Swin-Conv-Dspp and Global Local Transformer for Remote Sensing Image Semantic Segmentation

Youda Mo, Huihui Li, Xiangling Xiao, Huimin Zhao, Xiaoyong Liu, and Jin Zhan

***Abstract*—Compared with the traditional method based on hand-crafted features, deep neural network has achieved a certain degree of success on remote sensing (RS) image semantic segmentation. However, there are still serious holes, rough edge segmentation, and false detection or even missed detection due to the light and its shadow in the segmentation. Aiming at the above problems, this article proposes a RS semantic segmentation model SCG-TransNet that is a hybrid model of Swin transformer and Deeplabv3+, which includes Swin-Conv-Dspp (SCD) and global local transformer block (GLTB). First, the SCD module which can efficiently extract feature information from objects at different scales is used to mitigate the hole phenomenon, reducing the loss of detailed information. Second, we construct a GLTB with spatial pyramid pooling shuffle module to extract critical detail information from the limited visible pixels of the occluded objects, which alleviates the problem of difficult object recognition due to occlusion effectively. Finally, the experimental results show that our SCG-TransNet achieves a mean intersection over union of 70.29% on the Vaihingen datasets, which is 3% higher than the baseline model. It also achieved good results on POSDAM datasets. These demonstrate the effectiveness, robustness, and superiority of our proposed method compared with existing state-of-the-art methods.**

***Index Terms*—Global local transformer block (GLTB), remote sensing (RS) image, semantic segmentation, Swin transformer, Swin-Conv-Dspp (SCD).**

I. Introduction

SEMANTIC segmentation provides pixel-level classification and is applied in many real applications. In the field of remote sensing (RS), semantic segmentation is also known as land use and land cover type classification [1]. In addition, RS technology can provide rich data sources for Earth observation. At present, RS images have been widely used in urban planning [2], [3], [4], housing planning [5], road detection [6], and forest protection [7], [8].

In recent years, with the rapid development of deep learning technology, segmentation models based on convolutional neural networks (CNN) and full convolutional networks (FCN) [9] have gradually become the most advanced image processing technology. In the process of this development, the encoder–decoder [10] structure showed extremely good segmentation performance, which also made it gradually become the basic architecture of many excellent models in the future. As a well-known encoder–decoder network model, UNet [11] fuses the feature information of deep granularity and shallow granularity through skip connections, which effectively alleviates the feature information lost by upsampling and downsampling. In addition, the well-known DeeplabV3+ [12] also follows the encoder–decoder structure, which is mainly improved on DeeplabV3 [13]. It extracts the information from different scales of the deep feature map in the encoder by using the hole spatial pyramid pooling, and fuses it with the shallow feature information in the decoder stage. Finally, it achieves very good performance.

However, RS images have complex imaging, redundant information, high similarity between classes, and are easily affected by the particularity of light intensity, light incident angle, and ground objects (small scale [14], high similarity [15], and mutual occlusion [16]). We have summarized two main issues, as shown in Fig. 1. From the examples, we can see that they are filled with a large number of objects occluded by shadows, easily leading to rough segmentation of the target edge and serious holes in the segmentation, which resulted it faces huge challenges in the application process. How to effectively utilize the occluded objects which have extremely small number of pixels has become the key to RS image segmentation. In traditional CNN, the encoder often uses multiple downsampling to reduce the amount of computation while increasing the receptive field. But multiple downsampling tends to lose a lot of valuable information, especially for occluded objects. If such a small number of precious pixels is lost, the effect of identifying occluded objects will become very bad. And CNN have inductive biases of locality and weight sharing [17], which lead to their inevitable constraints in learning long-range dependencies [18] and spatial correlations [19].

Manuscript received 27 November 2022; revised 17 February 2023, 23 February 2023, and 24 April 2023; accepted 23 May 2023. Date of publication 26 May 2023; date of current version 16 June 2023. This work was supported by the National Natural Science Foundation of China under Grant 62006049, Grant 62072122, and Grant 62172113, in part by The Ministry of Education of Humanities and Social Science Project under Grant 18JDGC012, in part by Guangdong Science and Technology Project under Grant KTP20210197 and Grant 2017A040403068, in part by Project of Education Department of Guangdong Province under Grant 2022KTSCX068 and Grant 2022ZDZX1013, and in part by Guangdong Science and Technology Innovation Strategy Special Fund Project (Climbing Plan) under Grant pdjh2022b0302 and Grant pdjh2022a0290. *(Corresponding authors: Huihui Li; Xiangling Xiao.)*

Youda Mo, Xiangling Xiao, Huimin Zhao, and Jin Zhan are with the School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou 510665, China (e-mail: 330462897@qq.com; 979432400@qq.com; zhaohuimin@gpnu.edu.cn; jinerzhan@163.com).

Huihui Li is with the School of Computer Science and Guangdong Provincial Key Laboratory of Intellectual Property and Big Data, Guangdong Polytechnic Normal University, Guangzhou 510665, China (e-mail: 29777562@qq.com).

Xiaoyong Liu is with the School of Data Science and Engineering, Guangdong Polytechnic Normal University, Guangzhou 510665, China, and also with the Institute of GPNU, Heyuan 517002, China (e-mail: 35643506@qq.com).

The code will be available at https://github.com/yuwxyun275/SCG-TransNet.

Digital Object Identifier 10.1109/JSTARS.2023.3280365

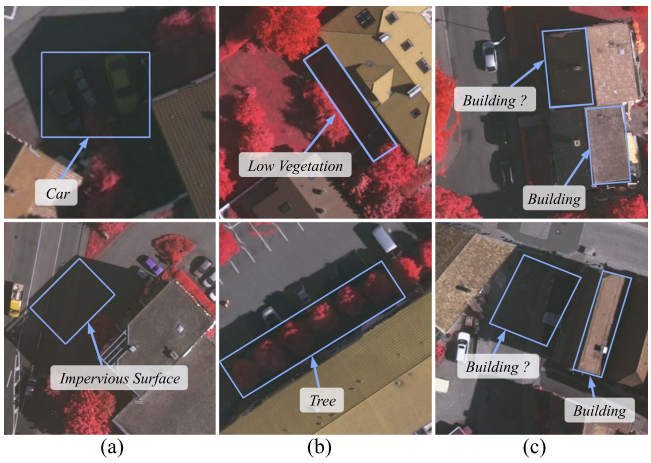

Fig. 1. Examples of the characteristics of RS images, which are taken from the Vaihingen dataset. 1) Affected by the shadow of the light. Like (a) and (b), it is hard to recognize the "Car," "Low vegetation," "Impervious surface," and "Tree"; 2) Interclass similarity and intraclass variability. Due to uneven lighting, the "Building" in (c) are difficult to identify.

Recently, vision transformer (VIT) [20] has been brilliant in the application of computer vision, and various transformer variants applied in the field of computer vision are emerging one after another. For example, the pyramid vision transformer (PVT) proposed by Wang et al. [21], as one of the representative models of transformer applied to the visual field, has achieved excellent results in many tasks. Although PVT reduces the consumption of computing resources to a certain extent, its computational complexity has a square relationship with the sequence length. In order to reduce the computational cost, Liu et al. [22] proposed a Swin transformer based on a shifted window strategy, which limits the computation of multihead self-attention to nonoverlapping windows while allowing cross-window information interaction. It broke through the problem of very high computational complexity of the transformer in vision tasks. Lin et al. [23] proposed cross attention in the transformer, a novel cross-attention mechanism to capture local as well as global information. Shao et al. [24] proposed a local transformer network embedded in a multiscale structure to explicitly learn the correspondence between multimodal inputs. The network can effectively and accurately capture the correspondence between long and short distances. Zhang et al. [25] proposed the Swin-Conv module, which combined the residual convolution and the capabilities of the Swin transformer then inserted it into the UNet [26] architecture. They also designed a practical noise degradation module, which was used in image denoising. Guo et al. [27] proposed a new visual network architecture, CNNs meet transformers. By simply combining traditional convolution and transformer, the network performance can achieve good performance, which is superior to Swin transformer and so on. Zhu et al. [28] proposed a unified framework to segment objects by considering contextual information and boundary artifacts. Azad et al. [29] proposed a new architecture based on a pure transformer named TransDeeplab, which combined transformer and deeplab architecture for the first time, and achieved the effect of state-of-the-art((SOTA) in medical image segmentation.

Nonetheless, the computational complexity of these VIT variants is still very high. The amount of parameters is very large, and the local and global context information is not sufficiently combined. This is not conducive to solving the problems of rough edge segmentation and serious holes in segmentation caused by shadow occlusion.

