Swin Transformer Embedding UNet for Remote Sensing Image Semantic Segmentation

Xin He, Yong Zhou, Jiaqi Zhao, *Member, IEEE*, Di Zhang, Rui Yao, *Member, IEEE*, and Yong Xue, *Senior Member, IEEE*

***Abstract*— Global context information is essential for the semantic segmentation of remote sensing (RS) images. However, most existing methods rely on a convolutional neural network (CNN), which is challenging to directly obtain the global context due to the locality of the convolution operation. Inspired by the Swin transformer with powerful global modeling capabilities, we propose a novel semantic segmentation framework for RS images called ST-U-shaped network (UNet), which embeds the Swin transformer into the classical CNN-based UNet. ST-UNet constitutes a novel dual encoder structure of the Swin transformer and CNN in parallel. First, we propose a spatial interaction module (SIM), which encodes spatial information in the Swin transformer block by establishing pixel-level correlation to enhance the feature representation ability of occluded objects. Second, we construct a feature compression module (FCM) to reduce the loss of detailed information and condense more small-scale features in patch token downsampling of the Swin transformer, which improves the segmentation accuracy of small-scale ground objects. Finally, as a bridge between dual encoders, a relational aggregation module (RAM) is designed to integrate global dependencies from the Swin transformer into the features from CNN hierarchically. Our ST-UNet brings significant improvement on the ISPRS-Vaihingen and Potsdam datasets, respectively. The code will be available at https://github.com/XinnHe/ST-UNet.**

***Index Terms*— Global information embedding, remote sensing (RS), semantic segmentation, Swin transformer.**

I. Introduction

With the rapid development of aerospace technology and sensor technology, researchers can easily collect a large number of high-quality remote sensing (RS) images, which reflects the state of the ecological environment and traces of human activities [1]. Learning the knowledge contained in these images and effectively screening the information of interest has become the focus of intelligent interpretation of RS images [2]. Semantic segmentation has received widespread attention as a feasible solution. Its core goal is to identify the semantic category of each pixel in the image. At present, RS image semantic segmentation is applied in many real-world scenarios, such as urban planning [3]–[6], disaster assessment [7], [8], and agricultural production [9], [10].

In recent years, the rapid development of a convolutional neural network (CNN) has provided technical support for semantic segmentation. In particular, the fully convolutional network (FCN) [11] has played a vital role. Subsequently, the researchers put forward a lot of innovative works. In this process, the encoder–decoder structure exhibits excellent segmentation performance and has gradually become a popular structure configuration in semantic segmentation networks [12], [13]. The encoder is used to extract features, and the decoder restores the image resolution as finely as possible while fusing high-level semantic and low-level spatial information. For example, the U-shaped network (UNet) [14] utilized a decoder to learn the spatial correlation of the corresponding encoding stage by skip connections. Deeplab V3+ [15] introduced a decoder based on Deeplab V3 [16] to integrate spatial features, significantly improving network performance.

However, the particularities of ground objects (small scale, high similarity, and mutual occlusion) pose new challenges to semantic segmentation for RS images, as shown in Fig. 1. CNN-based models perform feature downsampling in the feature extraction process to reduce the amount of calculation, which easily causes small-scale features to be discarded [17], [18]. Ground objects with different semantic categories may have similar size, material, and spectral characteristics, which are difficult to distinguish. Besides, the occlusion problem usually leads to semantic ambiguity. Therefore, more global context information and fine spatial features are requested as clues for semantic reasoning [19].

CNN has advantages in spatial position representation, but it is difficult to model global semantic interaction and context information directly due to the locality of the convolution operation [20]. Existing methods apply the attention mechanism to solve this problem. DANet [21] constructed

Manuscript received September 1, 2021; revised December 10, 2021; accepted January 14, 2022. Date of publication January 19, 2022; date of current version March 14, 2022. This work was supported in part by the Natural Science Foundation of Jiangsu Province under Grant BK20201346; in part by the National Natural Science Foundation of China under Grant 61806206, Grant 62172417, and Grant 41871260; and in part by the Six Talent Peaks Project in Jiangsu Province under Grant 2015-DZXX-010 and Grant 2018-XYDXX-044. *(Corresponding author: Yong Zhou.)*

Xin He, Yong Zhou, Jiaqi Zhao, Di Zhang, and Rui Yao are with the School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China, and also with the Engineering Research Center of Mine Digitization, Ministry of Education of the People's Republic of China, Xuzhou 221116, China (e-mail: tb21170005b3ld@cumt.edu.cn; yzhou@cumt.edu.cn; jiaqizhao@cumt.edu.cn; zhang_di@cumt.edu.cn; ruiyao@cumt.edu.cn).

Yong Xue is with the School of Environment Science and Spatial Informatics, China University of Mining and Technology, Xuzhou 221116, China, and also with the School of Electronics, Computing and Mathematics, University of Derby, Derby DE22 1GB, U.K. (e-mail: 5878@cumt.edu.cn).

Digital Object Identifier 10.1109/TGRS.2022.3144165

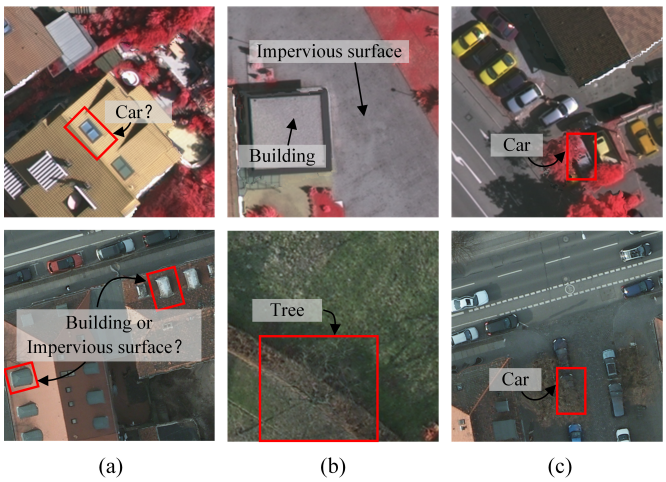

Fig. 1. Examples of the characteristics of RS images, where the first row of images is taken from the ISPRS Vaihingen dataset, and the second row is taken from the ISPRS Potsdam dataset. (a) Skylights on the roof have a similar appearance to "Car" and "Impervious Surface." (b) "Building" and "Impervious Surfaces" have the same material, and "Tree" is almost invisible among "Low Vegetation." (c) Large part of "Car" is obscured by "Tree."

long-range dependencies through parallel channel attention and position attention, while [22] proposed a "criss-cross" attention. In addition, a multiscale feature fusion strategy is also a common method. Zhao *et al.* [23] used different levels of feature summation (the same as FCN [11]) to describe comprehensive features. PSPNet [24] and its improved version UperNet [25] exploited contextual information through the pyramid pooling module. These methods aggregate the global information from the local features obtained by CNN instead of directly encoding the global context. Therefore, it is difficult to gain clear global scene information from RS images with complex backgrounds [26].

Recently, transformer's success has opened up new research ideas for modeling global relationships [27]. A transformer is a popular sequence prediction model in the field of natural language processing. Carion *et al.* [28] proposed DETR, which employed the encoder–decoder structure of the transformer to model the interaction between elements in the sequence. Analogous to CNN-based models, Chen *et al.* [29] designed a two-branch transformer structure to learn features of different scales. It is proven that multiscale feature representation is also effective to visual transformer (ViT). Following this structure, Swin transformer [30] is constructed and shows great potential in several dense prediction tasks. At present, the Swin transformer-based models have made great progress in medical image segmentation [31], [32], but its segmentation potential on RS images has not been confirmed.

In this article, to alleviate the deficiencies of CNN in global modeling, we propose a novel network framework for RS image semantic segmentation called ST-UNet, which utilizes the Swin transformer to assist UNet. As mentioned earlier, UNet is a U-shaped decoder-encoder network based on CNN, which realizes feature fusion between encoder and decoder by skip-connection layers. We take the encoder in UNet as the main encoder and the Swin transformer as the auxiliary encoder to form a parallel dual-encoder structure. Specifically, we build a unidirectional information stream from the auxiliary encoder to the main encoder through a well-designed relational aggregation module (RAM), which is the key component of our ST-UNet. Furthermore, the SIM is attached to the Swin transformer to explore the spatial correlation of global features, and the FCM is used to improve the segmentation accuracy of small-scale objects.

The main contributions of this article are given as follows.

1) We construct the spatial interaction module (SIM) to focus on the pixel-level feature correlation in the spatial dimension, thereby alleviating the semantic ambiguity caused by ground object occlusion. Besides, SIM compensates for the global modeling capabilities of the Swin transformer limited by its window mechanism.
2) We propose a feature compression module (FCM) in the auxiliary encoder to alleviate the omission of small-scale features during patch token downsampling. FCM can gather more features about small-scale objects and reduce the loss of detailed information.
3) To extract discriminative features in RS images, we design an RAM, which extracts channel-related information from the auxiliary encoder as a global clue to guide the main encoder. RAM can effectively distinguish ground objects with high similarity.

II. Related Work

A. Remote Sensing Image Semantic Segmentation Based on CNN

Due to the publication of some datasets [33]–[35] and contests, such as the IEEE Geoscience and Remote Sensing Society (IGARSS) data fusion contest,[1] the SpaceNet competition,[2] the DeepGlobe contest,[3] and the International Society for Photogrammetry and Remote Sensing (ISPRS) Benchmarks,[4] semantic segmentation for RS images based on CNN has received widespread attention. Zhang *et al.* [36] adopted the multibranch parallel convolution structure in HRNet [37] to generate multiscale feature maps and designed an adaptive spatial pooling module to aggregate more local contexts. Maggiori *et al.* [38] introduced a multilayer perceptron (MLP) into the segmentation network to produce better segmentation results. Qi *et al.* [39] proposed a spatial information inference structure based on recurrent neural network and 3-D convolution to learn the global spatial context and local visual features, which effectively solves the occlusion problem in road detection.

