TCNet: Multiscale Fusion of Transformer and CNN for Semantic Segmentation of Remote Sensing Images

Xuyang Xiang, Wenping Gong, Shuailong Li, Jun Chen, *Member, IEEE*, and Tianhe Ren

***Abstract*—Semantic segmentation of remote sensing images plays a critical role in areas such as urban change detection, environmental protection, and geohazard identification. Convolutional Neural Networks (CNNs) have been excessively employed for semantic segmentation over the past few years; however, a limitation of the CNN is that there exists a challenge in extracting the global context of remote sensing images, which is vital for semantic segmentation, due to the locality of the convolution operation. It is informed that the recently developed Transformer is equipped with powerful global modeling capabilities. A network called TCNet is proposed in this article, and a parallel-in-branch architecture of the Transformer and the CNN is adopted in the TCNet. As such, the TCNet takes advantage of both Transformer and CNN, and both global context and low-level spatial details could be captured in a much shallower manner. In addition, a novel fusion technique called Interactive Self-attention is advanced to fuse the multilevel features extracted from both branches. To bridge the semantic gap between regions, a skip connection module called Windowed Self-attention Gating is further developed and added to the progressive upsampling network. Experiments on three public datasets (i.e., Bijie Landslide Dataset, WHU Building Dataset, and Massachusetts Buildings Dataset) depict that TCNet yields superior performance over state-of-the-art models. The IoU values obtained by TCNet for these three datasets are 75.34% (ranked first among 10 models compared), 91.16% (ranked first among 13 models compared), and 76.21% (ranked first among 13 models compared), respectively.**

***Index Terms*—Convolutional Neural Network (CNN), feature fusion, remote sensing images, semantic segmentation, Transformer.**

I. Introduction

With the rapid advancement of aerospace and sensor technology [1], plenty of high-quality remote sensing images can be accessed by the public; based on these images, the state of the environment and human activity traces might be easily disclosed [2], [3], [4]. Several techniques have been explored for extracting relevant information in remote sensing images [5]. Among the various techniques available, semantic segmentation, the aim of which is to determine the semantic category of each pixel in the image of interest [6], has attracted increasing popularity over the past few years [7]. It should be noted that semantic segmentation of remote sensing images has been successfully implemented in various application scenarios such as environmental protection [8], urban change detection [9], and geohazard identification [10].

The Convolutional Neural Network (CNN) has always been deemed the most popular deep learning model, because of its remarkable ability in feature extraction and high level of automation [11]. Indeed, the Fully Convolutional Network (FCN) created by Long et al. [12] could be taken as the prototype of most CNN models for semantic segmentation that are available in the literature. Among the various modifications of FCN, the encoder–decoder structure, which exhibits excellent segmentation performance, has become the most popular structure configuration [13]. For example, UNet [14] utilizes a decoder to learn the spatial correlation of image features in encoding stages; and DeepLab V3+ [15] involves a decoder based on DeepLab V3 [16] to integrate spatial features, and the network performance is thus considerably improved. Note that while the CNN models may achieve good performance, multiscale information of the concerned images cannot be fully utilized due to the design limitations of the decoder [17]. Nonetheless, due to the complicated backgrounds and the existence of a lot of noise in the images, global context information and reasoning capabilities for fine spatial features must be enabled for effective semantic segmentation of remote sensing images. In other words, the conventional CNN models are not perfect for the semantic segmentation of remote sensing images [18], and further improvement is warranted.

To address this limitation of the conventional CNN models, attention mechanisms and multiscale feature fusion strategies, such as channel attention and position attention module [19], criss-cross attention module [20], pyramid pooling module [21], and multilevel feature fusion strategy [22], have often been adopted. However, the global information captured by these approaches is not encoded from the global modeling directly [6]; rather, it is mainly constructed of the local features captured with the existing CNN models; as such, the global scene information

Manuscript received 15 September 2023; revised 1 December 2023; accepted 31 December 2023. Date of publication 4 January 2024; date of current version 18 January 2024. This work was supported in part by the Outstanding Youth Foundation of Hubei Province, China under Grant 2022CFA102, and in part by the National Natural Science Foundation of China under Grant 41977242. *(Corresponding author: Wenping Gong.)*

Xuyang Xiang, Wenping Gong, Shuailong Li, and Tianhe Ren are with the Faculty of Engineering, China University of Geosciences, Wuhan 430074, China (e-mail: xiangxuyang@cug.edu.cn; wenpinggong@cug.edu.cn; li_shuailong@cug.edu.cn; rentianhe@cug.edu.cn).

