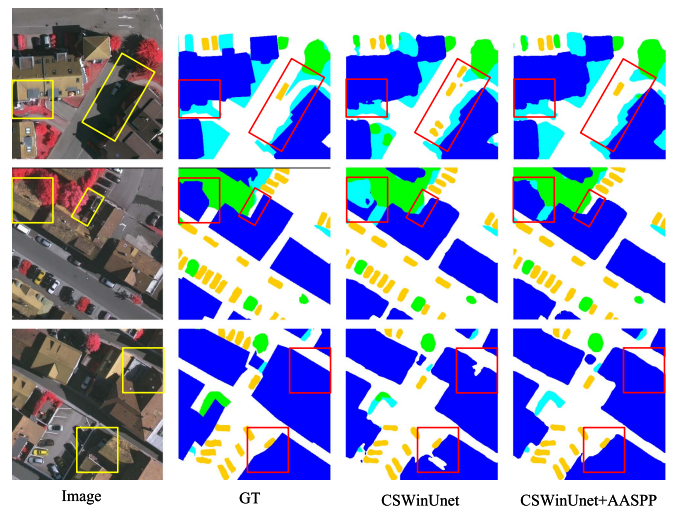


Fig. 11. Visualizing and comparing segmentation results of AASPP ablation experiments.

challenges in distinction. However, incorporating the FAFM facilitates more accurate differentiation between these categories, further demonstrating the validity of FAFM.

3) *Effects of the AASPP*: The experimental findings presented in Table V demonstrate that the inclusion of the AASPP yields a noteworthy enhancement of 1.1% in mIoU, underscoring the effectiveness of this module. Notably, the most significant improvements in segmentation accuracy are observed in the “Building” and “Car” categories, with corresponding increases in IoU of 1.5% and 1.4%, respectively. Given the multiscale sizes of “Building” in RS images and the requirement of multiscale context understanding for small objects, such as “Car,” achieving accurate segmentation necessitates a model capable of leveraging multiscale receptive fields. By incorporating the AASPP for multiscale feature extraction, the segmentation performance of these two categories can be significantly enhanced.

TABLE V
ABLATION EXPERIMENTS ON THE PROPOSED THREE MODULES FAFM, AASPP, AND ESEG H IN THE VAIHINGEN DATASET

Methods	IoU (%)					Evaluation Metrics
	Impervious Surface	Building	Low Vegetation	Tree	Car	mIoU(%)
CSWin+Unet	82.9	88.0	68.4	77.4	75.7	78.5
+FAFM	83.7	89.3	69.8	78.8	77.8	79.9
+AASPP	83.8	89.5	69.5	78.1	77.1	79.6
+ESegH	83.6	88.9	70.2	79.0	77.2	79.8
+FAFM+AASPP	84.8	89.8	71.4	79.8	78.2	80.8
+FAFM+ESegH	85.9	90.1	71.6	80.2	80.3	81.6
+AASPP+ESegH	85.8	90.6	72.1	80.3	80.0	81.8
+FAFM+AASPP+ESegH (EGUNet)	87.3	92.3	73.8	81.8	81.0	83.2

TABLE VI
QUANTITATIVE RESULTS OF ABLATION EXPERIMENTS ON AASPP MODULES ON THE VAIHINGEN DATASET

Methods	IoU(%)					Evaluation Metrics
	Impervious Surface	Building	Low Vegetation	Tree	Car	mIoU(%)
baseline	82.9	88.0	68.4	77.4	75.7	78.5
baseline+ASPP	82.8	87.5	69.0	78.4	76.1	78.8
baseline+AASPP	83.8	89.5	69.5	78.1	77.1	79.6

ASPP stands for disabling the AFM module in AASPP.