Moreover, some researchers paid attention to the feature extraction of small-scale features. Kammpffmeyer *et al.* [40] assembled patch-based pixel classification and pixel-to-pixel segmentation, which introduced uncertain mapping to achieve high performance on small-scale objects. Inspired by UNet, Dong *et al.* [17] proposed DenseU-Net, which realizes the aggregation of small-scale features through a dense fusion strategy. FactSeg [41] proposed a symmetrical dual-branch decoder consisting of a foreground activation branch and

[1] https://www.igarss2021.com/
[2] https://spacenetchallenge.github.io/
[3] http://deepglobe.org/challenge.html
[4] https://www.isprs.org/education/benchmarks.aspx

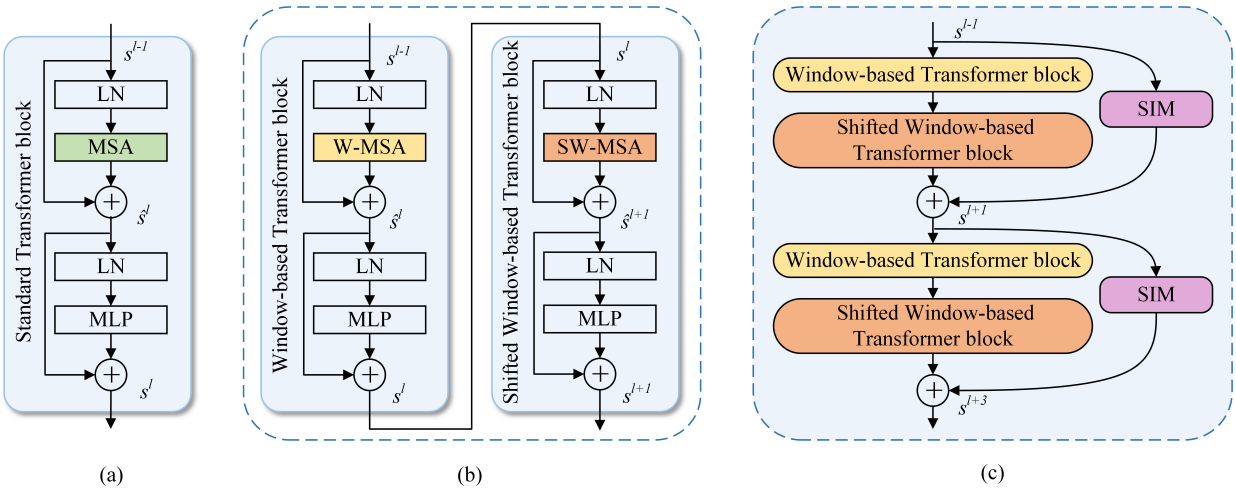

Fig. 2. (a) Structure of the standard transformer block [27]. (b) Two consecutive Swin transformer blocks [30] (renamed W-Trans block and SW-Trans block, respectively). (c) Swin transformer blocks with the proposed SIM.

a semantic refinement branch. The two branches performed multiscale feature fusion through skip connection, thereby improving the accuracy of small-scale object segmentation. Chen *et al.* [18] proposed an adaptive receptive field convolutional network to achieve a compromise between the feature extraction of large-scale objects and small-scale objects.

The prediction of object boundary is also a noteworthy aspect. Marmanis *et al.* [42] explicitly introduced the edge detection module [43] in a semantic segmentation network to supervise the boundary feature learning. ERN [44] suggested two simple edge loss reinforced modules to enhance the preservation of object boundaries.

The abovementioned CNN-based methods have promoted the development of RS image semantic segmentation. Instead of abandoning CNN, such as the recent transformer-based methods [28]–[30], we are committed to proposing a new framework to inherit the advantages of both.

B. Self-Attention Mechanism

Recently, the self-attention mechanism has been prevalent in computer vision tasks. Zhao *et al.* [45] and Li *et al.* [46] presented region-level attention and frame-level attention for video captioning. Zhao *et al.* [47] explored the effectiveness of pairwise self-attention and patchwise self-attention in image recognition. SENet [48] expressed the relationship between channels through the global average pooling layer to understand the importance of different channels automatically. CBAM [49] applied channel-level attention and spatial-level attention to adaptive feature refinement. Ding *et al.* [19] proposed a patch attention module to highlight the focus area of the feature map. Panboonyuen *et al.* [50] introduced channel attention blocks at each stage of the GCN [51] framework to hierarchically optimize the feature map. Su *et al.* [52] focused on similar objects in a small batch of images and encoded the interactive information between them through the self-attention mechanism. Different from the above methods, we consider vertical and horizontal pixel-level self-attention and combine multiple pooling layers to extract channel dependencies.

C. Vision Transformer

The transformer was first proposed for machine translation tasks [53] and surpassed the previous sequence transduction model based on complex recurrent or CNNs. Subsequently, transformer-based models have been widely used in various fields of natural language processing [54]. The standard transformer block consists of a multihead self-attention (MSA), a multiple layer perceptron (MLP), and a layer normalization (LN), as illustrated in Fig. 2(a). MSA played a key role in establishing global dependencies between input and output sequences.

Recent studies have indicated that a transformer is also suitable for computer vision tasks. Dosovitskiy *et al.* [27] was the first to apply transformer to image classification, turning image data into a sequence of tokens through splitting and flattening. Such a ViT achieved state-of-the-art performance under pretraining with large-scale datasets. Following that, Chen *et al.* [55] explored a universal transformer-based pretraining approach for image processing tasks. T2T-ViT [56] recursively aggregated adjacent tokens into one token to model the local structural features represented by surrounding tokens and reduce the number of tokens. Touvron *et al.* [57] proposed a new token-based distillation strategy to improve the training efficiency of the original ViT on the smaller Imagenet-1k dataset.

Nevertheless, ViT still has a huge training cost for intensive prediction tasks. It only outputs a low-resolution feature that cannot match the prediction target (the same resolution as the input image). Some work modified the ViT architecture to adapt to dense prediction tasks, such as semantic segmentation and object detection. SETR [58] regarded the transformer as an encoder that models the global context in each layer combined with a simple decoder to form a semantic segmentation network. Imitating the characteristics of the CNN backbones, PVT [59] introduced the pyramid structure into ViT to obtain multiscale feature maps. In detail, it flexibly controls the length of the transformer sequence through the patch embedding layer. Although PVT reduces the consumption of computing resources to a certain extent, its complexity is still quadratic

to image size. Therefore, Liu *et al.* [30] proposed Swin transformer based on the shifted window strategy, which limits the calculation of MSA to nonoverlapping windows while allowing cross-window information interaction. The Swin transformer only has linear computational complexity and achieves advanced performance in various vision tasks, including image classification, object detection, and semantic segmentation. Abandoning the MSA in the traditional transformer, it leverages window-based MSA (W-MSA) and shifted W-MSA (SW-MSA), as shown in Fig. 2(b).

With the Swin transformer as the backbone, Cao *et al.* [31] and Lin *et al.* [32] developed the U-shaped encoder–decoder framework for medical image semantic segmentation. In particular, Lin *et al.* [32] utilized the Swin transformer-based dual encoder with two input image scales to extract feature representations of different semantic scales.

However, TransUNet [20] and TransFuse [60] pointed out that the pure transformer segmentation network produces unsatisfactory results because transformer only focuses on global modeling and lacks positioning capabilities. Thus, they created a hybrid structure of CNN and transformer. TransUNet sequentially stacked CNN and transformer to form a new encoder structure, while TransFuse executed both in parallel and tried to fuse the two features. In addition, TransFuse used simple progressive upsampling [58] in the decoder of the transformer branch to restore the spatial resolution.

Inspired by these excellent works, we adopt the auxiliary encoder composed of the Swin transformer blocks to provide global context information for the CNN-based main encoder. As far as we know, the proposed ST-UNet is the first to apply the Swin transformer to the RS image segmentation task, which makes up for the shortcomings of pure CNNs and improves the segmentation accuracy.

III. Method

In this section, we first introduce the overall structure of the proposed ST-UNet and describe the Swin transformer involved. After this, three important modules in ST-UNet are introduced, namely, RAM, SIM, and FCM.

A. Network Structure

The overall architecture of our ST-UNet is shown in Fig. 3. As a hybrid of the Swin transformer and UNet, our ST-UNet follows the excellent structure of UNet, in which skip connection layers connect the encoder and the decoder. In particular, ST-UNet constructs the dual encoder composed of a CNN-based residual network and the Swin transformer, which transmits information through RAM to fully obtain the discriminative features of RS images. In addition, we design SIM and FCM to further improve the performance of the Swin transformer.

For a given RS image $X \in \mathbb{R}^{H\times W\times 3}$, ViT divided the image data into nonoverlapping patches to analogize the "tokens" of sequence data. However, there is no internal correlation between tokens, while image patches are just the opposite. The pixels of the same object in a patch are usually clustered together, which has a strong semantic correlation. Therefore, to avoid losing the continuity of semantic information in the initial input stage, we obtain overlapping patch tokens from each image through convolution. In the experiment, the patch size is 8×8, and the overlap rate is 50%. Then, the linear embedding layer flattens and projects these patches into dimension C_1. These patch tokens are put into the auxiliary encoder stacked by Swin transformer blocks. The auxiliary encoder has four feature extraction stages, and the output of each stage is defined as $\mathbf{S}_n$, where $n = 1, 2, 3, 4$. Standard Swin transformer blocks include two types, namely, window-based transformer (W-Trans) and shifted W-Trans (SW-Trans). Especially, we propose the SIM to establish pixel-level information exchange, which is attached to the Swin transformer blocks. SIM can effectively compensate for the limitations of window-based self-attention and alleviate the problem of semantic ambiguity caused by occlusion. Besides, to obtain multiscale features while matching with the feature resolution of the main encoder, we build the FCM to form a four-stage hierarchical feature encoding structure by shortening the length of the patch token. In this process, the proposal of FCM can reduce the omission of small-scale object features. The output resolution of stage n is $(H/(2^{n+1}) \times (W/(2^{n+1}))$, and the dimensions are $2^{n-1}C_1$.

The original RS image X is first fed to ResNet50 with half the compression on the channel to obtain the deep features in the main encoder. The output feature map of the nth residual block can be expressed as $\mathbf{A}_n \in \mathbb{R}^{(H/(2^{n+1}))\times(W/(2^{n+1}))\times 2^{n-1}C_2}$. Here, $C_2 = 128$. Then, $\mathbf{A}_n$ and the output $\mathbf{S}_n$ from the corresponding stage of the auxiliary encoder are fed into the RAM, and the fusion result is returned to the main encoder. As a bridge between the main and auxiliary encoders, the RAM module establishes the connection through deformable convolution and the channel attention mechanism.

After the above four coding stages, we get feature $F \in \mathbb{R}^{(H/32)\times(W/32)\times 1024}$, which is fed into the decoder after a convolutional layer. Then, we input it into a 2×2 deconvolution layer to expand the resolution. Following UNet, ST-UNet utilizes skip-connection layers to concatenate the encoder and decoder features while reducing the number of channels through the 3×3 convolutional layer. Here, each convolutional layer is accompanied by a batch normalization layer and an ReLU layer. The above process is executed four times; feature F is gradually expanded to $F' \in \mathbb{R}^{(H/2)\times(W/2)\times 64}$. Finally, we apply a 3×3 convolutional layer and linear interpolation upsampling on feature F' to obtain the final prediction mask.