Jun Chen is with the School of Automation, China University of Geosciences, Wuhan 430074, China (e-mail: chenjun71983@163.com).

Digital Object Identifier 10.1109/JSTARS.2024.3349625

might not be captured yet [23]. The Transformer, well-known in natural language processing (NLP) for its exceptional capability in capturing global relationships, offers a viable answer to semantic segmentation [24], and astounding performance could be achieved by the Transformer in the field of computer vision. For example, a Transformer encoder–decoder architecture was developed to model the relations between the objects and the global image context [25], and a dual-branch Transformer structure was designed to learn the multiscale feature of images [26]. Further, the Swin Transformer was constructed, and it exhibits great potential in image classification and dense prediction tasks [27]. It is noted that while Transformer-based models have been widely adopted in other image segmentation and huge progress has been achieved [6], their application in semantic segmentation of remote sensing images still needs to be explored. It is noted that although the Transformer can capture the long-range dependence of global features effectively, the local feature information is frequently disregarded. In order to effectively segment remote sensing images using semantics, it is crucial to create a model that can benefit from both the Transformer and the CNN, and thus, both the global context and fine spatial details could be effectively captured [28].

In this study, a network called TCNet is proposed and a parallel-in-branch architecture is utilized. In the context of the TCNet, a Transformer branch is employed to obtain the global context, whereas a CNN branch is adopted to capture the low-level spatial details. In addition, a novel fusion technique called Interactive Self-attention (ISa) is advanced to fuse the multilevel features extracted from both branches; and, to bridge the semantic gap between different regions, a skip connection module called Windowed Self-attention Gating (WSaG) is developed and added to the progressive upsampling network (PUN).

The main contributions of this article are given as follows.

1) A novel TCNet is proposed to achieve accurate semantic segmentation. The parallel-in-branch architecture of the TCNet uses Transformer and CNN to extract local and global feature information. The ISa effectively couples the information that is extracted from the parallel-in-branch architecture. To bridge the semantic gap between various regions, the linked multiscale feature information improves the interaction between features using the WSaG-based PUN, thus improving the accuracy of segmentation.
2) To effectively couple the coding features from CNN and Transformer, ISa is designed. ISa contains the attention residual (ARE) block, the efficient self-attention (ESA) module, the interactive efficient self-attention (IESA) module, and the Residual block, which can help focus on focal area information.
3) The WSaG-based PUN is used as a decoder to ensure effective multiscale feature interaction. WSaG is used to create a within windows and cross-window transfer of features, making more efficient use of features. Features between different levels are connected by upsampling, as such, bottom-up information interactions can be realized.

The rest of this article is organized as follows. First, the studies on semantic segmentation of remote sensing images are reviewed. Second, the methodology of the TCNet is detailed. Third, illustrative applications are presented based on three public datasets (i.e., Bijie Landslide Dataset, WHU Building Dataset, and Massachusetts Buildings Dataset). Fourth, comparisons are conducted to depict the superiority of the TCNet over state-of-the-art models. Fifth, the model efficiency and the significance of each branch are discussed. Finally, the concluding remarks are drawn.

II. Related Work

CNNs and Transformers are two types of methods that could be used for the semantic segmentation of remote sensing images. A short literature review of these two methods is presented in this section.

A. CNN-Based Semantic Segmentation of Remote Sensing Images

CNNs were mainly developed based on artificial neural networks. Because of its exceptional performance, CNN has gained increased popularity in various areas [29]. The FCN, created by Long et al. [12], could be taken as the prototype of most CNN models used in the existing semantic segmentation of remote sensing images [30], [31], [32], [33]. In comparison to the classical CNN, the fully connected layers are replaced by convolutional layers in the FCN; as such, the FCN is equipped with the capability to make predictions on arbitrary-sized inputs and the pixels-to-pixels mapping can be learned by the networks, without extracting the region proposals [12], [34], [35]. The existing FCN-based models are best suited for local tasks, not global tasks, due to their particular structure [35]. For example, rather than object classification, the FCN-based models could be more suitable for semantic segmentation or object detection.