TABLE VII
ABLATION EXPERIMENTS ON THE LOCATION OF SKIP CONNECTIONS WITH THE ADDITION OF AASPP

Location of adding AASPP	IoU(%)					Evaluation Metrics
	Impervious Surface	Building	Low Vegetation	Tree	Car	mIoU(%)
No adding	82.9	88.0	68.4	77.4	75.7	78.5
Adding on skip4	82.7	87.8	69.4	77.2	76.2	78.7
Adding on skip4 and skip3	83.8	89.5	69.5	78.1	77.1	79.6
Adding on skip4, skip3 and skip2	83.5	88.7	69.7	78.2	76.3	79.3
Adding on all skip connections	82.5	88.1	69.5	77.7	76.0	78.8

As depicted in Fig. 11, in both the first and second rows, the edge of “Building” is integrated with “Low Vegetation” and “Tree,” posing a challenge for the model to precisely segment the “Building” and its boundaries. By incorporating the AASPP, the model’s segmentation performance for the “Building” category and its boundaries is significantly enhanced. Another challenging aspect is the impact of lighting conditions. For instance, in the first row, “Car” is entirely covered by the shadow of “Building.” Similarly, in the third row, the shadows of the buildings overshadow the low buildings and the road background. By visualizing the results, we can see that introducing AASPP effectively improves the occlusion problem of the “Building” and “Car” categories.

To further validate the effectiveness of the proposed module of AFM in AASPP, we adopt the Unet network of CSWin Transformer as the baseline network and use ASPP as the control group for ablation experiments. Initially, we introduce ASPP to skip connections in the third and fourth stages, as summarized in Table VI. These modifications yield segmentation results that exhibit a modest enhancement of 0.3% in terms of mIoU. Nevertheless, after adding AFM to ASPP, the mIoU is significantly improved by 1.1% over the baseline, underscoring the effectiveness of AFM and, consequently, AASPP.

Subsequently, the optimal placement for the introduced AASPP is further analyzed. Table VII gives that we conduct four experiments in which AASPP is added at skip 4, skips 3 and 4, skips 2–4, and all skips. The experimental findings

conclusively demonstrate that the best results were obtained by adding AASPP at skips 3 and 4, with a mIoU of 79.6%. Accordingly, we added AASPP at skips 3 and 4.

4) *Effects of the ESegH Module:* The results presented in Table V demonstrate a noticeable improvement in the mIoU of the model, with an increase of 1.3% following the integration of the ESegH module. Remarkably, the data highlight the substantial impact of ESegH on the “Low Vegetation,” exhibiting a notable increase of 1.9% in IoU. Further analysis, supported by the visualization results depicted in Fig. 12, elucidates the efficacy of the ESegH in handling challenging scenarios characterized by mutual obscuration. Specifically, in the first and third rows, “Low Vegetation” is adjacent to “Building” and shadowed by “Building,” rendering accurate segmentation challenging. However, the ESegH effectively mitigates this issue by enhancing the accuracy of object segmentation, even in scenarios with mutual obscuration. Furthermore, in the second row, the similarity between the appearance of the “Low Vegetation” and the vegetation-covered roof of the “Building” compromises segmentation accuracy. Nevertheless, using ESegH, the semantic area of the vegetation-covered roof can be more precisely delineated, significantly improving the overall segmentation results.

In addition, we investigate the combined effect of modules under the EGUNet. As demonstrated in Table V, the mIoU is improved by 2.3% after adding FAFM and AASPP modules, 3.1% after adding FAFM and ESegH modules, and 3.3% after adding AASPP and ESegH modules. Notably, when all three modules (FAFM, AASPP, and ESegH) are employed

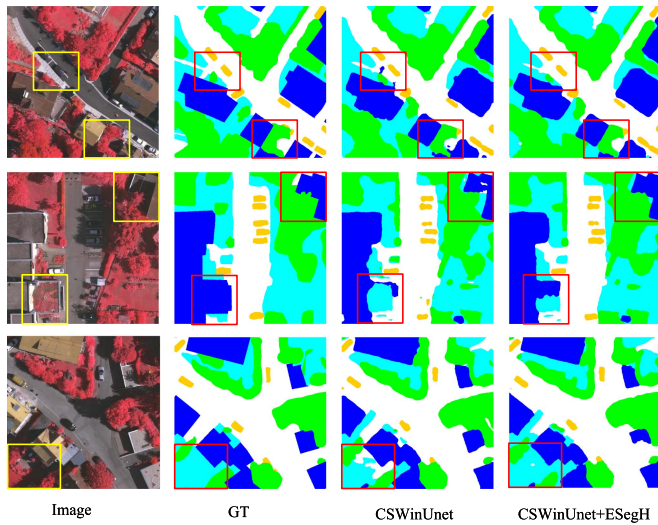


Fig. 12. Visualizing and comparing segmentation results of ESegH ablation experiments.

simultaneously, the most significant mIoU improvement is achieved, surpassing the CSWinUnet baseline by 4.7%. Consequently, our EGUNet effectively harnesses the synergistic benefits of the FAFM, AASPP, and ESegH modules, thereby yielding optimal segmentation performance.