B. Swin Transformer Block

As previously mentioned, the standard transformer block consists of MSA, MLP and LN [see Fig. 2(a)]. Therefore, we can express the output $\mathbf{s}^l$ of layer l as follows:

$$\begin{aligned}\hat{\mathbf{s}}^l &= \mathrm{MSA}\big(\mathrm{LN}\big(\mathbf{s}^{l-1}\big)\big) + \mathbf{s}^{l-1},\\ \mathbf{s}^l &= \mathrm{MLP}\big(\mathrm{LN}\big(\hat{\mathbf{s}}^l\big)\big) + \hat{\mathbf{s}}^l.\end{aligned} \tag{1}$$

The standard transformer block used MSA to calculate the global self-attention among all tokens, which leads to the quadratic computational complexity of the number of tokens

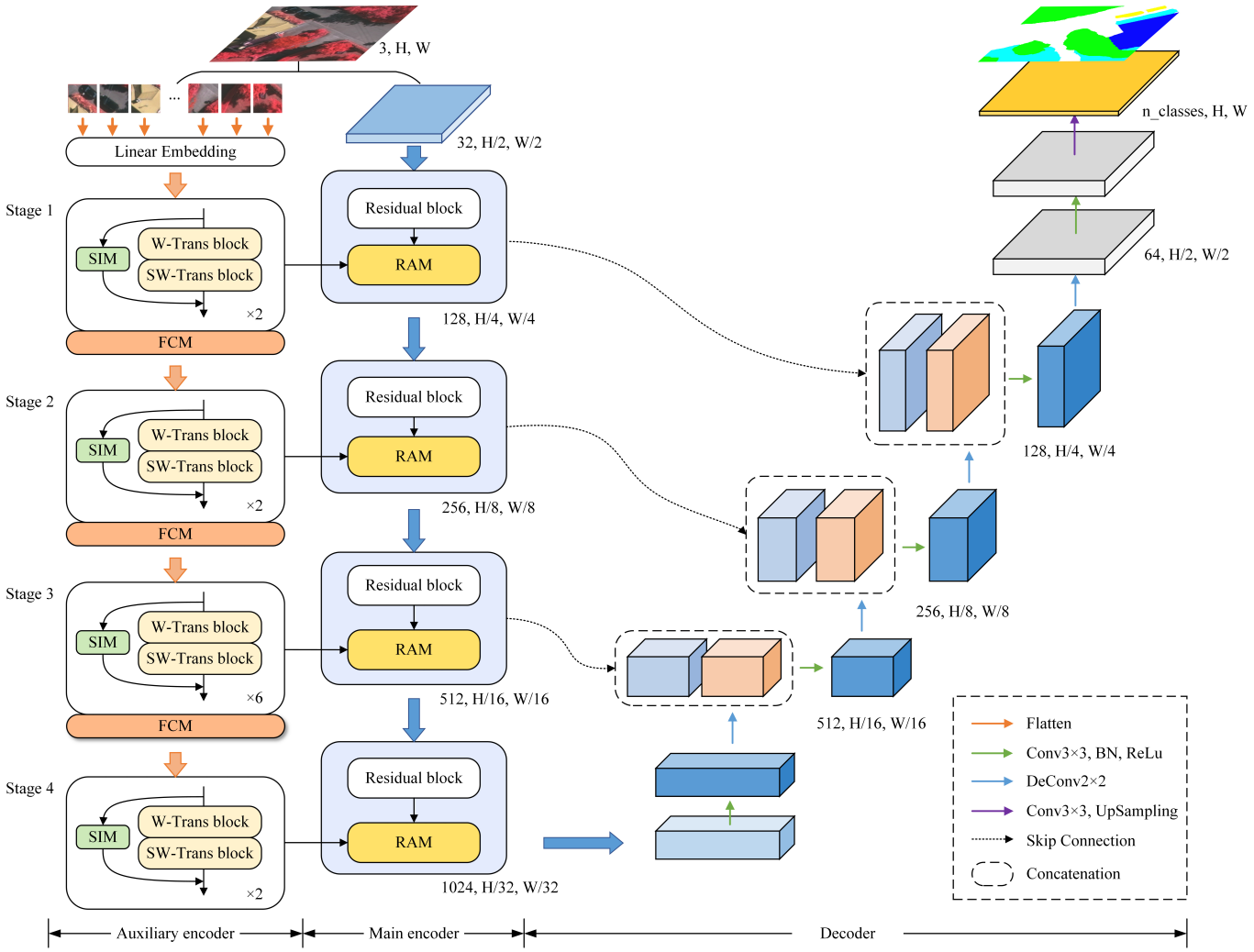

Fig. 3. Architecture of our proposed ST-UNet. ST-UNet contains three important modules: RAM, SIM, and FCM.

and limits its application scope, especially for some dense prediction tasks or based on high-resolution images (videos) tasks. For efficient modeling, the Swin transformer proposed the W-MSA with two partitioning configurations to replace ordinary MSA, namely, regular window configuration (W-MSA) and shifted window configuration (SW-MSA). They performed self-attention within windows but ignored the tokens outside the window, where each window only covers $D \times D$ patches. In the experiment, for the convenience of calculation, we set D to 8. As presented in Fig. 2(b), W-MSA and SW-MSA are alternately executed in consecutive Swin transformer blocks to enhance the information connection across windows. To facilitate the distinction, we rename the two Swin transformer blocks as W-Trans block and SW-Trans block. They are represented as follows:

$$\begin{aligned}
\hat{\mathbf{s}}^{l} &= \text{W-MSA}\left(\text{LN}\left(\mathbf{s}^{l-1}\right)\right) + \mathbf{s}^{l-1} \\
\mathbf{s}^{l} &= \text{MLP}\left(\text{LN}\left(\hat{\mathbf{s}}^{l}\right)\right) + \hat{\mathbf{s}}^{l} \\
\hat{\mathbf{s}}^{l+1} &= \text{SW-MSA}\left(\text{LN}\left(\mathbf{s}^{l}\right)\right) + \mathbf{s}^{l} \\
\mathbf{s}^{l+1} &= \text{MLP}\left(\text{LN}\left(\hat{\mathbf{s}}^{l+1}\right)\right) + \hat{\mathbf{s}}^{l+1}
\end{aligned} \tag{2}$$

where $\mathbf{s}^{l}$ represents the output feature of the W-Trans block and $\mathbf{s}^{l+1}$ represents the output feature of the SW-Trans block.

C. Spatial Interaction Module

The Swin transformer block establishes the relationship of patch tokens within a limited window, effectively reducing memory overhead. However, this approach weakens the global modeling capabilities of the transformer to a certain extent, even if it adopts the alternate execution strategy of the regular and shifted window. Besides, the occlusion of ground objects in RS images leads to blurred boundaries, which requires some spatial information to eliminate. Hence, we propose the SIM across the W-Trans and SW-Trans blocks to further enhance the information exchange while encoding more precise spatial information. SIM introduces attention in two spatial dimensions to consider the relationship between pixels, not just patch tokens, making the transformer more suitable for image segmentation tasks. The components of SIM are illustrated in Fig. 4.

Given stage n, we first reshape the input feature $\mathbf{s}^{l-1} \in \mathbb{R}^{(h\times w)\times c_1}$ of the W-Trans block into $\mathbf{z} \in \mathbb{R}^{h\times w\times c_1}$.

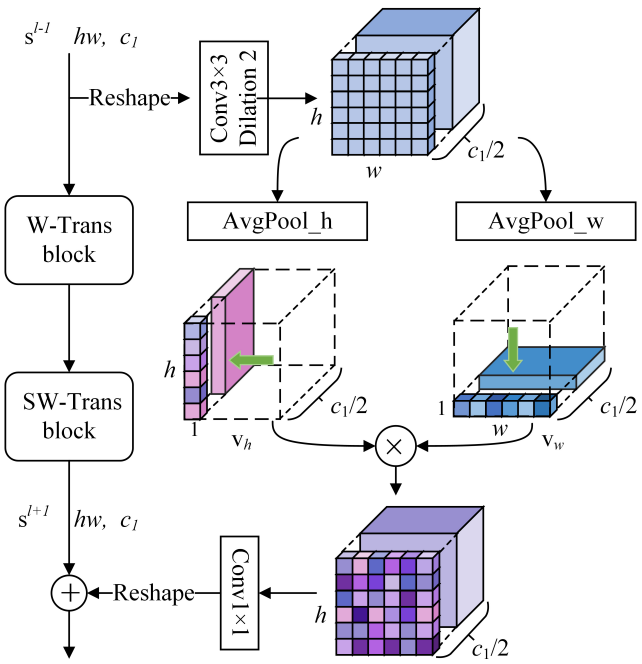

Fig. 4. Structure of SIM.

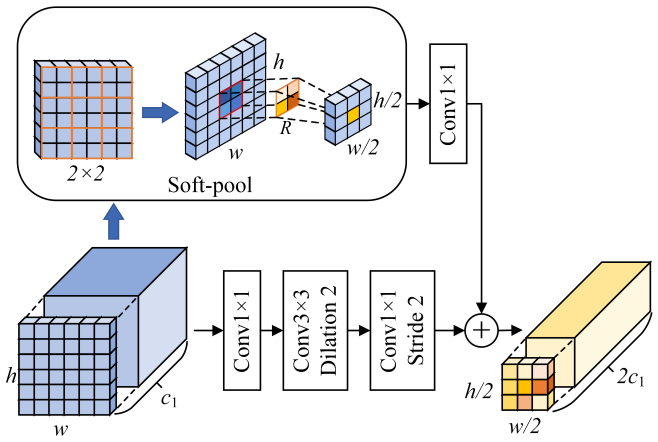

Fig. 5. Structure of FCM.

Here, $c_1 = 2^{n-1}C_1, h = (H/(2^{n+1}))$ and $w = (W/(2^{n+1}))$. The feature $\mathbf{z}$ is fed into a 3×3 dilated convolution layer with dilation rate 2 to reconstruct the structural information of the feature map through a large receptive field. In addition, to reduce the computational cost, the number of channels is shrunk to $c_1/2$. Then, the global average pooling operation is applied to obtain the statistics of the feature map in the spatial direction (vertical and horizontal). Specifically, the calculation formula of the elements in each direction is denoted as follows:

$$v_{h_i}^k = \frac{1}{w}\sum_{j=0}^{w-1}\hat{\mathbf{z}}^k(i,j)$$
$$v_{w_j}^k = \frac{1}{h}\sum_{i=0}^{h-1}\hat{\mathbf{z}}^k(i,j) \tag{3}$$

where i, j, and k are the indexes of the vertical direction, the horizontal direction, and the channel. Here, $0 \le i < h, 0 \le j < w, 0 \le k < c_1/2$. The feature $\hat{\mathbf{z}} = f(\mathbf{z})$, and $f(\cdot)$ is the dilated convolution layer with batch normalization and the GELU activation function. Therefore, we denote the aggregate tensor in the vertical and horizontal directions obtained by (3) as $\mathbf{v}_h \in \mathbb{R}^{h\times 1\times(c_1/2)}$ and $\mathbf{v}_w \in \mathbb{R}^{1\times w\times(c_1/2)}$, respectively. $\mathbf{v}_h$ and $\mathbf{v}_w$ converge the pixel-level weights of the feature map in spatial, so we multiply the two to obtain the attention map $\mathbf{M}$ related to the position, $\mathbf{M} \in \mathbb{R}^{h\times w\times(c_1/2)}$. Finally, the output feature map $\mathbf{F}$ of SIM is obtained by adding $\mathbf{M}$ and the output of SW-Trans block $\mathbf{s}^{l+1}$. It should be noted that the dimension of $\mathbf{M}$ needs to be increased through a convolutional layer to match the dimension of feature $\mathbf{s}^{l+1}$. The feature $\mathbf{F} \in \mathbb{R}^{h\times w\times c_1}$ can be expressed as follows:

$$\mathbf{F} = \mathbf{s}^{l+1} \oplus \varphi(\mathbf{v}_h \otimes \mathbf{v}_w) \tag{4}$$

where $\otimes$ stands for matrix multiplication and $\oplus$ stands for element-level addition. $\varphi(\cdot)$ means a 1×1 convolution layer with batch normalization and GELU.