The resolution of the predictions generated by the FCN is low due to the intrinsic limitation of the simple decoder used, and the boundaries of the object recognized are fuzzy. The encoder–decoder structure was subsequently created by building symmetrical decoders like UNet [14] and SegNet [36] to overcome this problem; and, the spatial resolution of extracted features could then be restored progressively. Further, to improve the effectiveness of the encoder–decoder structure in capturing richer contextual features and reducing the loss of feature information, various enhancing techniques such as deep deconvolution network [37] and atrous convolution [38] have been developed and included in the modified CNNs. The encoder–decoder structure has emerged as a dominating structure configuration in semantic segmentation [13]. However, the improved CNN models are not yet capable of correctly identifying complex objects in remote sensing images [17]. In such a situation, an attempt, based on attention mechanisms and multiscale feature fusion strategies, has been conducted to improve the segmentation precision through exploiting the contextual information. For example, a linear attention mechanism was constructed and added to each skip connection to establish long-term dependencies of the feature map [39]; a top-down strategy was advanced to fuse high-level features with shallow low-level features, acquired by the deep and shallow layers, respectively [33]; and a multiscale skip connection network was designed to realign semantic features of different levels [40].

Recent advances in CNNs have promoted the semantic segmentation of remote sensing images. The CNN models mentioned above have been successfully applied in various areas, such as environmental protection [8], urban change detection [9], and geohazard identification [10]. However, the CNN models modified are mainly based on convolution operations, which are not liberated from the original CNN structure. In summary, the CNN models are not perfect, and limitations in acquiring the global information of remote sensing images, intraclass differences of which are large, whereas interclass differences are small, are evident [41], [42].

B. Transformer-Based Semantic Segmentation of Remote Sensing Images

The Transformer was initially developed for NLP, and it has been shown that its precision is greater than that of conventional sequence transduction models based on complex recurrent networks or CNNs [43]. Then, several Transformer-based models have been developed for semantic segmentation of remote sensing images [24]. Transformer-based models typically perform better in global context modeling than the CNN models mentioned above due to their strong capabilities in sequence-to-sequence modeling [17].

It should be noted that the majority of Transformer-based semantic segmentation models employ an encoder–decoder structure, and these models can be broadly classified into two categories. The first category is solely based on Transformers and the representative models are Segmenter [44], SegFormer [45], CrackFormer [46], and SwinUNet [47]. As the Transformer primarily focuses on global modeling and lacks localization capabilities, the investigations by Wang et al. [5], Chen et al. [48], Zhang et al. [49], Long et al. [50], and Zhang et al. [51] show that the pure Transformer-based segmentation networks may generate unacceptable performance. On the other hand, hybrid architectures are oftentimes adopted in the second category. For example, a dual-branch encoder, which is based on the Transformer and the CNN, was created for urban scene understanding [17]. By integrating the Swin Transformer into the traditional CNN-based UNet, a new dual encoder structure was created [6]; and, to model both global and local information more effectively, a Transformer-based decoder was designed and the lightweight ResNet-18 was selected as the encoder [5]. It is noted that although Transformer-based models have been widely used in the segmentation of medical images [6], their use in the semantic segmentation of remote sensing images has been limited. Inspired by the studies discussed above, a network called TCNet, in which the parallel-in-branch architecture is utilized, is proposed in this study. In the context of the proposed TCNet, a Transformer branch is employed to obtain the global context, whereas a CNN branch is adopted to capture the low-level spatial details.

III. Methodology of the TCNet

In this part, the architecture and key modules of the proposed TCNet are described in depth. The ISa module and the WSaG-based PUN are introduced after outlining the overall structure of the TCNet.

A. Overall Structure of the TCNet

As shown in Fig. 1, two parallel feature extraction branches are adopted in the TCNet, in which, the branch of ResNet-34 [52] is employed to encode the local features, whereas that of eight-layer DeiT-Small (DeiT-S) [53] is adopted to encode the global features. Note that the ResNet networks can improve the link between different layers of the network, allowing for more plentiful expression of high-resolution features, whereas a teacher–student strategy is taken by the DeiT; thus, much fewer data are demanded for model training and better convergence performance could be achieved. Further, both ResNet-34 and DeiT-Small are lightweight models. In summary, a parallel-in-branch architecture, which is based on the ResNet-34 and DeiT-Small, is adopted in the proposed TCNet.