In order to solve the above problems, this article proposes a new RS image semantic segmentation network framework SCG-TransNet, which combines the network structure of the Swin transformer and Deeplabv3+. Deeplabv3+ is a network based on CNN that employs spatial pyramid pooling. The SCG-TransNet framework uses a Swin transformer as the encoder and decoder to extract features from high-resolution information. In the final stage of the encoder, Swin-Conv-Dspp (SCD) is used to capture multiscale feature information, and suppress the negative effects of high interclass similarity and intraclass difference caused by light factors, so as to alleviate the hole phenomenon in segmentation. In addition, a global local transformer block (GLTB) module is added before each visual upsampling to capture local feature information and global feature information to explore the spatial correlation between global and local features, improve target edge localization blur and alleviate segmentation blur caused by target occlusion. The main contributions of this article are as follows.

1) We propose an SCG-TransNet architecture that combines the Swin transformer with Deeplabv3+, which is applied to RS image segmentation for the first time.
2) We propose a SCD module to alleviate the hole phenomenon generated during segmentation. SCD can be helpful to extract feature information from objects at different scales and suppress the negative effects of noise such as chromatic aberration caused by light.
3) To extract discriminative information better, especially for small objects, we construct a GLTB with spatial pyramid pooling shuffle module (SPPS), which improves the accuracy of target edge localization.

II. Related Works

A. Semantic Segmentation of RS Images Based on CNN

FCN [9], a framework proposed by Jonathan et al. in 2015 for images semantic segmentation, has dominated the semantic segmentation tasks in RS in the subsequent years. However, the results obtained from fully convolutional neural networks are not fine-grained and sensitive to detail, and lack spatial consistency by lacking consideration of pixel-to-pixel relationships.

To better address these issues, an encoder–decoder network Deeplabv3+ [12] based on atrous spatial pyramid pooling (ASPP) is proposed. ASPP mines the contextual information of features of different resolutions through receptive field pooling of different sizes, while the encoder–decoder can better capture the edge information of different targets by gradually reconstructing the spatial information. Subsequently, the encoder–decoder architecture has been widely used in the framework of RS image semantic segmentation. Liu et al. [30] adopted a dual attention mechanism algorithm to improve the Deeplabv3+ network, which effectively enhanced the edge localization of

images and the accuracy of segmentation. Baheti et al. [31] adopted the idea of a two-stage attention mechanism and firstly proposed Attention Deeplabv3+ by assigning weights to each channel to capture the relationship between channels of a set of feature maps. Akcay et al. [32] developed an end-to-end two-stream architecture considering geospatial imagery based on the DeepLabv3+ architecture. Wang et al. [33] proposed a class feature attention mechanism fused with the improved Deeplabv3+ network CFAMNet for semantic segmentation of common features in RS images and achieved good segmentation results. Wang et al. [34] proposed a road segmentation method based on the receptive field and improved Deeplabv3+, innovatively used the initialization method to extract the layer backbone network in the network structure, and better extracted the characteristics of the road in the RS image. In addition, Li et al. [35] proposed a semisupervised semantic segmentation strategy for RS images, which improves the problems existing in the semisupervised semantic segmentation method of confrontation network by using a consistent self-training framework.

B. Semantic Segmentation of RS Images Based on Transformer

In recent years, VIT [20] has made great achievements in the field of RS image semantic segmentation. The traditional transformer structure is mainly used to process word vectors in the field of natural language [36], while the VIT is compatible with the transformer framework architecture into the field of computer vision. It can still achieve very good results on RS image segmentation tasks without relying on convolution. The convolution operation often causes the network to focus too much on the local features of the feature map, while the attention mechanism in the transformer can consider the global semantic feature information. Benefiting from the transformer's strong modeling ability for sequences, it has achieved advanced results in many basic vision tasks.

With such excellent results, many RS image researchers have also applied transformer to the semantic segmentation of high-resolution RS images. However, most of the existing transformer architecture networks for semantic segmentation still used the encoder–decoder architecture. For example, Li et al. [37] proposed a multistream RS spatiotemporal fusion network (MSNet) based on transformer and convolution, which achieved excellent segmentation accuracy on multiple RS datasets. Gao et al. [38] designed an adaptive fusion module, and proposed STransFuse by adopting a self-attention mechanism to adaptively fuse the semantic feature information of feature maps of different resolutions, which improved the segmentation quality of various RS images. Chen et al. [39] creatively proposed a new algorithm based on Swin transformer and linear spectral mixture theory for high-resolution RS images, and achieved state-of-the-art results in multiple public datasets. Li et al. [40] designed a modified transformer to capture global spatial location features across different scales, and demonstrated on object detection in optical remote sensing images (DIOR) [41] and northwestern polytechnical university very high resolution -10 (NWPU VHR-10) [42] high-resolution RS image datasets' excellent segmentation accuracy. Wang et al. [43] proposed a Swin transformer-based densely connected feature aggregation module by recovering resolution and generating segmentation maps by designing shared spatial attention and shared channel attention. It enhanced the relationship between semantic features in space and channels, and effectively alleviated the problems of multiscale and confusing geospatial targets that often appear in high-resolution RS images. Kaselimi et al. [44] proposed a multilabel visual transformer model ForestVIT, which applied transformer with a self-attention mechanism to the detection of deforestation. Zhang et al. [45] proposed a hybrid deep neural network based on transformer and CNN for semantic segmentation of ultrahigh-resolution RS images. Sun et al. [46] proposed a spectral spatial feature tokenized transformer based on the transformer framework, which can effectively capture spectral spatial features and advanced semantic features so that the model can better extract deep semantic features.

C. Attention Mechanism

In order to improve the defect that the CNN network focuses too much on local features due to convolution and cannot capture global information well, many scholars have begun to integrate attention into the network. Li et al. [47] adaptively refine features by integrating lightweight spatial and channel attention modules. Chen et al. [48] proposed a feature map attention mechanism for image super-resolution reconstruction. By using features of different resolutions to adaptively adjust the channel features, we recover more details and relieve the network from focusing too much on local areas.

Li et al. [49] proposed a high-resolution RS image change detection model with a multiscale attention mechanism. By applying the attention mechanism to feature maps of different resolution scales, feature representations of various scales are generated and then improved of the defect of over-focusing on local context. Liu et al. [50] proposed a self-attention negative feedback network applied to real-time image segmentation, which reconstructed more realistic and clearer real-time images. Hu et al. [51] proposed a dual-region learning network applied to high-resolution image reconstruction to extract continuous and fine pixel-level features through the spatial spectrum module with efficient feature fusion. Xia et al. [52] proposed a new deformable self-attention module, which can select the positions of key and value pairs in self-attention according to different dependencies of the data. This self-attention mechanism can focus on the associated regions and capture more informative features. Zhang et al. [53] proposed a lightweight multiscale attention block to build attention between feature maps of different resolutions, achieving better results. Sun et al. [54] proposed a successive pooling attention network including a successive pooling attention module and a feature fusion module, which effectively alleviates the difficulty of accurately segmenting small-scale objects and object boundaries in RS images.

Nonetheless, the computational complexity of the proposed state-of-the-art transformer-based encoders tends to be very large, and the extraction of global contextual information is still insufficient. This will still lead to missed detection due to the hole

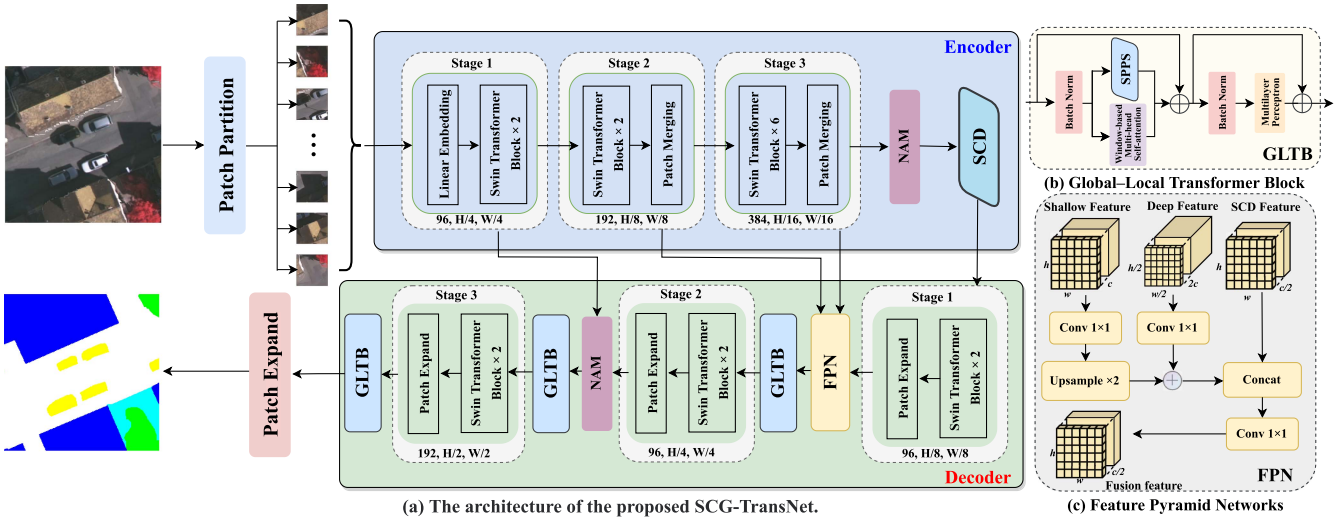

Fig. 2. (a) Architecture of our proposed SCG-TransNet. SCG-TransNet contains two important modules: SCD and GLTB with SPPS; (b) Components of the GLTB; (c) Components of the FPN.