D. Feature Compression Module

Previous work on transformer formed a hierarchical network by flattening and projecting image patches [27], [59] or merging the features of 2×2 adjacent patches and performing linear processing [30]. However, these methods easily lead to the loss of many details and structural information, which is not conducive to the semantic segmentation of RS images with dense and small-scale objects. Therefore, we design the FCM in the patch token downsampling of the Swin transformer to avoid the above problems, thereby improving the segmentation effect of small-scale objects.

Specifically, FCM has two branches, as illustrated in Fig. 5. One is a bottleneck block with the dilated convolution, which broadly gathers the features and structural information of small-scale objects by expanding the receptive field of the convolution. In the bottleneck block, the first 1×1 convolutional layer increases the dimension, the middle 3×3 dilated convolution layer is utilized to obtain extensive structural information, and the last 1×1 convolutional layer reduces the feature scale. Given the output $\mathbf{s}$ of stage n, the output of this branch is $\mathbf{F_1} \in \mathbb{R}^{(h/2)\times(w/2)\times 2c_1}$.

Another branch introduces soft-pool [61] operation to obtain finer downsampling. Soft-pool can activate the pixels in the pooling kernel in an exponentially weighted manner to preserve more detailed information. For each pixel in a specific kernel neighborhood $\mathbf{R}$, the calculation method of soft-pool is shown in the following equation:

$$\tilde{\mathbf{s}} = \sum_{i\in\mathbf{R}} \frac{e^{\mathbf{s}_i} * \mathbf{s}_i}{\sum_{j\in\mathbf{R}} e^{\mathbf{s}_j}}. \tag{5}$$

Then, the feature after soft-pool is input to a convolutional layer (increased dimension) to obtain the output $\mathbf{F_2} \in \mathbb{R}^{(h/2)\times(w/2)\times 2c_1}$. $\mathbf{F_2}$ can be expressed as follows:

$$\mathbf{F_2} = \varphi(\text{SoftPool}(\mathbf{s})). \tag{6}$$

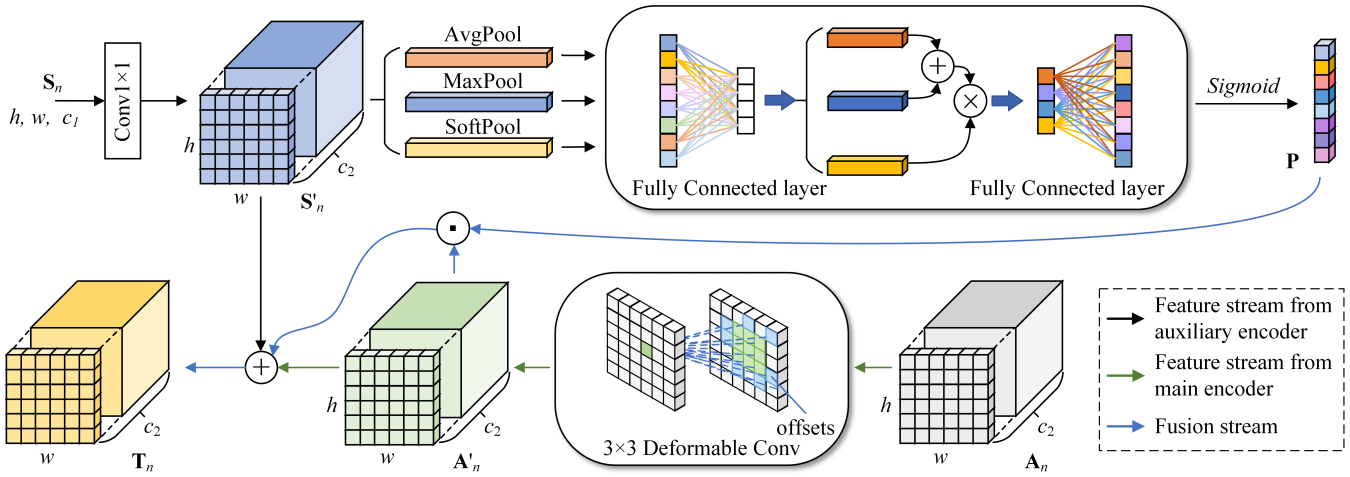

Fig. 6. Structure of the RAM.

In short, the function of one branch is to obtain small-scale features, and the function of the other is to preserve details, both of which are equally important and indispensable. Therefore, the two branches are merged in equal proportions as the output $\mathbf{L}$ of FCM. This process can be denoted as follows:

$$\mathbf{L} = \mathbf{F_1} \oplus \mathbf{F_2} \tag{7}$$

where $\oplus$ stands for element-level addition.

E. Relational Aggregation Module

The CNN-based main encoder extracts the local information restricted by the convolution kernel in the spatial dimension but lacks explicit modeling of the relationship between the channel dimensions [48], which can cause confusion when objects share similar distribution patterns but different channels [26]. Some methods (e.g., [21] and [62]) have proved that encoding the dependence of channel dimensions can improve feature discrimination. Therefore, we propose the RAM, whose detailed structure is presented in Fig. 6. To emphasize the important and more representative channels from the entire feature map, we extract the channel dependence from the global features of the auxiliary encoder and then embed it into the local features obtained from the main encoder. In addition, RAM introduces the deformable convolution [63] to adapt to differently shaped object regions and further refines the features of the main encoder. Through RAM, we can encode more global discriminative features to improve the segmentation accuracy of ground objects with high similarity in RS images.

As mentioned earlier, $\mathbf{A}_n$ and $\mathbf{S}_n$, respectively, represent the output of the main and auxiliary encoders at stage n. On the one hand, we input $\mathbf{A}_n$ into the deformable convolution to adapt to the geometric diversity of RS objects, which can be denoted as $\mathbf{A}'_n = \delta(\mathbf{A}_n)$. Here, δ is a 3×3 deformable convolution. On the other hand, $\mathbf{S}_n$ is sent to a convolutional layer to change the dimension and get the feature $\mathbf{S}'_n = \varphi(\mathbf{S}_n)$. Since each channel of the feature map can be regarded as a feature detector [64], the channel dependence focuses on the "meaningful content" in the image [49]. We adopt three pooling strategies to obtain more comprehensive channel dependence. First, we apply average- and max-pool layers to calculate the statistical characteristics of the feature map on the channel and send them to a shared fully connected layer. Then, $\mathbf{P}_{\mathrm{A\&M}} \in \mathbb{R}^{1\times1\times(c_1/2)}$ is obtained by adding the two. At the same time, a soft-pool with exponential weight is introduced to calculate the global weight descriptor, and it is also put into a fully connected layer, denoted by $\mathbf{P}_{\mathrm{S}}$. This process can be expressed by the following equation:

$$\begin{aligned}\mathbf{P}_{\mathrm{A\&M}} &= \sigma\left(\$_1\left(\mathrm{AvgPool}\left(\mathbf{S}'_n\right)\right)\right) + \sigma\left(\$_1\left(\mathrm{MaxPool}\left(\mathbf{S}'_n\right)\right)\right)\\ \mathbf{P}_{\mathrm{S}} &= \sigma\left(\$_1\left(\mathrm{SoftPool}\left(\mathbf{S}'_n\right)\right)\right)\end{aligned} \tag{8}$$

where σ stands for the ReLu function and $\$_1$ is set as a fully connected layer whose size is halved, taking into account the amount of calculation. Then, we optimize the descriptor of each channel by multiplying $\mathbf{P}_{\mathrm{A\&M}}$ and $\mathbf{P}_{\mathrm{S}}$, as shown in the following equation:

$$\mathbf{P} = \delta\left(\$_2(\mathbf{P}_{\mathrm{A\&M}} \odot \mathbf{P}_{\mathrm{S}})\right) \tag{9}$$

where δ stands for the sigmoid function, $\$_2$ is a fully connected layer with increased size, and $\odot$ represents element-level multiplication. We multiply the channel dependence $\mathbf{P}$ as the weight and the result $\mathbf{A}'_n$ of deformable convolution operation to obtain the refined features. Finally, the output feature $\mathbf{T}_n$ of the RAM is formed by connecting the refined features and the residual structure, which can be denoted as follows:

$$\mathbf{T}_n = \mathbf{A}'_n \oplus \mathbf{S}'_n \oplus \left(\mathbf{P} \odot \mathbf{A}'_n\right). \tag{10}$$

IV. Experiments

A. Datasets

1) Vaihingen Dataset: The Vaihingen dataset [65] contains 33 true orthophoto (TOP) images collected by advanced airborne sensors, covering 1.38 km^2 area of Vaihingen. The ground sampling distance (GSD) is about 9 cm. Each TOP image has infrared (IR), red (R), and green (G) channels.

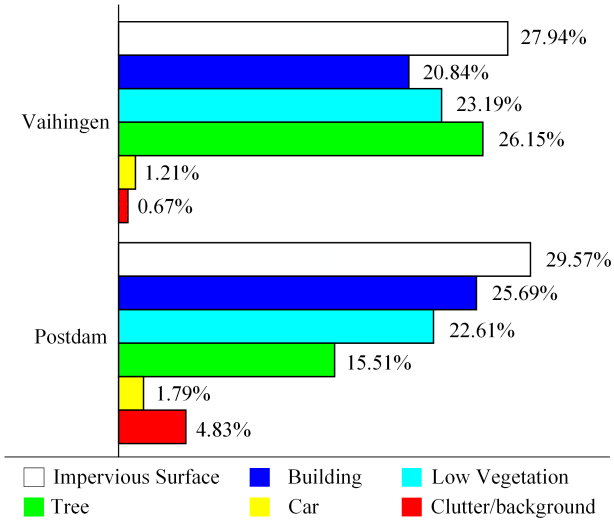

Fig. 7. Proportion of each semantic label in the Vaihingen and Potsdam datasets.

These images are labeled into sic categories for semantic segmentation. Following [38] and [66]–[68], we select 11 images for training (image IDs: 1, 3, 5, 7, 13, 17, 21, 23, 26, 32, and 37), fiveimages for testing (image IDs: 11, 15, 28, 30, and 34), and crop them to 256 × 256, respectively.