There are five blocks embedded in the ResNet-34, each block downsamples the feature maps by a factor of 2. The outputs derived from the fourth ($g^0 \in R^{\frac{H}{16}\times\frac{W}{16}\times 256}$), third ($g^1 \in R^{\frac{H}{8}\times\frac{W}{8}\times 128}$), and second ($g^2 \in R^{\frac{H}{4}\times\frac{W}{4}\times 64}$) blocks of the CNN branch ResNet-34 are fused with the results derived from the Transformer branch of eight-layer DeiT-Small. Note that a typical encoder structure is taken in the eight-layer DeiT-Small. Specifically, an input image $F \in R^{H\times W\times 3}$ is first equally divided into $N = \frac{H}{16}\times\frac{W}{16}$ patches, which are then flattened and passed to a linear embedding layer with an output dimension of D_0; and, as a result, raw embedding sequence $z^0 \in R^{N\times D0}$ can be derived. The resulting embeddings $z^0 \in R^{N\times D0}$ are inputted to the encoder of eight-layer DeiT-Small, which contains eight layers of multiheaded self-attention (MSA) and multilayer perceptron (MLP). It should be noted that a layer normalization exists in front of the MSA and MLP (of each layer). The output of the encoder is further reshaped to a feature map $t^0 \in R^{\frac{H}{16}\times\frac{W}{16}\times 384}$. Finally, the spatial resolution of the reshaped feature map $t^0 \in R^{\frac{H}{16}\times\frac{W}{16}\times 384}$ is recovered with the aid of two consecutive standard upsampling-convolutional layers; as an outcome, the feature maps $t^1 \in R^{\frac{H}{8}\times\frac{W}{8}\times 128}$ and $t^2 \in R^{\frac{H}{4}\times\frac{W}{4}\times 64}$ are sequentially obtained. The feature maps of different scales (i.e., t^0, t^1, and t^2) are then fused with those obtained from the CNN branch (i.e., g^0, g^1, and g^2), as illustrated in Fig. 1.

It is noted that the feature maps obtained from the Transformer branch (i.e., t^0, t^1, and t^2) and those from the CNN branch (i.e., g^0, g^1, and g^2), in the proposed TCNet, are fused with the module of ISa. Here, the local texture feature g^i and global context feature t^i ($i = 0$, 1, and 2) are correspondingly inputted to the ISa module. As both local and global features are fused, the representation of the context in images can be more complete and more compact. Four blocks are involved in the ISa module proposed, including the ARE block, the ESA module [54], the IESA module, and the Residual block. It is noted that the ARE block, IESA module, and Residual block are specially developed in this article. The contextual semantics within large neighborhoods of the local and global context features (i.e., inputs to the ISa module) are first extracted with the aid of the ARE block; and, the resulting semantics are then cross-fused with the modules of ESA and IESA separately. Finally, the results obtained from the ESA module and those from the IESA module are processed by the Residual block; as such, the features of multiple scales could be selectively emphasized. Afterward, the fused feature map

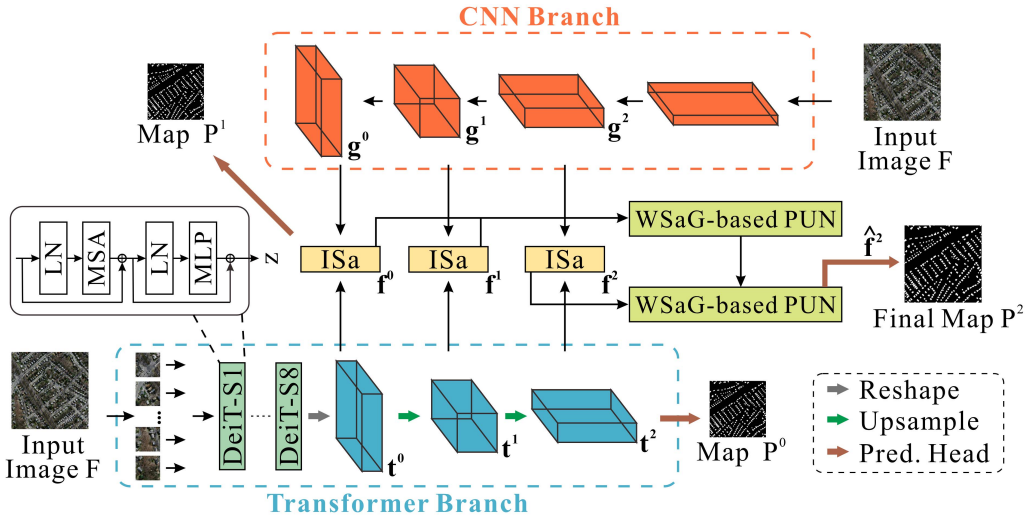

Fig. 1. Structure of the proposed TCNet.