phenomenon. Therefore, in order to fully extract the global context information, we propose a hybrid of Swin transformer and Deeplabv3+ as the encoder–decoder for efficient segmentation of RS images. Specifically, for the proposed SCG-TransNet, we avoid the problem of the loss of detail information and discontinuous pixel segmentation due to high-fold direct upsampling by using feature pyramid networks (FPN). In the final stage of the encoder, we use SCD to efficiently extract feature information of objects of different scales while suppressing noise. In the decoder, we introduce the GLTB with SPPS before each visual up-sampling, and finally achieve a high-precision segmentation effect.

III. Method

In this section, we detail the overall structure of the proposed SCG-TransNet and introduce the involved Swin transformer. Subsequently, two important modules in SCG-TransNet, namely SCD and GLTB with SPPS, are introduced.

A. Overview

The overall architecture of the proposed SCG-TransNet is shown in Fig. 2. As a hybrid of Swin transformer and Deeplabv3+, our SCG-TransNet follows the encoder–decoder paradigm. In the encoder stage, we adopt the Swin transformer as the backbone network for feature extraction, and introduce the SCD module in the final stage of the encoder. In the decoder, FPN is used to fuse the features of different resolutions generated by Stage 2 and Stage 3, followed by stacking on the channel with the feature map twice upsampled after SCD, which enhances the communication of multiscale features and solves the problem of loss of important pixel information caused by direct high-multiple upsampling, and effectively enhances the continuity of pixel information. In addition, a normalization-based attention module (NAM) attention mechanism [55] is added before SCD and before concat of shallow and deep features to redistribute the weights of multiscale feature maps for better feature extraction. Finally, a GLTB module is added before each visual upsampling.

B. Swin Transformer Based Encoder and Decoder

The encoder is mainly composed of the Swin transformer backbone network and SCD. Swin Transformer is used to extract hierarchical feature maps, and SCD is used to capture multiscale contextual information. The decoder is mainly composed of the Swin transformer block, FPN, and GLTB. The FPN is used to fuse feature maps of different depths. The GLTB is used to capture global and local semantic information of feature maps. This process can be expressed as

$$e_i = Encoder_{Swim-Trans}(images) \tag{1}$$

$$d_i = Decoder_{Swim-Trans}\left(e_i\right). \tag{2}$$

The Swin transformer block is the core of the Swin transformer backbone network. The computational complexity of the traditional VIT on the global receptive field is quadratic. In order to reduce the computational complexity, Liu et al. designed the Swin transformer. Between successive self-attention layers, a multihead self-attention (MSA) module in transformer is replaced by a shift-window-based module. By sequentially concatenating the window-based multihead self-attention (W-MSA) block with a shifted window-based multihead self-attention (SW-MSA) block, the context information of the global space is obtained in a more efficient manner. For the specific calculation process refer to [22].

Under this shifted window partitioning scheme, a W-MSA module and a SW-MSA module are applied in series to the transformer block. The first Swin transformer block is a W-MSA block. The input feature $\mathbf{x}^{l-1}$ passes through the LayerNorm and W-MSA layers and establishes a residual connection to obtain $\hat{\mathbf{x}}^{l}$. After that, $\hat{\mathbf{x}}^{l}$ passes through the LayerNorm and multi-layer perceptron (MLP) layers and establishes a residual connection again to obtain $\mathbf{x}^{l}$. The SW-MSA block has only half the window

size offset in the calculation of the W-MSA layer, and the other structures are almost the same as the W-MSA block. This process can be expressed as

$$\hat{\mathbf{x}}^{l} = W_{\mathrm{MSA}}\left(\mathrm{LN}\left(\mathbf{x}^{l-1}\right)\right) + \mathbf{x}^{l-1}. \tag{3}$$

$$\mathbf{x}^{l} = \mathrm{MLP}\left(\mathrm{LN}(\hat{\mathbf{x}}^{l})\right) + \mathbf{x}^{l}. \tag{4}$$

$$\hat{\mathbf{x}}^{l+1} = \mathbf{SW}_{\mathrm{MSA}}\left(\mathbf{LN}\left(\hat{\mathbf{x}}^{l}\right)\right) + \mathbf{x}^{l}. \tag{5}$$

$$\mathbf{x}^{l+1} = \mathbf{MLP}\left(\mathbf{LN}\left(\hat{\mathbf{x}}^{l+1}\right)\right) + \mathbf{x}^{l+1}. \tag{6}$$

Specific details of the calculations can be found in [22].

Compared with the backbone network based on CNN, the Swin transformer is a sequence-to-sequence model, which makes it easier to combine multimodal data. Its long-range modeling capability from the attention mechanism releases the limitations of traditional CNN-based models. The Swin transformer does not contain inductive biases, so it does a good job of capturing long-range spatial dependencies in images. Second, compared with other transformer-based backbone networks, the computational complexity of the Swin transformer is lower, and the speed of recognition and reasoning will be faster.

C. Swin-Conv-Dspp

Since atrous convolution [56] easily leads to the loss of continuous information in space, it is not conducive to capture object features of different scales. To solve this problem, ASPP in Deeplabv3+ uses multiple parallel atrous convolutional layers with different sampling rates to obtain information of different scales of objects. And in the case of reducing the loss of information as much as possible, the construction of the feature extraction network is strengthened by increasing the receptive field. RS images often contain a lot of noise [57], such as light intensity and light incident angle. How to effectively suppress the negative effects of these noises has become the key to semantic segmentation of RS urban scenes. Atrous convolution is extracted across pixels in feature point extraction, which is a sparse sampling method. This will inevitably lead to the loss of pixel information, resulting in a lack of correlation between the results obtained by long-distance convolution, which is not conducive to suppressing noise. This will make it difficult to identify targets with too high interclass similarity or too large intraclass differences due to light incident angle and light intensity, and eventually lead to the appearance of holes.

Therefore, we combined the characteristics of CNN and the Swin transformer to design a dual-space pyramid pooling layer, using Swin transformer's strong information extraction ability in the global context to make up for the key details lost by using atrous convolution information, and strengthen the ability to extract global context feature information to alleviate the difficulty of ASPP to capture the long-range dependence of semantic information. The proposed SCD is shown in Fig. 3. Atrous convolution is essentially a superposition of many high-pass filters [58], which continuously enhances high-frequency information, so it tends to be better at extracting high-frequency information of features. The transformer is essentially a low-pass filter [59], which continuously strengthens the underlying

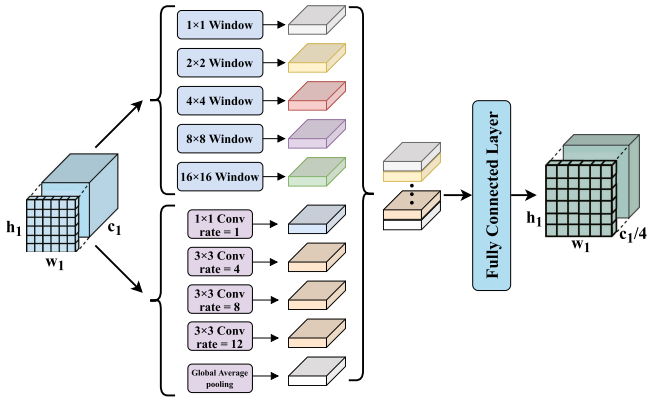

Fig. 3. Structure of the SCD module, of which the top branch is the Swin transformer branch and the bottom branch is the CNN branch.

semantic information of the image, so it is often better at extracting low-frequency information of features. By combining the advantages of the two to reduce the differences within classes and expand the frequency of information between classes, the negative effects of various noises in RS images are effectively suppressed. It improves the phenomenon that it is difficult to distinguish due to excessive intraclass differences or high interclass similarity, and alleviates the problem of hole phenomenon.