V. CONCLUSION

In this article, we focus on developing a semantic segmentation method suitable for VHR urban scene imagery, aiming to achieve efficient global understanding with low computational complexity and memory requirements. By integrating CSWin Transformer as the encoder backbone network into Unet architecture, we construct a CNNs–Transformer hybrid RS image semantic segmentation network called EGUNet. Our proposed EGUNet incorporates critical components, including FAFM, AASPP, and ESegH. Specifically, the FAFM enables the adaptive fusion of local detail information and deep semantic information, while AASPP effectively learns multiscale semantic features. Furthermore, ESegH improves semantic segmentation accuracy through a lightweight fusion from each encoder stage.

Despite achieving SoTA results on the ISPRS Vaihingen and Potsdam datasets, there are still many limitations. For example, EGUNet does not segment the edges of objects well, and the boundaries are not precisely aligned with the shape of the objects. Future work will focus on enhancing the accuracy of semantic segmentation by exploring methods to refine boundary segmentation for diverse object types in RS images.

REFERENCES

- [1] F. Mohammadimanesh, B. Salehi, M. Mahdianpari, E. Gill, and M. Molinier, "A new fully convolutional neural network for semantic segmentation of polarimetric SAR imagery in complex land cover ecosystem," *ISPRS J. Photogrammetry Remote Sens.*, vol. 151, pp. 223–236, 2019.
- [2] R. Li et al., "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2021, Art. no. 5607713.
- [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 growthExamining 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] P. Shamsolmoali, M. Zareapoor, H. Zhou, R. Wang, and J. Yang, "Road segmentation for remote sensing images using adversarial spatial pyramid networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 4673–4688, Jun. 2021.
- [5] H. Bi, F. Xu, Z. Wei, Y. Xue, and Z. Xu, "An active deep learning approach for minimally supervised PolSAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9378–9395, Nov. 2019.
- [6] X. Chen et al., "Adaptive effective receptive field convolution for semantic segmentation of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3532–3546, Apr. 2021.
- [7] R. Liu, L. Mi, and Z. Chen, "AFNet: Adaptive fusion network for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7871–7886, Sep. 2020.
- [8] 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, 2022, Art. no. 4408715.
- [9] L. Ding, H. Tang, and L. Bruzzone, "LANet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 426–435, Jan. 2021.
- [10] C. Zhang, P. M. Atkinson, C. George, Z. Wen, M. Diazgranados, and F. Gerard, "Identifying and mapping individual plants in a highly diverse high-elevation ecosystem using UAV imagery and deep learning," *ISPRS J. Photogrammetry Remote Sens.*, vol. 169, pp. 280–291, 2020.
- [11] Z. Zheng, Y. Zhong, J. Wang, and A. Ma, "Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4096–4105.
- [12] 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.
- [13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.
- [14] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.
- [15] L.-C. Chen, "Rethinking atrous convolution for semantic image segmentation," *Comput. Res. Repository*, 2017.
- [16] 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, 2022, Art. no. 4408820.
- [17] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.
- [18] X. Yang et al., "An attention-fused network for semantic segmentation of very-high-resolution remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 177, pp. 238–262, 2021.
- [19] L. Mou, Y. Hua, and X. X. Zhu, "Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7557–7569, Nov. 2020.
- [20] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.
- [21] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 418–434.
- [22] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Representations*, 2021.
- [23] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.