f^i ($i = 0$, 1, and 2), the output of the ISa module, is transferred to the WSaG-based PUN. As an outcome, a final segmentation map P^2 could be generated by a simple head (Pred. Head). To optimize the network of the TCNet, two auxiliary heads are adopted to generate two segmentation maps P^0 and P^1 as the intermediate products [55]. The operation of the TCNet is expressed as follows:

$$g^i = \mathrm{CNN}(F) \tag{1}$$

$$t^i = \mathrm{Transformer}(F) \tag{2}$$

$$f^i = \mathrm{ISa}(g^i, t^i) \tag{3}$$

$$P^0 = \mathrm{Pred.Head}(t^2) \tag{4}$$

$$P^1 = \mathrm{Pred.Head}(f^0) \tag{5}$$

$$P^2 = \mathrm{Pred.Head}(\mathrm{PUN}(\mathrm{PUN}(f^0, f^1), f^2)). \tag{6}$$

B. Module of ISa

To couple the local and global features, obtained by CNN and Transformer, respectively, more effectively, an ISa module is developed and employed in the proposed TCNet. As mentioned above, four blocks, in terms of the ARE block, the ESA module, the IESA module, and the Residual block, are involved in the ISa module, as shown in Fig. 2. Inspired by the Attention U-net proposed by Oktay et al. [56], atrous convolutions (AConv) are adopted in the ARE block to capture contextual semantics within a large neighborhood; and, the context contrast in the obtained feature map is enhanced with the aid of the ESA module. Further, the relationships between different feature maps obtained by the CNN and Transformer branches are learned by the IESA module. Finally, the contextual information obtained by the ESA module and that obtained by the IESA module are processed by the Residual block. Thus, the information interference from irrelevant regions is suppressed, the features of multiple scales could be selectively emphasized, and the number of channels is reduced.

The procedures for the feature fusion with the ISa module are summarized as follows.

1) The local texture feature g^i and global context feature t^i ($i = 0$, 1, and 2) are inputted to the ARE block, and the inputted feature g^i and t^i ($i = 0$, 1, and 2) are processed by three different convolutional layers (a 1×1 convolutional layer and two 3×3 AConv layers) separately. To improve the convergence and generalization of the TCNet, the feature obtained from each convolutional layer is further processed by a batch normalization (BN) layer. Finally, an elementwise sum operation is conducted to refine the obtained features.
2) The feature maps obtained from the ARE block are cross-fused with the modules of ESA and IESA separately. The modules of ESA and IESA are, respectively, based on two novel self-attentions. A pyramid pooling operation is included in the novel self-attentions, in comparison to the original self-attentions. The feature obtained from the novel self-attentions is further processed by a feedforward (FF) layer and a reshaping operation sequentially. For simplicity, the residual addition between the input and the output of FF is omitted. Finally, to learn the relationships between different feature maps obtained from the CNN and Transformer branches, a joint mechanism is created and included in the IESA module.
3) The results obtained from the ESA module and those from the IESA module are concatenated via a channel-wise concatenation operation, and the outcome is then inputted to the Residual block. The number of channels of the inputted feature map is first reduced by two 1×1 convolutional layers; and, the derived feature map is further processed by three different convolutional layers (a 3×3 Depthwise Conv layer and two 3×3 AConv layers) separately. The contextual semantics obtained from the three convolutional layers are refined through a matrix multiplication operation; and, the refined feature map f^i ($i = 0$, 1, and 2), the output of the ISa module, is finally transferred to the WSaG-based PUN.