Specifically, SCD has Swin transformer branch and convolution branch. As shown in Fig. 3, shiftable windows of different sizes are used to better extract semantic information between patches with different distances to capture multiscale information. Smaller window scales aim to capture local information, while larger windows aim to capture global contextual information. The convolution branch utilizes 1, 4, 8, and 12 atrous convolutions with different dilation rates, and it broadly extracts objects of different scales by expanding the receptive field of the convolution. Try to expand the receptive field to extract feature information of different scales without reducing the loss of information. By combining the strong local feature extraction ability of convolution and the excellent capture ability of the transformer in global context and long-range dependencies, SCD shows excellent antinoise performance, effectively alleviating the problem of holes caused by excessive similarity between classes.

D. Global-Local Transformer Block

The proposed GLTB is mainly composed of the following two branches: 1) the global context branch and 2) the local context branch.

In RS images, the distribution of cars is often clustered, such as parking lots, temporary parking spaces on roads, and the parking locations and distributions are often regular. If the global context information can be effectively extracted, learnt, and captured car parking the rules of the model, the accuracy of the car class prediction will be greatly improved. The same is true for the building class, the distribution structure of the house is strongly related to the location. Houses are always arranged in a determinant layout, with buildings arranged in parallel to form a regular determinant. And the shape of the house is often

rectangular, square, and rarely other shapes. Although global context information is very important in semantic segmentation of complex urban scenes, local information also plays a pivotal role in the rich spatial details of images. Only using traditional convolution to extract semantic features is not ideal because it is too limited to local information. Local information is very important for RS segmentation, and only using the global context capture ability of transformer will result in the inability to effectively extract local information features of high-resolution RS images. The advantages of both can be combined to effectively extract local context information and global context information.

In the global branch, which is mainly captured by window-based multihead autonomous attention, we first use standard 1×1 convolution to expand the channel dimension of the input 2-D feature map $\in R^{B\times C\times H\times W}$ by a factor of 3. Next, the 1-D sequence $\in R\left(3 \times B \times \frac{H}{W} \times \frac{W}{W} \times \frac{h}{h}\right) \times (w \times w) \times \frac{C}{h}$ is converted into Q, K, and V vectors using the window division operation. For details, the channel dimension is set to 64, and the window size and attention head are both set to 8. Details of window-based multihead self-attention can be found in [60]. Although self-attention based on shiftable windows can capture feature information across windows, the amount of computation is greatly increased. Therefore, we introduce the context interaction module of the cross-shaped window to fuse the two feature maps generated by the horizontal average pooling layer and the vertical average pooling layer, so as to capture the global context efficiently. Details of the computation of GLTB in the global branch can be found in [61].

In addition, in RS images, a certain category is often obscured by shadows or other objects. For example, houses on both sides of the road can easily occlude parked or moving cars on the road. In this way, it is easy to cause problems such as blurred boundaries, false detections or even missed detections during segmentation. In order to solve the problem that the target to be recognized is occluded, Li et al. [62] proposed spatial pyramid convolutional shuffle in you only look once (YOLO), hoping to solve the recognition of the occluded human body. However, it does not take into account the loss of detailed information caused by the use of parallel convolution, and is still limited to the capture of local information by convolution, lacking the extraction of global context information, which obviously cannot be directly applied to the field of RS. Inspired by it, in the local branch, we propose an SPPS whose structure is shown in Fig. 4. Specifically, it covers different ranges by using four parallel convolution kernels of different sizes and dilation rate convolutions, and then stacks the results obtained by each branch convolution on the channel. Since the limited visible pixels of the occluded object are very rare, it is inevitable that some pixel information will be lost after passing through the convolution kernels of different sizes. For the occluded object, these rare pixel information often becomes the final occlusion whether objects can be correctly identified. Therefore, we use a global average pooling layer and skip connections to compensate for the loss of key details in utilizing different kernel sizes. Then, we use PixelShuffle to combine the adjacent elements. PixelShuffle will combine the information of the same position extracted from convolution kernals of different sizes such as blue, green, yellow, and purple in the figure in an adjacent manner. In the feature graph output by SPPS, the information of adjacent combination of features extracted from the same position in the original feature graph by convolution kernel of different sizes and expansion rate is called a cell. Each cell contains information extracted by different kernel sizes at the same position in the original feature map. These information can provide multilevel information extracted from the same position of the original feature map, and realize the extraction of multireceptive field information from the same position of the feature map. SPPS can improve the ability to extract key details and generate distinguishable features from the limited visible pixels of occluded objects. What is more, it further enhances the extraction ability of local information, and refines the global and local feature information of the feature map when upsampling restores the feature map. At the same time, the module can be plug-and-play and can be efficiently migrated to other models for use.

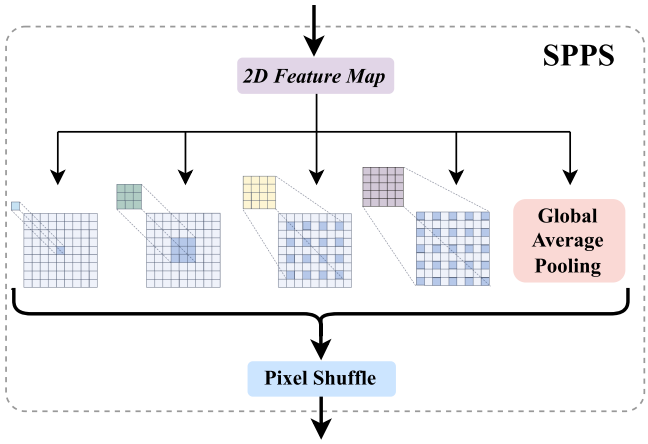

Fig. 4. Structure of the SPPS module.

Algorithm 1: Training Process of SCG-TransNet.

Input: Vaihingen or Potsdam dataset D;

```
for epoch < epochs do
    Extract features by (1) with SCD module;
    Fusing features by (2) with GLTB module;
    Get segmentation maps;
    Update the parameters of the module;
end for
```

Output: Trained SCG-TransNet;

Furthermore, we provide Algorithm 1 to describe our proposed SCD-TransNet in detail.

IV. Experiments

A. Datasets

1) Vaihingen Dataset: The Vaihingen dataset [63] contains 33 remotely sensed images of different sizes, which extracted from a very large top-level orthophoto image, covering more than 1.38 km^2 of the city of Vaihingen. The RS image format

is an 8-b tag image file format (TIFF) file consisting of the following three bands: 1) near infrared; 2) red and 3) green. The digital surface model (DSM) is a single-band TIFF file with the gray level (corresponding to the DSM height) encoded as a 32-b floating point value. In our experiments, we cropped them each to a size of 256 × 256, and the details of the experiments are given in [64].

2) POTSDAM Dataset: The POTSDAM dataset [63] has 38 remotely sensed images all of 6000 × 6000 resolution in size. The dataset covers 3.42 km^2 of complex buildings and dense settlement structures. The dataset has six categories for semantic segmentation. Again, we cropped each of them to a size of 256 × 256.

We ignore the category of "background" in the quantitative evaluation of the two datasets.

B. Implementation Details

1) Training Setup: Our network model is built on Pytorch's deep learning framework. For fast convergence, we use Adam as the optimizer and set the propulsion to 0.9 to train the model. The initial learning rate was set to 0.001 and the learning rate was adjusted using a step strategy. All experiments were deployed on NVIDIA GTX 2060 and NVIDIA GTX 3090. The batch size was set to 10 and the maximum epoch was 150.

2) Loss Function: Due to the category imbalance in the Vaihingen and POTSDAM datasets, the model training focused on the larger categories and ignored the smaller categories. To improve this problem, we used the joint loss of dice Loss [65] and cross entropy (CE) Loss, with the joint loss L denoted as

$$L = L_{\text{CE}} + L_{\text{Dice}}. \tag{7}$$

3) Evaluation Metrics: We use the mean cross-merge ratio (MIoU) and the mean F1 (Ave F1) score to evaluate model performance. These two evaluation metrics are based on confusion matrices and contain the following four terms:

1) true positive (TP);
2) false positive (FP);
3) true negative (TN);
4) false negative (FN).