2) Potsdam Dataset: The Potsdam dataset [65] has 38 patches of the same size (6000 × 6000), which are all extracted from the very high-resolution TOP mosaic, and the GSD is 5 cm. The dataset covers 3.42 km^2 of Potsdam with complex buildings and dense settlement structures. The dataset is annotated with six categories for semantic segmentation research. Each image has three combinations of channels, namely, IR-R-G, R-G-B, and R-G-B-IR. In the experiment, we refer to the previous works [26], [38] to utilize 14 images with R-G-B for testing (image IDs: 2_13, 2_14, 3_13, 3_14, 4_13, 4_14, 4_15, 5_13, 5_14, 5_15, 6_14, 6_15, and 7_13), and the remaining 24 images with R-G-B for training. Similarly, we cut these original images to 256 × 256.

Fig. 7 shows the proportion of each semantic label in the above two datasets. Following [19] and [69], we ignore the category "Clutter/background" when performing quantitative evaluation on the two datasets.

B. Implementation Details

1) Training Settings: Our network is built with the Pytorch framework. We use the SGD optimizer with a momentum item of 0.9 and a weight decay of $1e-4$ to train the model. In addition, we set the initial learning rate to 0.01 and adopt the "Poly" decay strategy. All experiments are implemented on an NVIDIA Geforce RTX 2080 Ti 11-GB GPU. The batch size is set to 8, and the maximum epoch is 100.

2) Loss Function: As shown in Fig. 7, the category proportions are imbalanced in the Vaihingen and Potsdam datasets, which leads to the model training focusing on the categories with a larger proportion, while "disregarding" the categories with a smaller proportion [70]. To alleviate this problem, we adopt the joint loss with dice loss [71] L_{Dice} and the cross-entropy loss L_{CE} to supervise the model following [72] and [73]. The joint loss L is expressed as follows:

$$L = L_{\text{CE}} + L_{\text{Dice}}. \tag{11}$$

3) Evaluation Index: We employ the mean intersection over union (MIoU) and average F1 (Ave.F1) score to evaluate model performance. These two evaluation indicators are based on the confusion matrix, which contains four items: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). For each category, IOU is defined as the ratio of intersection and union of the predicted value and the true value, calculated as follows:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}. \tag{12}$$

The F1 score of each category is calculated as follows:

$$\text{F1} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{13}$$

where precision $= \text{TP}/(\text{TP}+\text{FP})$ and recall $= \text{TP}/(\text{TP}+\text{FN})$. Furthermore, MIoU represents the average of IoU from all categories, and the Ave.F1 score is the average of F1 from all categories.

TABLE I
ABLATION EXPERIMENT OF DUAL ENCODER STRUCTURE ON THE VAIHINGEN DATASET

Network Structure	Evaluation index	
	MIoU (%)	Average F1 (%)
Baseline UNet	66.35	79.13
Swin Transformer-UNet (Add_{LS})	67.05	79.82
Swin Transformer-UNet (Add_{ES})	67.32	79.94

C. Ablation Study

To evaluate the performance of the proposed network structure and three important modules, we apply UNet as the baseline network to perform ablation experiments on the Vaihingen dataset. In addition, we also study the effect of the loss function on the proposed network. In our ST-UNet, the main encoder employs half-compressed ResNet50, and the auxiliary encoder adopts Swin transformer with "Tiny" configuration (i.e., the hidden size $C_1 = 96$, the window size is 8, the number of layers corresponding to each stage is $\{2, 2, 6, 2\}$, and the number of heads corresponding to each layer is $\{3, 6, 12, 24\}$).

1) Effect of Dual Encoder Structure: The results are demonstrated in Table I. We can see that the introduction of the Swin transformer can effectively improve the segmentation performance of the baseline UNet. Specifically, we discuss the effect of the joint mode of the Swin transformer-based auxiliary encoder and the CNN-based main encoder on the segmentation model (for the convenience of discussion, we mainly focus on MIoU). One is Add_{LS}, that is, the features of the auxiliary encoder and the main encoder are merged only in the last stage of encoding. The other is Add_{ES}, which combines the

TABLE II
ABLATION EXPERIMENT OF THE PROPOSED MODULES ON THE VAIHINGEN DATASET

Model Name	Modules			IoU (%)					Evaluation index	
	RAM	SIM	FCM	Impervious Surface	Building	Low Vegetation	Tree	Car	MIoU (%)	Ave.F1 (%)
STransU				74.44	81.24	55.86	72.05	53.00	67.32	79.94
STransU+RAM	✓			75.07	81.64	57.98	71.42	53.50	67.92	80.42
STransU+SIM		✓		74.34	81.32	56.50	72.48	57.53	68.43	80.85
STransU+FCM			✓	75.20	82.03	56.14	72.07	57.71	68.63	80.97
STransU+RAM+SIM	✓	✓		76.03	83.11	57.21	72.68	59.94	69.79	81.81
STransU+SIM+FCM		✓	✓	75.87	82.50	57.74	72.69	58.46	69.45	81.57
STransU+RAM+FCM	✓		✓	76.02	82.98	56.90	72.03	61.18	69.82	81.85
STransU+RAM+FCM+SIM	✓	✓	✓	76.36	82.98	57.79	72.53	61.48	70.23	82.15

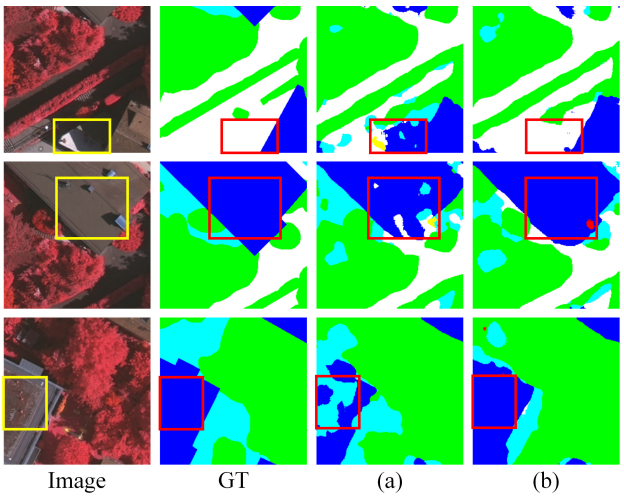

Fig. 8. Comparison of segmentation results before and after using RAM in the STransU framework. (a) STransU (Add_{ES}). (b) STransU (Add_{ES}) + RAM.

features of the auxiliary encoder and the main encoder at each stage of encoding by element-wise addition. Under the settings of Add_{LS} and Add_{ES}, the model increases the MIoU by 0.70% and 0.97%, respectively, compared to the baseline. It is proven that dual encoders can aggregate more information that is conducive to semantic prediction by cascading hierarchically. We refer to this network structure as STransU for short.

2) Effect of Relational Aggregation Module: Table II illustrates that the segmentation results increase by 0.60% on MIoU and 0.48% on Ave.F1 when the RAM is considered in the STransU framework. In particular, the segmentation accuracy of the category "Low Vegetation" has increased the most with an increase of 2.12% IoU, followed by the category "Impervious Surface" with an increase of 0.63%. More intuitively, the comparison of the visual segmentation results is shown in Fig. 8. In the first and second rows, the segmentation errors caused by light changes and the skylight on the roof are avoided after using RAM. In the third row, plants are growing on the top of the "Building," which is very similar to the "Low Vegetation," and they are still distinguished after using RAM. It is demonstrated that, after using RAM to embed more global context information, the segmentation accuracy of objects with high similarity is effectively improved.

3) Effect of Spatial Interaction Module: As shown in Table II, after using the SIM independently, the model achieves 1.11% and 0.91% improvements on MIoU and Ave.F1, respectively. The effectiveness of SIM in our network is verified. Since the categories "Car" and "Low Vegetation" are easily occluded by "Tree" in RS images, extracting and recognizing their semantic features are difficult. The model after using the SIM increases by 4.53% on the category "Car," 0.64% on the "Low Vegetation," and 0.43% on the "Tree." As shown in the first row of Fig. 9, a large part of the "Car" is blocked by the tall "Tree." The model fails to judge the boundary of cars and mistakenly recognizes the two cars as one car (the phenomenon of boundary blending occurs). In the third row, "Tree" and "Low Vegetation" have a similar appearance, and "Tree" is embedded in "Low Vegetation," which makes the model unable to determine the region of "Tree" exactly. It can be detected from visualization results in Fig. 9(b) that the introduction of SIM effectively diminishes the negative impact of mutual occlusion of objects.

4) Effect of Feature Compression Module: Table II reflects that after using FCM on the STransU framework, MIoU achieves growth of 1.31%, and Ave.F1 increases by 1.03%. From the numerical results in Table II, the model with FCM has the most obvious effect on "Car," which increases 4.71% IoU. Fig. 10 illustrates the visualized comparison results. The small-scale "Car" and "Building" in the first and second rows are extracted and segmented. Although the "Car" is surrounded by "Low Vegetation" in the third row, FCM can effectively segment its semantic region. This result indicates that FCM is beneficial to improve the segmentation accuracy of small-scale ground objects.

Moreover, the joint effect between modules is studied under the STransU framework, which is shown in Table II. Simultaneously introducing RAM and SIM, MIoU and Ave.F1 increase by 2.12% and 1.63%, respectively. When both RAM and FCM are included, the segmentation result is improved by 2.50% MIoU and 1.91% Ave.F1. When both SIM and FCM are considered, MIoU gets increments of 2.13%, and Ave.F1 obtains increments of 1.63%. Our ST-UNet with three

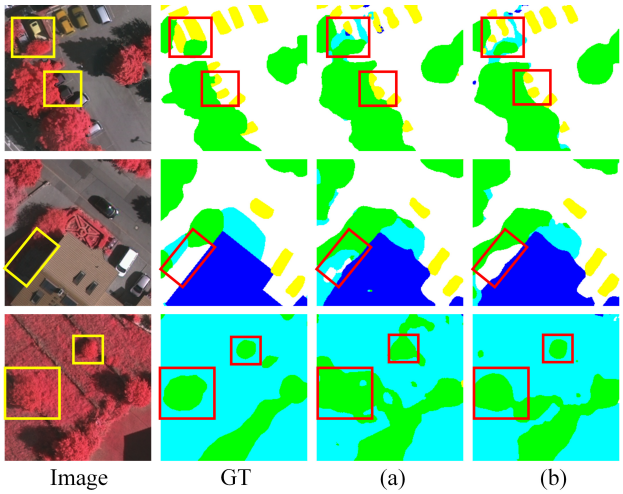

Fig. 9. Comparison of segmentation results before and after using SIM in the STransU framework. (a) STransU (Add_{ES}). (b) STransU (Add_{ES}) + SIM.