- [24] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6881–6890.
- [25] S.-M. Hu, D. Liang, G.-Y. Yang, G.-W. Yang, and W.-Y. Zhou, "Jittor: A novel deep learning framework with meta-operators and unified graph executionJittor: A novel deep learning framework with meta-operators and unified graph execution," *Sci. China Inf. Sci.s*, vol. 63, no. 222103, pp. 1–21, 2020.
- [26] 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.
- [27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Comput. Sci.*, 2014.
- [28] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 60–77, 2018.
- [29] M. Wurm, T. Stark, X. X. Zhu, M. Weigand, and H. Taubenböck, "Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks," *ISPRS J. Photogrammetry Remote Sens.*, vol. 150, pp. 59–69, 2019.
- [30] Y. Wei, K. Zhang, and S. Ji, "Simultaneous road surface and centerline extraction from large-scale remote sensing images using CNN-based segmentation and tracing," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8919–8931, Dec. 2020.
- [31] N. He, L. Fang, and A. Plaza, "Hybrid first and second order attention UNet for building segmentation in remote sensing images," *Sci. China Inf. Sci.s*, vol. 63, pp. 1–12, 2020.
- [32] 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.
- [33] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5693–5703.
- [34] H. Zhang et al., "ResNeSt: Split-attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2736–2746.
- [35] K. Yue, L. Yang, R. Li, W. Hu, F. Zhang, and W. Li, "TreeUNet: Adaptive tree convolutional neural networks for subdecimeter aerial image segmentation," *ISPRS J. Photogrammetry Remote Sens.*, vol. 156, pp. 1–13, 2019.
- [36] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.
- [37] 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. Photogrammetry Remote Sens.*, vol. 162, pp. 94–114, 2020.
- [38] K. Nogueira, M. Dalla Mura, J. Chanussot, W. R. Schwartz, and J. A. Dos Santos, "Dynamic multicontext segmentation of remote sensing images based on convolutional networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7503–7520, Oct. 2019.
- [39] R. Niu, X. Sun, Y. Tian, W. Diao, K. Chen, and K. Fu, "Hybrid multiple attention network for semantic segmentation in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2021, Art. no. 5603018.
- [40] X. Hu, P. Zhang, Q. Zhang, and F. Yuan, "GLSANet: Global-local self-attention network for remote sensing image semantic segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023, Art. no. 6000105.
- [41] 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.
- [42] H. Li, K. Qiu, L. Chen, X. Mei, L. Hong, and C. Tao, "SCAttNet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 905–909, May 2021.
- [43] J. Li, J. Xiu, Z. Yang, and C. Liu, "Dual path attention net for remote sensing semantic image segmentation," *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 10, 2020, Art. no. 571.
- [44] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Process.: Syst. Demonstrations*, 2020, pp. 38–45.
- [45] H. Chen et al., "Pre-trained image processing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12299–12310.
- [46] L. Yuan et al., "Tokens-to-token ViT: Training vision transformers from scratch on ImageNet," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 558–567.
- [47] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 15908–15919, 2021.
- [48] X. Chu et al., "Twins: Revisiting the design of spatial attention in vision transformers," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 9355–9366, 2021.
- [49] K. Wu et al., "TinyViT: Fast pretraining distillation for small vision transformers," in *Proc. 17th Eur. Conf. Comput. Vis.*, 2022, pp. 68–85.
- [50] X. Chen, Q. Cao, Y. Zhong, J. Zhang, S. Gao, and D. Tao, "DearKD: Data-efficient early knowledge distillation for vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12052–12062.
- [51] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.
- [52] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.
- [53] Z. Liu et al., "Swin transformer v2: Scaling up capacity and resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12009–12019.
- [54] S. K. Roy, A. Deria, C. Shah, J. M. Haut, Q. Du, and A. Plaza, "Spectral-spatial morphological attention transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023, Art. no. 5503615.
- [55] L. Wang, S. Fang, X. Meng, and R. Li, "Building extraction with vision transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022, Art. no. 5625711.
- [56] Z. Zheng, Y. Zhong, S. Tian, A. Ma, and L. Zhang, "ChangeMask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection," *ISPRS J. Photogrammetry Remote Sens.*, vol. 183, pp. 228–239, 2022.
- [57] L. Ding et al., "Looking outside the window: Wide-context transformer for the semantic segmentation of high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022, Art. no. 4410313.
- [58] H. Wang, X. Chen, T. Zhang, Z. Xu, and J. Li, "CCTNet: Coupled CNN and transformer network for crop segmentation of remote sensing images," *Remote Sens.*, vol. 14, no. 9, 2022, Art. no. 1956.
- [59] H.-F. Zhong, Q. Sun, H.-M. Sun, and R.-S. Jia, "NT-Net: A semantic segmentation network for extracting lake water bodies from optical remote sensing images based on transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022, Art. no. 5627513.
- [60] 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, 2022, Art. no. 6517505.
- [61] L. Cui, X. Jing, Y. Wang, Y. Huan, Y. Xu, and Q. Zhang, "Improved Swin transformer-based semantic segmentation of postearthquake dense buildings in urban areas using remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 369–385, 2022.
- [62] Y. Liu, Y. Zhang, Y. Wang, and S. Mei, "BiTSRS: A bi-decoder transformer segmentor for high-spatial-resolution remote sensing images," *Remote Sens.*, vol. 15, no. 3, 2023, Art. no. 840.
- [63] Z. Zhang, X. Huang, and J. Li, "DWin-HRFormer: A high-resolution transformer model with directional windows for semantic segmentation of urban construction land," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–14, 2023, Art. no. 5400714.
- [64] A. Kirillov, R. Girshick, K. He, and P. Dollar, "Panoptic feature pyramid networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6399–6408.
- [65] 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.
- [66] I. Vaihingen, "2D semantic labeling dataset," Accessed: Apr. 2018.
- [67] I. Potsdam, "2D semantic labeling dataset," Accessed: Apr. 2018.
- [68] X. Li et al., "PointFlow: Flowing semantics through points for aerial image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4217–4226.
- [69] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2018.
- [70] L. Fidon et al., "Generalised Wasserstein Dice score for imbalanced multi-class segmentation using holistic convolutional networks," in *Brainlesion: Glioma, Mult. Sclerosis, Stroke Traumatic Brain Injuries: Third Int. Workshop*, 2018, pp. 64–76.