In addition, we added giga floating-point operations per second (GFLOPs) and overall accuracy (OA) in the ablation experiment to evaluate the computational complexity and overall accuracy of the model, respectively. For each category, the intersection over union (IOU) is defined as the intersection of the predicted and true values and is calculated as follows:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}. \tag{8}$$

The F1 score for each category is calculated as follows:

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{9}$$

The overall accuracy rate OA is calculated as

$$\text{OA} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{10}$$

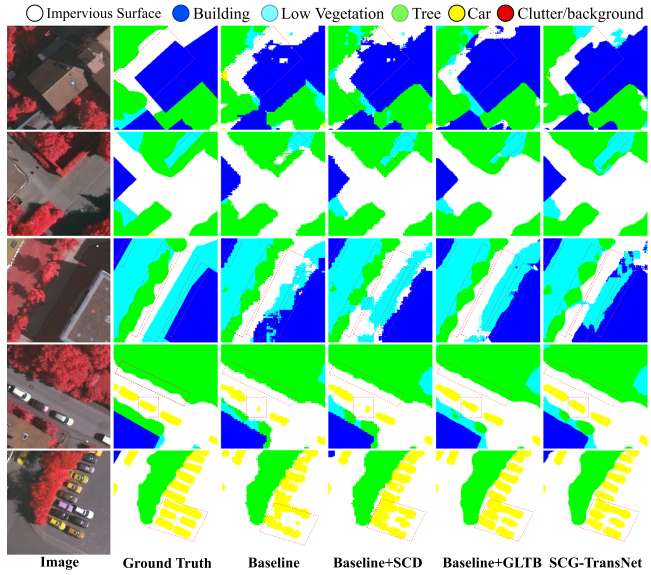

Fig. 5. Comparison of segmentation results before and after using SCD and GLTB in the SCG-TransNet framework.

follows where precision = TP/(TP + FP) recall = TP/(TP + FN). In addition, MIoU represents the average of IoU across all categories, and Ave.F1 score is the average of F1 across all categories.

C. Ablation Experiments

To evaluate the performance of the proposed network structure and two important modules, we used SCG-TransNet without the addition of SCD and GLTB as the baseline network and conducted ablation experiments on the Vaihingen dataset. The boldface of all tables in the text represents the maximum value of each column. In our baseline network, the Swin transformer is used for both encoder and decoder. In a large number of experiments comparing different hyperparameters, we select the hyperparameters with the best results to set the baseline network. In the encoder, the ratio of mlp hidden dim to embedding dim is set to 4, the patch size is set to 4 × 4 size and the patch norm is used, the stochastic depth rate is set to 0.1, attention dropout rate is set to 0, the hidden size is set to 96, the window size is 8, the number of layers corresponding to each stage is 2, 2, and 6, and the number of heads corresponding to each layer is 3, 6, and 12. In the decoder, the number of layers corresponding to each stage is 2, 2, and 2, and the number of heads corresponding to each layer is 3, 3, and 3.

1) Effect of SCD: Table I shows that MIoU, OA, and average F1 improve by 1.02%, 0.35%, and 0.79%, respectively, when SCD is considered in the SCG-TransNet framework. The car class shows the largest improvement in segmentation accuracy, with a 3.07% increase in IoU, followed by the building class with a 0.96% increase, validating the effectiveness of SCD in the network. As Fig. 5 shows, in the first row of the ablation experiment, the color of the "building" is not consistent on both sides due to the angle of incidence of the light and the strong

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

Method	Parameters (MB)	GFLOPs	Modules		IoU(%)					Evaluation index		
			SCD	GLTB	Impervious surface	Building	Low Vegetation	Tree	Car	MIoU(%)	Average F1(%)	OA(%)
Baseline	32.43	23.17	×	×	73.87	81.10	55.25	71.48	54.73	67.29	79.96	82.69
Baseline+SCD	38.74	24.30	✓	×	74.46	82.06	55.91	71.34	57.80	68.31	80.75	83.04
Baseline+GLTB	35.67	52.86	×	✓	**76.30**	**83.23**	57.49	**71.78**	58.80	69.52	81.61	83.47
Baseline+SCD+GLTB	41.98	52.97	✓	✓	74.59	82.69	**58.54**	71.53	**64.10**	**70.29**	**82.27**	**83.62**

illumination. The shadow from the "building" also obscures the "impervious surface" above it, resulting in a high intraclass variability with little interclass variability. Before the addition of the SCD, the black side of the "building," which was obscured by the shadow, showed varying degrees of holes, which were well mitigated by the introduction of the SCD. In the second row, the shadow from the high vegetation obscures the low vegetation in its immediate vicinity, and because the obscured low vegetation has very few and discontinuous pixel points, the model also shows varying degrees of holes in the segmentation before the introduction of SCD. The introduction of SCD effectively suppresses the negative effects of light and accurately separates out the obscured low vegetation, while at the same time mitigating the holes caused by the discontinuity of pixel points caused by the obscured low vegetation. The visualization of Fig. 5 shows that the introduction of SCD effectively mitigates the holes in the segmentation and improves the accuracy of target recognition for high interclass similarity.

2) Effect of GLTB: As shown in Table I, when GLTB is considered in the SCG-TransNet framework, MIoU, OA, and average F1 are improved by 2.23%, 0.78%, and 1.65%, respectively. The car class has the most improvement in segmentation accuracy, with a 4.07% improvement in IoU. The IoU of the other four classes "impervious surface," "building," "low vegetation," and "tree" improved by 2.43%, 2.13%, 2.24%, and 0.30%, respectively. As shown in Fig. 5, in the third row of the ablation experiment, under the influence of oblique sidelight, the shadow produced by "building" almost completely covers the low vegetation category, which is certainly very challenging for the model to identify. In the fifth row, the close proximity of the cars leads to the problem of shadows within the class obscuring each other. In the third and fourth rows, it can be seen that before the introduction of GLTB, the model does a very poor job of recognizing the obscured objects, not only incorrect vvcly detecting the obscured low vegetation as houses, but also missing the obscured cars. With the introduction of GLTB, the model accurately segmented the obscured low vegetation and cars, effectively reducing the negative effects of shadows. The image in the fifth row was taken in a car park, with cars in close proximity to each other, which tested the model's performance in segmenting small target edges in a densely distributed space. Before the introduction of GLTB, the model was unable to refine the features of each car, resulting in the edge pixels of the "car" being mixed together and unable to distinguish between cars. With the introduction of GLTB, the model almost perfectly separates the pixels that were previously predicted to be mixed in the car park, and separates the different cars. Secondly, in rows 4–6, before the introduction of GLTB, the model showed jagged edges for the segmentation of different categories. In contrast, after the introduction of GLTB, the model segmented the edges of different targets very smoothly, with almost no jagged fuzzy edges, and the model's performance in segmenting the edges of targets improved substantially. As can be seen from the visualization of the ablation experiments in the fourth, fifth, and sixth rows, GLTB shows excellent edge segmentation capability, effectively improving the situation of false detection or missed detection due to shadow obscuration from oblique sidelight.

3) Joint Effect: Table I reflects that the joint effect between the two modules is studied under the SCG-TransNet framework. When SCD and GLTB are introduced simultaneously, MIoU, OA, and average F1 are improved by 3.00%, 0.93%, and 2.31%, respectively. It is obvious that after adding GLTB, the IOU of the "car" class is increased by nearly 9.37%, followed by the "low vegetation" class, the segmentation accuracy is improved by 3.29%, and the remaining three classes "impervious surface," "building," the IoU of "tree" has increased by 0.72%, 1.59%, and 0.05%, respectively. From Fig. 5, we can clearly see that in the first row, the model effectively alleviates the hole phenomenon caused by high intensity light, which is caused by large intraclass difference and high interclass similarity. In the second, third, and fourth rows, the model effectively suppresses the negative effects of interclass and intraclass mutual occlusion caused by shadows generated by oblique side lights. In the fifth row, the model also achieves excellent segmentation performance in dense and complex small-object aggregation scenarios. In addition, we can clearly see that the SCG-TransNet combining SCD and GLTB has greatly improved the edge information localization ability of the model, and almost smoothes all segmentation edges.

D. Comparison With Other Methods

1) Results on the Vaihingen Dataset: Table II lists the experimental results of different existing methods. Our proposed SCG-TransNet achieves 70.29% MIoU and 82.27% average F1 segmentation, outperforming the other methods. In traditional CNN, the UNet network combines high-level semantic features from decoder and low-level features from encoder corresponding scales by using skip connections, Deeplabv3+ uses atrous convolutions with different dilation rates to build spatial pooling pyramids, and experimental data show that the segmentation effect is better than other traditional CNN methods. Compared to UNet, our model improves 3.94% on MIoU and 3.14% on average F1, and compared to Deeplabv3+, our model improves 3.13% on MIoU and 2.56% on average F1. UperNet and DANet with pyramidal structure are not as good as our SCG-TransNet in extracting global contextual information. Swin-UNet uses a pure transformer structure, which is not ideal in segmentation.