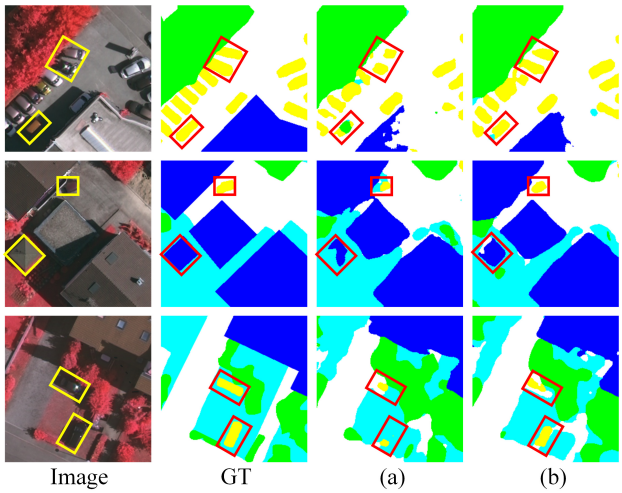

Fig. 10. Comparison of segmentation results before and after using FCM in the STransU framework. (a) STransU (Add_{ES}). (b) STransU (Add_{ES}) + FCM.

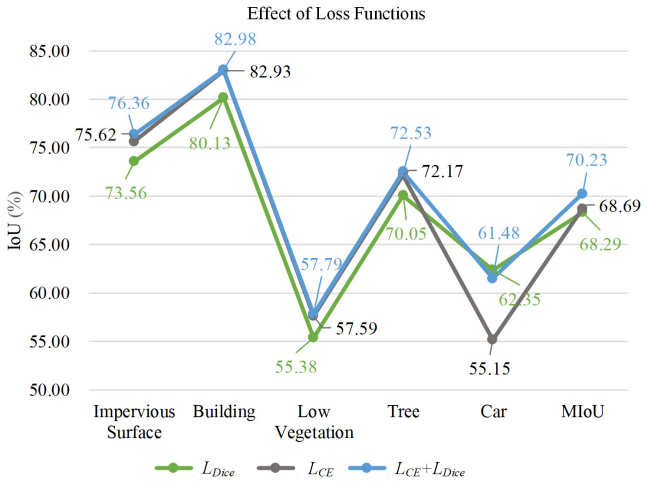

Fig. 11. Ablation experiment of loss functions on the Vaihingen dataset.

important modules (RAM, FCM, and SIM) brings an increase of 2.91% on MIoU and 2.21% on Ave.F1 compared with STransU.

5) Effect of Loss Functions: The ablation experiment about the effect of loss functions is implemented on the Vaihingen dataset. Fig. 11 intuitively shows the experimental results through lines and values. In contrast to only applying the cross-entropy loss L_{CE}, dice loss L_{Dice} significantly improves the segmentation result of "Car," which accounts for the smaller proportion. However, it weakens the supervision of categories other than "Car." When both L_{CE} and L_{Dice} are used (blue line versus gray line), the IoU of each category is improved.

D. Comparison With Other Methods

We compare the proposed ST-UNet with some existing methods, including FCN [11], UNet [14], Deeplab V3+ [15], unified perceptual parsing network (UperNet) [25], dual attention network (DANet) [21], TransUNet [20], and Swin-UNet [31]. The first five comparison methods are based on traditional CNN, and the latter two are based on transformer architecture. To be specific, TransUNet [20] is a hybrid of standard ViT and UNet, and Swin-UNet [31] is a UNet composed of pure Swin transformer blocks. It should be noted that TrandsUnet adopts the serial structure of transformer and CNN in the encoding stage, while our ST-UNet adopts the parallel mode.

The above comparison methods all use ResNet50 without pretraining as the backbone network. Due to memory limitations, ST-UNet utilizes half-compressed ResNet50 as the backbone of the main encoder, which also has not been pretrained.

1) Results on Vaihingen Dataset: Table III lists the numerical results of each semantic segmentation method. It is manifested that the proposed ST-UNet is superior to other methods in both MIoU and Ave.F1. Deeplab V3+ with dilated convolution and UperNet with pyramid structure obtain global context information by expanding the receptive field, while DANet uses the dual attention mechanism. Experimental data shows that UperNet and DANet with the dual attention mechanism are inferior to our ST-UNet in global context modeling. Deeplab V3+ uses the atrous spatial pyramid pooling module and the well-designed decoder structure to achieve better than other CNN-based models. UNet continuously integrates the spatial information from low-level features by skip connections, which produces slightly worse segmentation results than Deeplab V3+. Our ST-UNet improves MIoU by 3.53% and Ave.F1 by 2.78% compared with Deeplab V3+. In comparison with other methods, Swin-UNet utilizing a pure Swin transformer as the network architecture is not satisfactory. Although the Swin transformer blocks have better excellent global modeling capabilities, it is not enough for RS images to stack it similarly to Swin-UNet. TransUNet combines transformer and the standard convolutional layer sequentially, improving the segmentation accuracy by 9.10% MIoU compared to Swin-UNet. It demonstrated that the hybrid of CNN and transformer

TABLE III
COMPARISON OF SEGMENTATION RESULTS ON THE VAIHINGEN DATASET

Method	IoU (%)					Evaluation index	
	Impervious Surface	Building	Low Vegetation	Tree	Car	MIoU (%)	Average F1 (%)
FCN [11]	73.22	78.97	54.80	70.38	39.92	63.46	76.65
UNet [14]	72.91	81.68	57.23	71.63	48.29	66.35	79.13
Deeplab V3+ [15]	74.85	**83.01**	56.09	71.54	50.30	67.16	79.71
UperNet [25]	73.45	81.50	55.65	71.31	47.26	65.84	78.69
DANet [21]	73.54	81.40	56.88	71.21	42.68	65.14	78.00
TransUNet [20]	73.27	81.01	55.07	71.08	55.13	67.11	79.86
Swin-UNet [31]	69.31	73.37	49.48	67.12	30.78	58.01	72.02
ST-UNet	**76.36**	82.98	**57.79**	**72.53**	**61.48**	**70.23**	**82.15**

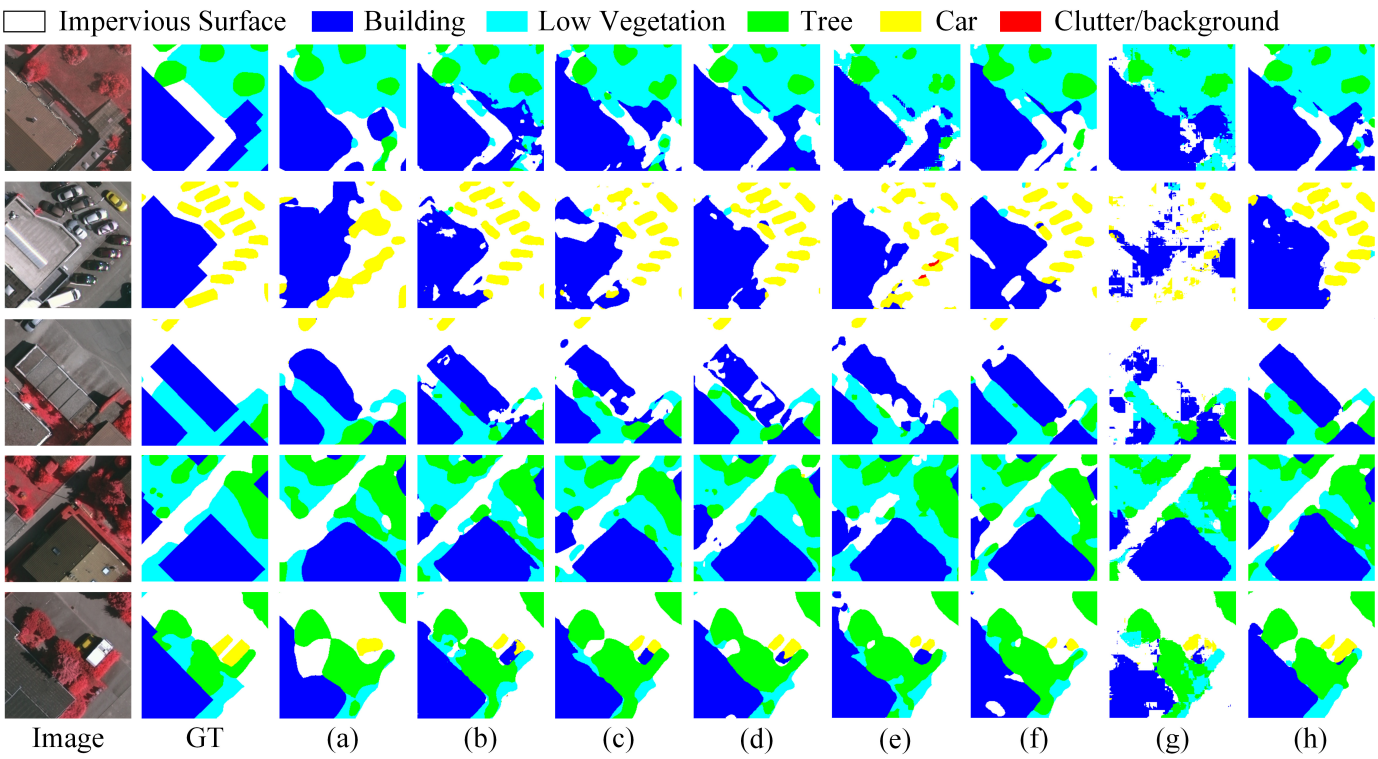

Fig. 12. Examples of semantic segmentation results on the Vaihingen dataset. (a) FCN. (b) UNet. (c) Deeplab V3+. (d) UperNet. (e) DANet. (f) TransUNet. (g) Swin-UNet. (h) ST-UNet.