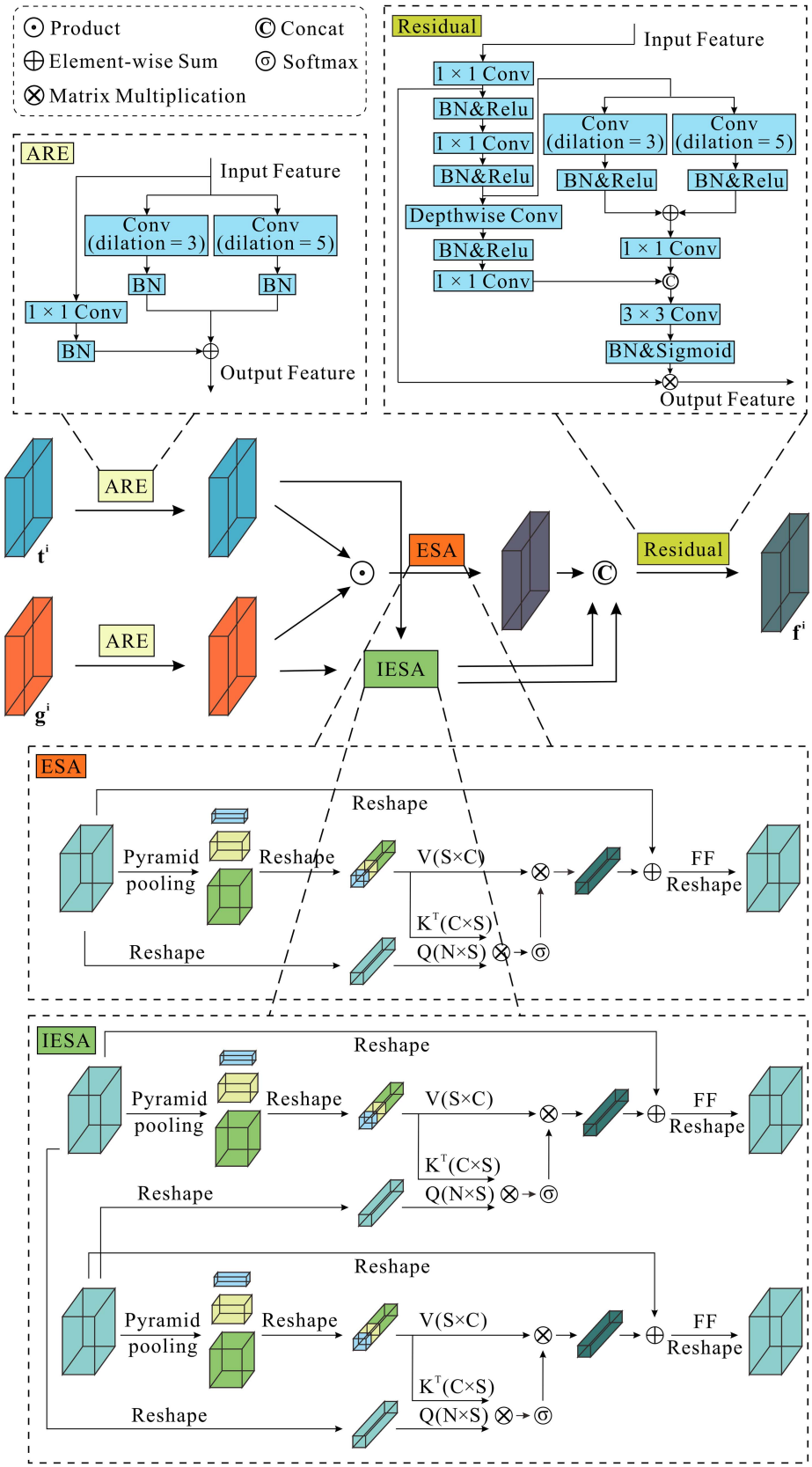

Fig. 2. Structure of the proposed ISa module.