TABLE II
COMPARISON OF SEGMENTATION RESULTS ON THE VAIHINGEN DATASET

Method	Parameters (MB)	GFLOPs	IoU(%)					Evaluation index	
			Impervious surface	Building	Low segetation	Tree	Car	MIoU(%)	Average F1(%)
FCN[9]	22.70	5.37	73.22	78.97	54.80	70.38	39.92	63.46	76.65
UNet[26]	25.13	10.72	72.91	81.68	57.23	71.63	48.29	66.35	79.13
DeeplabV3+[12]	38.48	20.78	74.85	**83.01**	56.09	71.54	50.30	67.16	79.71
UperNet[66]	102.13	26.03	73.45	81.50	55.65	71.31	47.26	65.84	78.69
DANet[67]	45.36	125.78	73.54	81.40	56.88	71.21	42.68	65.14	78.00
TransUNet[68]	100.44	35.84	73.27	81.01	55.07	71.08	55.13	67.11	79.86
Swin-UNet[69]	25.89	7.70	69.31	73.37	49.48	67.12	30.78	58.01	72.02
ST-UNet[64]	160.97	78.69	76.36	82.98	57.79	**72.53**	61.48	70.23	82.15
SCG-TransNet(ours)	41.98	52.97	74.59	82.69	**58.54**	71.53	**64.10**	**70.29**	**82.27**

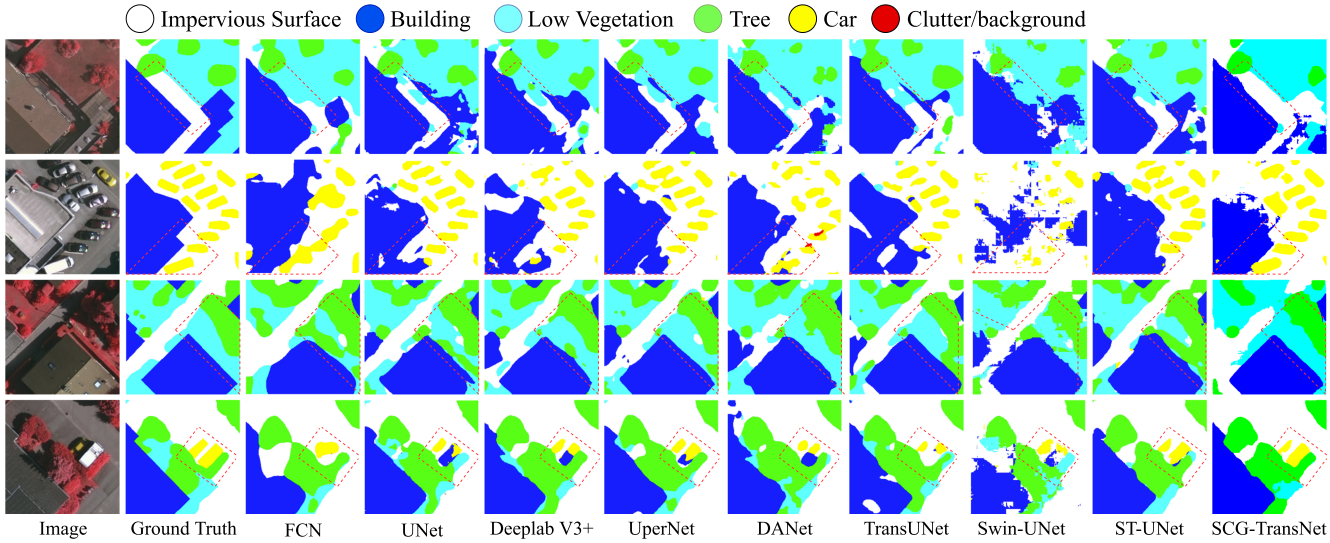

Fig. 6. Visualization results of the Vaihingen dataset.

TransUnet uses a transformer and CNN serial in the encoding stage structure for feature extraction, while ST-U-shaped network (ST-UNet) uses a parallel structure of transformer and CNN in the encoding phase. Compared with ST-UNet, our model improves 0.06% in MIoU and 0.12% in average F1, demonstrating the superiority of our proposed SCG-TransNet.

A visual comparison of several semantic segmentation methods used in Table II is shown in Fig. 6. Compared with other models, SCG-TransNet effectively alleviates the problem of poor segmentation accuracy caused by high similarity between classes due to differences in light intensity and light incident angle. As shown in the first line, the "impervious surface" in "building" and "low vegetation" is very similar to the adjacent "low vegetation" category due to light, and other models incorrectly identify "impervious surface" for "low vegetation", SCG-TransNet makes an accurate judgment. In the second row, "building" in the yellow box presents two colors of black and light gray due to the incident angle of the light. Especially in the yellow box, the white car has serious reflections under the action of strong light, which is very similar to the "building" class. Under such harsh environmental conditions, no other models can recognize the color of "building" in black and the "building" in strong reflection. "Car," while SCG-TransNet effectively eliminates the interference caused by light, and accurately recognizes the black "building" and the strongly reflective "car". In the third row, the shadows produced by high vegetation block nearby low vegetation. Under the influence of shadows, other models mistakenly identify the occluded low vegetation as "tree," while our model effectively extracts distinguishable feature information from the limited pixels of the occluded target, perfectly obstructed low vegetation is identified. In the fourth row, the white car also has reflected light, similar to the second row, and obscures the car next to it due to the difference in the height of the car and the angle of incidence of the light. Under the influence of reflected light, almost all other models misdetected the car as "building," but SCG-TransNet accurately identified and segmented it. As shown in Fig. 6, SCG-TransNet effectively identified the occluded target shows excellent segmentation performance.

2) Results on the POTSDAM Dataset: Table III shows the segmentation results of each method on the POTSDAM dataset. The proposed SCG-TransNet achieves 76.04% on MIoU and 86.20% on average F1, outperforming the results of other methods, demonstrating the superiority of the model. Among the traditional CNN models, Deeplabv3+ outperforms other traditional CNN segmentation models. Compared with Deeplabv3+, SCG-TransNet improves MIoU and average F1 by 1.56% and 1.07%, respectively. Compared with ST-UNet with 160.97 MB of parameters, SCG-TransNet with only 26% of ST-UNet parameters still surpasses 0.07% and 0.12% in MIoU and average

TABLE III
COMPARISON OF SEGMENTATION RESULTS ON THE POTSDAM DATASET

Method	Parameters (MB)	GFLOPs	IoU(%)					Evaluation index	
			Impervious surface	Building	Low vegetation	Tree	Car	MIoU(%)	Average F1(%)
FCN[9]	22.70	5.37	77.41	83.52	66.10	63.19	74.34	72.91	84.12
UNet[26]	25.13	10.72	77.10	82.83	64.59	65.44	76.16	73.22	84.35
DeeplabV3+[12]	38.48	20.78	79.01	84.76	67.53	63.05	78.05	74.48	85.13
UperNet[66]	102.13	26.03	76.95	83.93	65.65	60.40	76.57	72.70	83.91
DANet[67]	45.36	125.78	77.35	83.45	66.46	63.47	75.28	73.20	84.32
TransUNet[68]	100.44	35.84	78.61	85.60	67.16	64.10	79.33	74.96	85.44
Swin-UNet[69]	25.89	7.70	71.45	75.02	59.03	50.96	71.15	65.52	78.79
ST-UNet[64]	160.97	78.69	79.19	**86.63**	67.89	66.37	79.77	75.97	86.13
SCG-TransNet(ours)	41.98	52.97	78.41	86.11	**68.07**	**67.65**	**79.93**	**76.04**	**86.20**