TABLE IV
COMPARISON OF SEGMENTATION RESULTS ON THE POTSDAM DATASET

Method	IoU(%)					Evaluation index	
	Impervious Surface	Building	Low Vegetation	Tree	Car	MIoU (%)	Average F1 (%)
FCN [11]	77.41	83.52	66.10	63.19	74.34	72.91	84.12
UNet [14]	77.10	82.83	64.59	65.44	76.16	73.22	84.35
Deeplab V3+ [15]	79.01	84.76	67.53	63.05	78.05	74.48	85.13
UperNet [25]	76.95	83.93	65.65	60.40	76.57	72.70	83.91
DANet [21]	77.35	83.45	66.46	63.47	75.28	73.20	84.32
TransUNet [20]	78.61	85.60	67.16	64.10	79.33	74.96	85.44
Swin-UNet [31]	71.45	75.02	59.03	50.96	71.15	65.52	78.79
ST-UNet	**79.19**	**86.63**	**67.89**	**66.37**	**79.77**	**75.97**	**86.13**

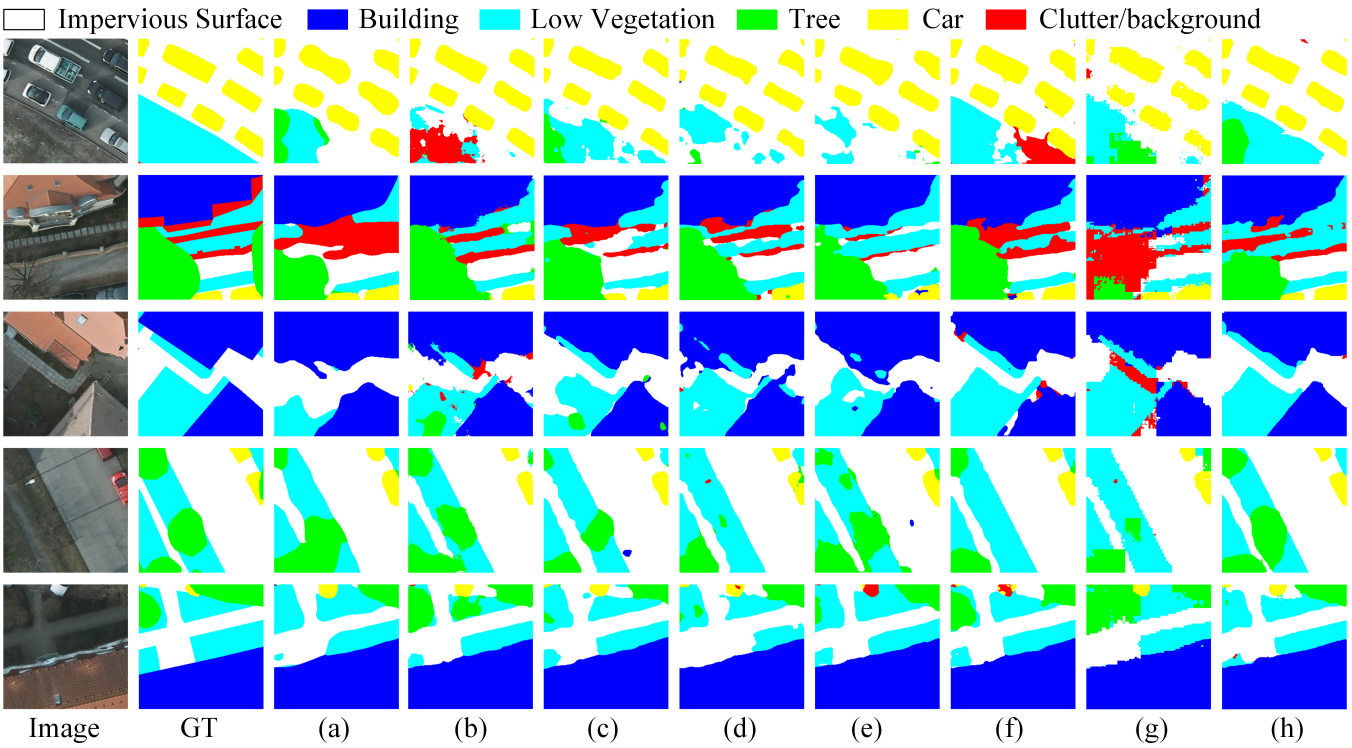

Fig. 13. Examples of semantic segmentation results on the Potsdam dataset. (a) FCN. (b) UNet. (c) Deeplab V3+. (d) UperNet. (e) DANet. (f) TransUNet. (g) Swin-UNet. (h) ST-UNet.

TABLE V
COMPARISON OF MODEL PARAMETERS, SPEED, AND ACCURACY

Method	Parameters	Vaihingen		Potsdam	
		Speed (FPS)	MIoU (%)	Speed (FPS)	MIoU (%)
FCN [11]	22.70 MB	380	63.46	381	72.91
UNet [14]	25.13 MB	219	66.35	229	73.22
Deeplab V3+ [15]	38.48 MB	70	67.16	71	74.48
UperNet [25]	102.13 MB	59	65.84	59	72.70
DANet [21]	45.36 MB	108	65.14	108	73.20
TransUNet [20]	100.44 MB	35	67.11	37	74.96
Swin-UNet [31]	25.89 MB	57	58.01	59	65.52
ST-UNet	160.97 MB	7	70.23	9	75.97

is feasible and can generate better results. Compared to TransUNet, our ST-UNet is more effective, increasing by 4.14% on MIou and 3.07% on Ave.F1, respectively.

Fig. 12 shows the prediction results of several semantic segmentation methods involved in Table III. It can be observed that Swin-UNet lacks spatial location information, resulting in many semantic fragments in its segmentation results. Compared with other models, ST-UNet reduces segmentation errors, especially for ground objects with high similarity. In the third row, other methods mistakenly identify " Building" as "Impervious Surface" due to their similar material, whereas our ST-UNet makes a relatively accurate judgment. In addition, the examples in the second row show that ST-UNet performs well as expected for dense and small-scale ground objects.

2) Results on Potsdam Dataset: Table IV reports the segmentation results of each method on the Potsdam dataset to prove the effectiveness of the proposed ST-UNet further. Our ST-UNet reaches 75.97% on MIoU and 86.13% on Ave.F1, which is superior to other methods. Due to the different sizes and data types, the segmentation accuracy on the Potsdam dataset is generally higher than that on the Vaihingen dataset. It should be noted that the segmentation accuracy of TransUNet with a hybrid structure surpasses the CNN-based models in Table IV. This verifies that the CNN-based models have certain limitations in describing global dependencies. However, Swin-UNet is still worse than the CNN-based model, which demonstrates that spatial information extraction is essential for large-scale RS images. Compared with Swin-UNet, ST-UNet improves MIoU by 10.45% and Ave.F1 by 7.34%. Compared with other comparison methods, our ST-UNet achieves an increase in the segmentation accuracy of each category.

We visualize the segmentation results in Fig. 13. Observing the first row, we can find that "Low Vegetation" and "Impervious Surface" have similar colors, and because of the existence of "Car," "Low Vegetation" is more likely to be considered as "Impervious Surface." With more discriminative features aggregated from the global context and local features, ST-UNet can still perform relatively accurate inference in this situation. In the third row, compared with other methods, ST-UNet accurately distinguishes "Impervious Surface" and "Low Vegetation" sandwiched between the crowded "Building." In the fourth row, our model better recognizes the "Tree" region in the "Low Vegetation" and the

elongated "Impervious Surface" sandwiched between the "Low Vegetation."

3) Efficiency Analysis: For the comprehensive comparisons, Table V lists the speed and parameters of all models in the same operating environment. In Table V, "Speed" indicates the number of images processed by the model per second, and its unit is frames per second (FPS) [69]. In terms of computational efficiency (speed), models with transformer or Swin transformer blocks are generally lower than other models with pure CNN structures. In particular, the speed of our ST-UNet on the Vaihingen and Potsdam datasets is only 7 and 9 FPS, respectively. Besides, since our ST-UNet is a parallel hybrid of the Swin transformer and UNet, its parameters are larger than other methods. Although the above two issues may limit the application of ST-UNet in some scenarios (such as small mobile devices), ST-UNet is still valuable in exploring the role of the Swin transformer in RS semantic segmentation.

V. Conclusion

In this article, we are committed to obtaining global contextual information in RS images to improve the feature discrimination of ground objects. We construct a semantic segmentation framework called ST-UNet with the dual encoder structure by combining Swin transformer and UNet. Specifically, the proposed relational aggregation module uses global features to guide the main encoder to obtain more discriminative features. Moreover, the SIM and the FCM are designed further to improve the global modeling capability of the Swin transformer. The SIM establishes pixel-level information exchange, eliminating the limitation of the window in the Swin transformer and alleviating the problem of semantic ambiguity caused by occlusion. The FCM retains as many detailed features as possible in the patch token downsampling for small-scale objects. However, our ST-UNet has shortcomings in the extraction of the boundary of the ground object, which is mainly manifested in that the segmentation result fails to fit the shape of the ground object completely, and the boundary lines are not smooth. We will further explore coding methods for boundary features o overcome this limitation. In addition, we will also put effort into model compression to improve inference efficiency.

References

[1] 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.

[2] X. Liu, L. Jiao, L. Li, X. Tang, and Y. Guo, "Deep multi-level fusion network for multi-source image pixel-wise classification," *Knowl.-Based Syst.*, vol. 221, Jun. 2021, Art. no. 106921.

[3] H. Luo, C. Chen, L. Fang, K. Khoshelham, and G. Shen, "MS-RRFSegNet: Multiscale regional relation feature segmentation network for semantic segmentation of urban scene point clouds," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8301–8315, Dec. 2020.

[4] J. Zhao, Y. Zhou, B. Shi, J. Yang, D. Zhang, and R. Yao, "Multi-stage fusion and multi-source attention network for multi-modal remote sensing image segmentation," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 6, pp. 1–20, Dec. 2021.

[5] L. Ding, J. Zhang, and L. Bruzzone, "Semantic segmentation of large-size VHR remote sensing images using a two-stage multiscale training architecture," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5367–5376, Aug. 2020.

[6] H. Bi, L. Xu, X. Cao, Y. Xue, and Z. Xu, "Polarimetric SAR image semantic segmentation with 3D discrete wavelet transform and Markov random field," *IEEE Trans. Image Process.*, vol. 29, pp. 6601–6614, 2020.

[7] L. Sahar, S. Muthukumar, and S. P. French, "Using aerial imagery and gis in automated building footprint extraction and shape recognition for earthquake risk assessment of urban inventories," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 9, pp. 3511–3520, Sep. 2010.

[8] G. Liu, L. Li, L. Jiao, Y. Dong, and X. Li, "Stacked Fisher autoencoder for SAR change detection," *Pattern Recognit.*, vol. 96, Dec. 2019, Art. no. 106971.

[9] Y. Yu *et al.*, "Crop row segmentation and detection in paddy fields based on treble-classification Otsu and double-dimensional clustering method," *Remote Sens.*, vol. 13, no. 5, p. 901, Feb. 2021.

[10] R. Sheikh, A. Milioto, P. Lottes, C. Stachniss, M. Bennewitz, and T. Schultz, "Gradient and log-based active learning for semantic segmentation of crop and weed for agricultural robots," in *Proc. Int. Conf. Robot. Automat.*, Paris, France, May/Aug. 2020, pp. 1350–1356.

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

[12] S. A. Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images: A review," *Artif. Intell. Rev.*, vol. 54, no. 1, pp. 137–178, 2020.

[13] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *CoRR*, vol. abs/2001.05566, pp. 1–22, Jan. 2020.

[14] 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. Intervent.*, vol. 9351, 2015, pp. 234–241.

[15] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," 2018, *arXiv:1802.02611*.

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

[17] R. Dong, X. Pan, and F. Li, "DenseU-net-based semantic segmentation of small objects in urban remote sensing imagess," *IEEE Access*, vol. 7, pp. 65347–65356, 2019.

[18] 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.

[19] 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.

[20] J. Chen *et al.*, "TransUNet: Transformers make strong encoders for medical image segmentation," *CoRR*, vol. abs/2102.04306, pp. 1–13, Feb. 2021.

[21] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 3146–3154.

[22] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. Int. Conf. Comput. Vis.*, Seoul, South Korea, Oct./Nov. 2019, pp. 603–612.