- [71] Q. Zhu, Y. Zheng, Y. Jiang, and J. Yang, "Efficient multi-class semantic segmentation of high resolution aerial imagery with dilated linknet," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 1065–1068.
- [72] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis.*, 2016, pp. 565–571.
- [73] J. Zhuang, J. Yang, L. Gu, and N. Dvornik, "ShelfNet for fast semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 847–856.
- [74] R. Li, S. Zheng, C. Zhang, C. Duan, L. Wang, and P. M. Atkinson, "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.
- [75] P. Hu et al., "Real-time semantic segmentation with fast attention," *IEEE Robot. Autom. Lett.*, vol. 6, no. 1, pp. 263–270, Jan. 2021.
- [76] X. Zheng, L. Huan, G.-S. Xia, and J. Gong, "Parsing very high resolution urban scene images by learning deep ConvNets with edge-aware loss," *ISPRS J. Photogrammetry Remote Sens.*, vol. 170, pp. 15–28, 2020.
- [77] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 7262–7272.
- [78] 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, no. 16, 2021, Art. no. 3065.
- [79] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16519–16529.
- [80] 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, 2022, Art. no. 6506105.
- [81] 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.



Jie Zhang received the M.S. degree in computer applications from the University of Chinese Academy of Sciences, Beijing, China, in 2019. She is currently working toward Ph.D. degree in geological engineering under the supervision of Prof. M. Shao with the School of Computer Science and Technology, China University of Petroleum (East China), Qingdao, China.

She is currently an Experimentalist with the Qingdao University of Technology, Qingdao, China. Her research interests include remote sensing image processing and deep learning.



Mingwen Shao (Member, IEEE) received the M.S. degree in mathematics from Guangxi University, Guangxi, China, in 2002, the Ph.D. degree in applied mathematics from Xi'an Jiaotong University, Xi'an, China, in 2005, and the Ph.D. degree in control science and engineering from Tsinghua University, Beijing, China, in 2008.

He is currently a Professor and a Doctoral Supervisor with the China University of Petroleum (East China), Dongying, China. His research interests include rough sets, data mining, machine learning, and generative adversarial learning.



Yuanjian Qiao received the M.S. degree in electrical engineering and automation from the Qilu University of Technology (Shandong Academy of Sciences), Jinnan, China, in 2021. He is currently working toward the Ph.D. degree in advanced scientific and engineering computing under the supervision of Prof. M. Shao with the School of Computer Science and Technology, China University of Petroleum (East China), Qingdao, China.

His research interests include image restoration and deep learning.



Xiangyong Cao (Member, IEEE) received the Ph.D. degree in mathematics and statistics from Xi'an Jiaotong University, Xi'an, China, in 2018.

From 2016 to 2017, he was a Visiting Scholar with Columbia University, New York, NY, USA. He is currently an Associate Professor with the School of Computer Science and Technology, Xi'an Jiaotong University. He is also with the Ministry of Education Key Laboratory For Intelligent Networks and Network Security, Xi'an. His research interests include statistical modeling and image processing.

processing and deep learning.