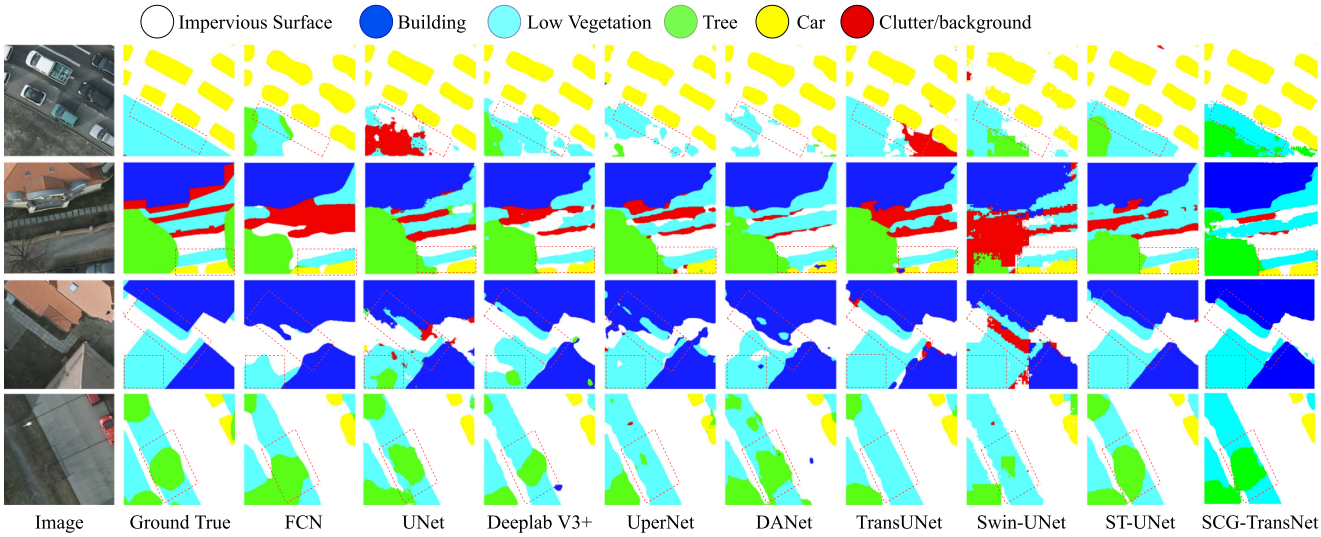

Fig. 7. Visualization results of the POTSDAM dataset.

F1, achieving SOTA results. Compared with Swin-UNet, SCG-TransNet improves MIoU and average F1 by 10.52% and 7.41%, respectively.

The visualization of segmentation results for each model is shown in Fig. 7. Looking at the first row, the similarity between "low vegetation" and its adjacent "impervious surface" is very high, and both appear dark gray. Obviously, the ability of the model to localize edge information is particularly important in the face of adjacent targets with such high interclass similarity. From the comparison of segmentation results of different models, we can clearly see that our model has the best performance for segmentation boundaries when the two classes are adjacent and the similarity between classes is so high. Looking at the third row, in the yellow box above, the sides of the "impervious surface" are surrounded by low vegetation. In the bottom yellow box, the shadows from the two houses cause the low vegetation between them to appear black, making it difficult for the model to be identified. Observe Fig. 7, some models mistakenly detect it as "tree," while our SCG-TransNet effectively suppresses the adverse effects of shadows, accurately segment the low vegetation that is obscured, and correctly distinguishes included between "building," "low vegetation," and "impervious surface." In the fourth row, the "tree" in the yellow box is almost the same color as its nearby low vegetation due to the light intensity. For this reason, the segmentation edges of the "tree" in this box are very blurred by other models and appear to have various degrees of false detection. And our SCG-TransNet effectively improves the phenomenon of false detection, blurred boundary, and boundary fault, showing strong segmentation performance.

3) Parameter and Computation Complexity Analysis: Analyzing in terms of parameters, the proposed model has only 41.98 MB of parameters, which is far less than the 160.97 MB parameter of the SOTA model, and only 26% of the SOTA model. And the obtained 70.29%, 76.04% Miou, and 82.27% on international society for photogrammetry and remote sensing (ISPRS)-Vaihingen dataset and ISPRS-Potsdam dataset, respectively, 86.20% average F1 surpassed SOTA's 70.23%, 75.97% Miou and 82.15%, 86.13% average F1. In addition, Swin-UNet has only 25.89 MB of parameters, although the parameter amount is very small, but due to its lack of convolution operation, it ignores the attention to local information, which is obviously not suitable for RS images with a large number of different scales, resulting in poor performance. The final performance of FCN with only 22.70 MB of parameters is also poor due to the lack of attention to global information. Analyzing in terms of model complexity, models with transformer or Swin transformer blocks usually have greater computational complexity than traditional CNN semantic segmentation models. For example, the GFLOPs of TransUNet and ST-UNet are 35.84 G and 78.69 G, respectively. Compared with the GFLOPs of the

traditional CNN semantic segmentation models DeeplabV3+ and UpperNet, which are only 20.78 G and 26.03 G, while the computational complexity GFLOPs of DANet with multiple attention is as high as 125.78 G. The computational complexity of our SCG-TransNet is 52.97 G. Although SCG-TransNet does not have an advantage in computational complexity when compared with the traditional CNN semantic segmentation models, it still has a great effect when applied to scenarios where the model efficiency requirements are not very high. At the same time, it is still valuable for exploring how Swin Transformer can be better applied to the field of RS with such a complex environment.

V. Conclusion

In this work, we propose SCG-TransNet, a semantic segmentation framework combining Swin transformer and Deeplabv3+. Compared with models based on CNN backbone network, the Swin transformer does not contain inductive bias, which allows for better representation of long-range dependencies. Compared with other transformers, Swin transformer has lower computational complexity, fewer parameters, and output of hierarchical feature maps. The proposed SCD captures multiscale information of features by combining the excellent local feature extraction ability of convolution and the powerful capture ability of Swin transformer in global context information, so as to obtain more discriminative features and effectively inhibits the noise caused by shadow occlusion caused by light. In addition, the GLTB with SPPS can make full use of the limited pixels of the occluded object to generate distinguishable representation information, effectively alleviate the situation of false detection or even missed detection caused by the occlusion of the target, and greatly improve the model's localization of edge information ability. In various RS image semantic segmentation, SCG-TransNet shows great potential for constructing long-range dependencies and outperforms other SOTA VITs in our experiments. In the future, we will continue to improve and optimize the model, expecting that the model can be more lightweight while ensuring the segmentation ability, and can be applied to a variety of different fields.

References

[1] H. Shafizadeh-Moghadam et al., "Google earth engine for large-scale land use and land cover mapping: An object-based classification approach using spectral, textural and topographical factors," *GISci. Remote Sens.*, vol. 58, no. 6, pp. 914–928, 2021.

[2] 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.

[3] J. Zhao, Y. Zhou, B. Shi, J. Yang, D. Zhang, and R. Yao, "Multistage 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.

[4] L. Ding, J. Zhang, and L. Bruzzone, "Semantic segmentation of largesize 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.

[5] D. Y. Chen et al., "Building extraction and number statistics in WUI areas based on UNet structure and ensemble learning," *Remote Sens.*, vol. 13, no. 6, 2021, Art. no. 1172.

[6] C. Ayala et al., "A deep learning approach to an enhanced building footprint and road detection in high-resolution satellite imagery," *Remote Sens.*, vol. 13, no. 16, 2021, Art. no. 3135.

[7] H. A. T. Nguyen et al., "Integrating remote sensing and machine learning into environmental monitoring and assessment of land use change," *Sustain. Prod. Consumption*, vol. 27, pp. 1239–1254, 2021.

[8] Z. Sun et al., "Use remote sensing and machine learning to study the changes of broad-leaved forest biomass and their climate driving forces in nature reserves of northern subtropics," *Remote Sens.*, vol. 14, no. 5, 2022, Art. no. 1066.

[9] 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.

[10] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[11] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.

[12] 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.

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

[14] J. Wu, Z. Pan, B. Lei, and Y. Hu, "FSANet: Feature-and-spatial-aligned network for tiny object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art no. 5630717.

[15] S. Lei and Z. Shi, "Hybrid-scale self-similarity exploitation for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–10, 2022.

[16] R. Yang, F. Pu, Z. Xu, C. Ding, and X. Xu, "DA2Net: Distraction-attention-driven adversarial network for robust remote sensing image scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[17] P. W. Battaglia et al., "Relational inductive biases, deep learning, and graph networks," 2018, *arXiv:1806.01261*.

[18] Q. Liu et al., "Self-constructing graph neural networks to model long-range pixel dependencies for semantic segmentation of remote sensing images," *Int. J. Remote Sens.*, vol. 42, no. 16, pp. 6184–6208, 2021.

[19] A. Wiacek, E. González, and M. A. L. Bell, "CohereNet: A deep learning architecture for ultrasound spatial correlation estimation and coherence-based beamforming," *IEEE Trans. Ultrasonics, Ferroelectrics, Freq. Control*, vol. 67, no. 12, pp. 2574–2583, Dec. 2020.

[20] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6836–6846.

[21] W. Wang et al., "PVT V2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, no. 3, pp. 415–424, 2022.