[23] B. Zhao, L. Hua, X. Li, X. Lu, and Z. Wang, "Weather recognition via classification labels and weather-cue maps," *Pattern Recognit.*, vol. 95, pp. 272–284, Oct. 2019.

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

[25] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. ECCV*, vol. 11209. Munich, Germany: Springer, Sep. 2018, pp. 432–448.

[26] 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, Dec. 2020.

[27] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Represent.* 2021, pp. 1–5.

[28] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. ECCV*, vol. 12346, Glasgow, U.K.: Springer, Aug. 2020, pp. 213–229.

[29] C. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," *CoRR*, vol. abs/2103.14899, pp. 1–12, Mar. 2021.

[30] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," *CoRR*, vol. abs/2103.14030, pp. 1–14, Mar. 2021.
[31] H. Cao *et al.*, "Swin-Unet: Unet-like pure transformer for medical image segmentation," *CoRR*, vol. abs/2105.05537, pp. 1–14, May 2021.
[32] A. Lin, B. Chen, J. Xu, Z. Zhang, and G. Lu, "DS-TransUNet: Dual swin transformer U-Net for medical image segmentation," *CoRR*, vol. abs/2106.06716, pp. 1–13, Jun. 2021.
[33] M. Zhang, X. Hu, L. Zhao, Y. Lv, M. Luo, and S. Pang, "Learning dual multi-scale manifold ranking for semantic segmentation of high-resolution images," *Remote Sens.*, vol. 9, no. 5, p. 500, May 2017.
[34] X.-Y. Tong *et al.*, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, Oct. 2020, Art. no. 111322.
[35] I. Nigam, C. Huang, and D. Ramanan, "Ensemble knowledge transfer for semantic segmentation," in *Proc. Winter Conf. Appl. Comput. Vis.*, Lake Tahoe, NV, USA, Mar. 2018, pp. 1499–1508.
[36] J. Zhang, S. Lin, L. Ding, and L. Bruzzone, "Multi-scale context aggregation for semantic segmentation of remote sensing images," *Remote. Sens.*, vol. 12, no. 4, p. 701, 2020.
[37] K. Sun *et al.*, "High-resolution representations for labeling pixels and regions," *CoRR*, vol. abs/1904.04514, pp. 1–14, Oct. 2019.
[38] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "High-resolution aerial image labeling with convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 7092–7103, Dec. 2017.
[39] J. Qi, C. Tao, H. Wang, Y. Tang, and Z. Cui, "Spatial information inference net: Road extraction using road-specific contextual information," in *Proc. Int. Geosci. Remote Sens. Symp.*, Yokohama, Japan, Jul. 2019, pp. 9478–9481.
[40] M. Kampffmeyer, A. 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. Conf. Comput. Vis. Pattern Recognit. Workshops*, Las Vegas, NV, USA, Jun. 2016, pp. 680–688.
[41] A. Ma, J. Wang, Y. Zhong, and Z. Zheng, "FactSeg: Foreground activation-driven small object semantic segmentation in large-scale remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5606216.
[42] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 135, pp. 158–172, Jan. 2018.
[43] S. Xie and Z. Tu, "Holistically-nested edge detection," *Int. J. Comput. Vis.*, vol. 125, nos. 1–3, pp. 3–18, Dec. 2017.
[44] S. Liu, W. Ding, C. Liu, Y. Liu, Y. Wang, and H. Li, "ERN: Edge loss reinforced semantic segmentation network for remote sensing images," *Remote Sens.*, vol. 10, no. 9, p. 1339, Aug. 2018.
[45] B. Zhao, X. Li, and X. Lu, "CAM-RNN: Co-attention model based RNN for video captioning," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5552–5565, Nov. 2019.
[46] X. Li, B. Zhao, and X. Lu, "MAM-RNN: Multi-level attention model based RNN for video captioning," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 2208–2214.
[47] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in *2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 10073–10082.
[48] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2017.
[49] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, vol. 11211. Cham, Switzerland: Springer, Oct. 2018, pp. 3–19.
[50] T. Panboonyuen, K. Jitkajornwanich, S. Lawawirojwong, P. Srestasathiern, and P. Vateekul, "Semantic segmentation on remotely sensed images using an enhanced global convolutional network with channel attention and domain specific transfer learning," *Remote Sens.*, vol. 11, no. 1, p. 83, 2019.
[51] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters—Improve semantic segmentation by global convolutional network," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 1743–1751.
[52] Y. Su, Y. Wu, M. Wang, F. Wang, and J. Cheng, "Semantic segmentation of high resolution remote sensing image based on batch-attention mechanism," in *Proc. Int. Geosci. Remote Sens. Symp.*, Yokohama, Japan, Jul. 2019, pp. 3856–3859.
[53] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 5998–6008.
[54] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Nov. 2020, pp. 38–45.
[55] H. Chen *et al.*, "Pre-trained image processing transformer," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jul. 2021, pp. 12299–12310.
[56] L. Yuan *et al.*, "Tokens-to-token ViT: Training vision transformers from scratch on imagenet," *CoRR*, vol. abs/2101.11986, pp. 1–10, Jan. 2021.
[57] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, Jul. 2021, pp. 10347–10357.
[58] S. Zheng *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," *CoRR*, vol. abs/2012.15840, pp. 1–12, Dec. 2020.
[59] W. Wang *et al.*, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," *CoRR*, vol. abs/2102.12122, pp. 1–15, Feb. 2021.
[60] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing transformers and CNNs for medical image segmentation," *CoRR*, vol. abs/2102.08005, pp. 1–11, Feb. 2021.
[61] A. Stergiou, R. Poppe, and G. Kalliatakis, "Refining activation downsampling with softpool," *CoRR*, vol. abs/2101.00440, pp. 1–21, Jan. 2021.
[62] Y. Huang, W. Jia, X. He, L. Liu, Y. Li, and D. Tao, "CAA: Channelized axial attention for semantic segmentation," *CoRR*, vol. abs/2101.07434, pp. 1–13, Jan. 2021.
[63] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets V2: More deformable, better results," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 9308–9316.
[64] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. ECCV*, vol. 8689. Zurich, Switzerland: Springer, Sep. 2014, pp. 818–833.
[65] *ISPRS 2D Semantic Labeling Dataset*. Accessed: Jun. 10, 2021. [Online]. Available: https://www2.isprs.org/commissions/comm2/wg4/benchmark/semantic-labeling/
[66] Y. Liu, D. M. Nguyen, N. Deligiannis, W. Ding, and A. Munteanu, "Hourglass-shapenetwork based semantic segmentation for high resolution aerial imagery," *Remote Sens.*, vol. 9, no. 6, p. 522, 2017.
[67] M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 881–893, Feb. 2017.
[68] D. Marcos, M. Volpi, B. Kellenberger, and D. Tuia, "Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 96–107, Nov. 2018.
[69] X. Li *et al.*, "PointFlow: Flowing semantics through points for aerial image segmentation," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 4217–4226.
[70] M. Kampffmeyer, A. Salberg, and R. Jenssen, "Urban land cover classification with missing data using deep convolutional neural networks," in *Proc. Int. Geosci. Remote Sens. Symp.*, Fort Worth, TX, USA, Jul. 2017, pp. 5161–5164.
[71] 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. (3DV)*, Stanford, CA, USA, Oct. 2016, pp. 565–571.
[72] Q. Zhu, Y. Zheng, Y. Jiang, and J. Yang, "Efficient multi-class semantic segmentation of high resolution aerial imagery with dilated linknet," in *Proc. Int. Geosci. Remote Sens. Symp.*, Yokohama, Japan, Jul. 2019, pp. 1065–1068.
[73] L. Fidon *et al.*, "Generalised wasserstein dice score for imbalanced multi-class segmentation using holistic convolutional networks," in *Proc. 3rd Int. Workshop*, vol. 10670, BrainLes, QC, Canada: Springer, 2017, pp. 64–76.

Xin He received the B.E. degree from the School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, China, in 2019, where she is currently pursuing the Ph.D. degree.

She is also a Member of the Engineering Research Center of Mine Digitization, Ministry of Education of the People's Republic of China, Xuzhou. Her research interests include computer vision and remote sensing image semantic segmentation.

Yong Zhou received the B.E. degree in industrial automation from Hohai University, Nanjing, China, in 1997, and the M.S. and Ph.D. degrees in control theory and control engineering from the China University of Mining and Technology, Xuzhou, China, in 2003 and 2006, respectively.

He is currently a Professor with the China University of Mining and Technology, and the Director of the Engineering Research Center of Mine Digitization, Ministry of Education of the People's Republic of China, Xuzhou. His research interests include artificial intelligence, deep learning, computer vision, and remote sensing image intelligent interpretation.

Jiaqi Zhao (Member, IEEE) received the B.E. degree in intelligence science and technology and the Ph.D. degree in circuits and systems from Xidian University, Xi'an, China, in 2010 and 2017, respectively.

From 2013 to 2014, he was an Exchange Ph.D. Student with the Leiden Institute for Advanced Computer Science (LIACS), Leiden University, Leiden, The Netherlands. He is currently an Associate Professor with the China University of Mining and Technology, Xuzhou, China, and a Member of the Engineering Research Center of Mine Digitization, Ministry of Education of the People's Republic of China, Xuzhou. His research interests include multiobjective learning and computer vision.

Di Zhang received the B.E. and M.S. degrees from the College of Computer Science and Engineering, Northwest Normal University, Lanzhou, China, in 2015 and 2018, respectively. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, China, and a Member of the Engineering Research Center of Mine Digitization, Ministry of Education of the People's Republic of China, Xuzhou.

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

Rui Yao (Member, IEEE) received the B.E. degree in computer science and technology from Henan Polytechnic University, Jiaozuo, China, in 2006, and the M.S. and Ph.D. degrees in computer science and technology from Northwestern Polytechnical University, Xi'an, China, in 2009 and 2013, respectively.

He is currently a Professor with the China University of Mining and Technology, Xuzhou, China, and a Member of the Engineering Research Center of Mine Digitization, Ministry of Education of the People's Republic of China, Xuzhou. His research interests include deep learning, computer vision, and pattern recognition.

Yong Xue (Senior Member, IEEE) received the B.Sc. degree in physics and the M.Sc. degree in remote sensing and geographic information systems (GISs) from Peking University, Beijing, China, in 1986 and 1989, respectively, and the Ph.D. degree in remote sensing from the University of Dundee, Dundee, U.K., in 1995.

He is currently a Professor with the China University of Mining and Technology (CUMT), Xuzhou, China, and a Professor of computation with the University of Derby, Derby, U.K. His research interests include geocomputation, aerosol optical depth retrieval from remotely sensed data, thermal inertia modeling, and heat exchange calculation for the boundary layer.

Prof. Xue is also an Editor of the *International Journal of Remote Sensing* and *International Journal of Digital Earth*, a Chartered Physicist, and a member of the Institute of Physics, U.K.