[22] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[23] H. Lin, X. Cheng, X. Wu, and D. Shen, "CAT: Cross attention in vision transformer," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2022, pp. 1–6.

[24] R. Shao et al., "Localtrans: A multiscale local transformer network for cross-resolution homography estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 14890–14899.

[25] K. Zhang et al., "Practical blind denoising via swin-conv-unet and data synthesis," 2022, *arXiv:2203.13278*.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, vol. 9351, pp. 234–241.

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

[28] F. Zhu, Y. Zhu, L. Zhang, C. Wu, Y. Fu, and M. Li, "A unified efficient pyramid transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 2667–2677.

[29] R. Azad et al., "TransDeepLab: Convolution-free transformer-based deeplab v3 for medical image segmentation," in *Proc. Int. Workshop PRedictive Intell. Med.*, Springer, Cham, 2022, pp. 91–102.

[30] W. Liu et al., "Remote sensing image segmentation using dual attention mechanism Deeplabv3 algorithm," *Trop. Geography*, vol. 40, pp. 303–313, 2020.

[31] B. Baheti et al., "Semantic scene segmentation in unstructured environment with modified DeepLabV3," *Pattern Recognit. Lett.*, vol. 138, pp. 223–229, 2020.

[32] O. Akcay, A. C. Kinaci, E. O. Avsar, and U. Aydar, "Semantic segmentation of high-resolution airborne images with dual-stream DeepLabV3," *ISPRS Int. J. Geo- Inf.*, vol. 11, no. 1, p. 23, 2021. [Online]. Available: https://doi.org/10.3390/ijgi11010023
[33] Z. Wang et al., "Semantic segmentation of high-resolution remote sensing images based on a class feature attention mechanism fused with Deeplabv3," *Comput. Geosci.*, vol. 158, 2022, Art. no. 104969.
[34] Y. Wang, S. Wang, and X. B. Hong, "Road extraction using high resolution satellite images based on Receptive Field and Improved Deeplabv3," in *Proc. J. Phys.: Conf. Ser.*, IOP Publishing, 2022, vol. 2320, no. 1, Art. no. 012021.
[35] J. Li, B. Sun, S. Li, and X. Kang, "Semisupervised semantic segmentation of remote sensing images with consistency self-training," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5615811.
[36] K. Chowdhary, "Natural language processing," in *Fundamentals of Artificial Intelligence*, Springer, pp. 603–649, 2020.
[37] W. Li et al., "MSNet: A multi-stream fusion network for remote sensing spatiotemporal fusion based on transformer and convolution," *Remote Sens.*, vol. 13, no. 18, 2021, Art. no. 3724.
[38] L. Gao et al., "STransFuse: Fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10990–11003, 2021.
[39] G. Chen et al., "SwinSTFM: Remote sensing spatiotemporal fusion using swin transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5410618.
[40] Q. Li, Y. Chen, and Y. Zeng, "Transformer with transfer CNN for remote-sensing-image object detection," *Remote Sens.*, vol. 14, no. 4, 2022, Art. no. 984.
[41] B. He et al., "UnityShip: A large-scale synthetic dataset for ship recognition in aerial images," *Remote Sens.*, vol. 13, no. 24, 2021, Art. no. 4999.
[42] W. Zhang, L. Jiao, Y. Li, Z. Huang, and H. Wang, "Laplacian feature pyramid network for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5604114.
[43] 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.
[44] M. Kaselimi, A. Voulodimos, I. Daskalopoulos, N. Doulamis, and A. Doulamis, "A vision transformer model for convolution-free multilabel classification of satellite imagery in deforestation monitoring," *IEEE Trans. Neural Netw. Learn. Syst.*, 2022, doi: 10.1109/TNNLS.2022.3144791.
[45] C. Zhang, W. Jiang, Y. Zhang, W. Wang, Q. Zhao, and C. Wang, "Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4408820.
[46] L. Sun, G. Zhao, Y. Zheng, and Z. Wu, "Spectral–spatial feature tokenization transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5522214.
[47] 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.
[48] Y. Chen et al., "Image super-resolution reconstruction based on feature map attention mechanism," *Appl. Intell.*, vol. 51, no. 7, pp. 4367–4380, 2021.
[49] J. Li et al., "Change detection for high-resolution remote sensing images based on a multi-scale attention siamese network," *Remote Sens.*, vol. 14, no. 14, 2022, Art. no. 3464.
[50] X. Liu et al., "Self-attention negative feedback network for real-time image super-resolution," *J. King Saud Univ.- Comput. Inf. Sci.*, vol. 34, no. 8, pp. 6179–6186, 2022.
[51] X. Hu et al., "HDNet: High-resolution dual-domain learning for spectral compressive imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17542–17551.
[52] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision transformer with deformable attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4794–4803.
[53] C. Zhang, H. Wan, X. Shen, and Z. Wu, "PatchFormer: An efficient point transformer with patch attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11799–11808.
[54] L. Sun, S. Cheng, Y. Zheng, Z. Wu, and J. Zhang, "SPANet: Successive pooling attention network for semantic segmentation of remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4045–4057, 2022, doi: 10.1109/JSTARS.2022.3175191.
[55] Y. Liu et al., "NAM: Normalization-based attention module," 2021, *arXiv:2111.12419*.
[56] L. C. Chen et al., "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[57] J. Gutter and J. Niebling, and X. X. Zhu, "Analyzing the interactions between Training Dataset Size, Label Noise and Model Performance in Remote Sensing Data," in *Proc. IEEE IGARSS Int. Geosci. Remote Sens. Symp.*, 2022, pp 303–306.
[58] C. Xiao et al., "Image inpainting detection based on high-pass filter attention network," *Comput. Syst. Sci. Eng.*, vol. 43, no. 3, pp. 1146–1154, 2022.
[59] A. I. Shaikh and S. S. Badroddin, "Noise reduction from L-band ALOSPALSAR data set using spatial domain Gaussian low-pass filter," *Int. Res. J. Adv. Sci. Hub*, vol. 3, pp. 87–93, 2021.
[60] L. Wang et al., "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogrammetry Remote Sensing*, 190, pp. 196–214, 2022.
[61] 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.
[62] X. Li et al., "SPCS: A spatial pyramid convolutional shuffle module for YOLO to detect occluded object," *Complex Intell. Syst.*, vol. 9, no. 1, pp. 301–315, 2022.
[63] I. Vaihingen, 2D semantic labeling dataset. Accessed: Apr. 2018.
[64] 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, 2022, Art. no. 4408715.
[65] 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, 2016, pp. 565–571.
[66] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "unified Perceptual Parsing for Scene Understanding," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany: Springer, 2018, vol. 11209, pp. 432–448.
[67] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 3146–3154.
[68] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.
[69] H. Cao et al., "Swin-Unet: Unet-like pure transformer for medical image segmentation," in *Proc. Comput. Vis.-ECCV 2022 Workshops*, Cham, Springer Nature Switzerland, 2023, pp. 205–218.

Youda Mo was born in June 2002 in Guangdong, China. He is currently a Junior Student studying data science and big data technology with the College of Guangdong Polytechnic Normal University. He will receive the B.S. degree in June, 2024.

His research interests include semantic segmentation of remote sensing images, especially the segmentation of high-resolution remote sensing images in complex environments.

Huihui Li received the Ph.D. degree in computer science and engineering from the South China University of Technology, Guangzhou, China, in 2019.

She is currently a Lecturer with the School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, China. Her current research interests include image processing, pattern recognition, and emotional computing.

Xiangling Xiao received the B.S. degree in faculty of intelligent manufacturing from Wuyi University, Jiangmen, China, in 2020. She is currently working toward the master's degree with the School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, China.

Her research interests include deep learning and computer vision.

Huimin Zhao was born in Shanxi, China, in 1966. He received the B.Sc. and M.Sc. degrees in signal processing from Northwestern Polytechnical University, Xi'an, China, in 1992 and 1997, respectively, and the Ph.D. degree in electrical engineering from the Sun Yat-sen University, Guangzhou, China, 2001.

He is currently a Professor with and the Dean of the School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, China. His research interests include image, video, and information security technology.

Xiaoyong Liu received the Ph.D. degree in library science from National Science Library, Chinese Academy of Sciences, Beijing, China, in 2011.

He is currently a Professor with the School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, China. His current research interests include image processing, pattern recognition, and natural language processing.

Jin Zhan received the Ph.D. degree in Computer Application from Sun Yat-sen University, Guangzhou, China, in 2015.

She is currently an Associate Professor with the School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, China. Her research interests include image and video intelligent analysis, machine learning, and computer vision.