The operation of the ISa module is expressed as follows:

$$\mathrm{ARE}(\mathrm{g}^i) = \mathrm{BN}(\mathrm{AConv}(\mathrm{g}^i)) + \mathrm{BN}(\mathrm{AConv}(\mathrm{g}^i)) + \mathrm{BN}(\mathrm{Conv}(\mathrm{g}^i)) \tag{7}$$

$$\mathrm{ARE}(\mathrm{t}^i) = \mathrm{BN}(\mathrm{AConv}(\mathrm{t}^i)) + \mathrm{BN}(\mathrm{AConv}(\mathrm{t}^i)) + \mathrm{BN}(\mathrm{Conv}(\mathrm{t}^i)) \tag{8}$$

$$\mathrm{ESA}(\mathrm{ARE}(\mathrm{t}^i) \odot \mathrm{ARE}(\mathrm{g}^i)) = \phi_0(\mathrm{concat}(\mathrm{head}^0, \ldots, \mathrm{head}^8)) \tag{9}$$

$$\mathrm{head}^j = \mathrm{Attention}(\phi_q^j(\mathrm{Q}), \phi_k^j(\mathrm{K}), \phi_v^j(\mathrm{V})) \tag{10}$$

$$\mathrm{Attention}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{softmax}(\frac{\mathrm{QK}^T}{\sqrt{\mathrm{d}_k}})\mathrm{V} \tag{11}$$

$$\mathrm{IESA}(\mathrm{t}^i, \mathrm{g}^i) = \phi_0(\mathrm{concat}(\mathrm{head}_1^0, \ldots, \mathrm{head}_1^8) \tag{12}$$

$$\mathrm{IESA}(\mathrm{g}^i, \mathrm{t}^i) = \phi_0(\mathrm{concat}(\mathrm{head}_2^0, \ldots, \mathrm{head}_2^8) \tag{13}$$

$$\mathrm{head}_1^j = \mathrm{Attention}(\phi_q^j(\mathrm{Q}_g), \phi_k^j(\mathrm{K}_t), \phi_v^j(\mathrm{V}_t)) \tag{14}$$

$$\mathrm{head}_2^j = \mathrm{Attention}(\phi_q^j(\mathrm{Q}_t), \phi_k^j(\mathrm{K}_g), \phi_v^j(\mathrm{V}_g)) \tag{15}$$

$$\mathrm{f}^i = \mathrm{Residual}(\mathrm{concat}(\hat{\mathrm{b}}^i, \hat{\mathrm{t}}^i, \hat{\mathrm{g}}^i)) \tag{16}$$

where ϕ_0 is the linear projection for the matrix of all attention heads (i.e., head^1, head^2, …, head^8); ϕ_q^j, ϕ_k^j, and ϕ_v^j are the linear projections for the matrix Q, matrix K, and matrix V of the jth head, respectively (j = 1, 2, …, 8); Attention() is the attention function; H_i, W_i, and C_i are the height, width, number of channels of the feature map $\mathrm{ARE}(\mathrm{g}^i)$ and $\mathrm{ARE}(\mathrm{t}^i)$ obtained by the ARE block, respectively (i = 0, 1, and 2); $\mathrm{Q} \in \mathrm{R}^{\mathrm{N}i \times \mathrm{C}i}$ ($\mathrm{N}_i = \mathrm{H}_i \times \mathrm{W}_i$) is the reshaped matrix of the input feature map g^i, t^i, and $\mathrm{ARE}(\mathrm{t}^i) \odot \mathrm{ARE}(\mathrm{g}^i)$($i$ = 0, 1, and 2), in which $\odot$ is the Hadamard product; $\mathrm{K} \in \mathrm{R}^{\mathrm{S} \times \mathrm{C}i}$ and $\mathrm{V} \in \mathrm{R}^{\mathrm{S} \times \mathrm{C}i}$ are matrices obtained from the pyramid pooling, reshaping, and concatenating (concat) operations (S = 1 × 1, 3 × 3, or 5 × 5); K^T is the transpose matrix of matrix K; softmax() is the softmax function; d_k is a scale factor that indicates the dimension of each attention head, the value of which is $\frac{\mathrm{C}_i}{8}$; $\hat{\mathrm{b}}^i$ is the feature map outputted from the ESA module, and the related inputs are t^i and g^i (i = 0, 1, and 2); $\hat{\mathrm{t}}^i$ is the feature map outputted from the IESA module, and the related input is the feature map t^i (i = 0, 1, and 2); and $\hat{\mathrm{g}}^i$ is the feature map outputted from the IESA module, and the related input is the feature map g^i (i = 0, 1, and 2).

C. WSaG-Based PUN

Inspired by the SEgmentation TRansformer model [57], the PUN is adopted as the decoder in the proposed TCNet. It is noted that convolutional layers and upsampling operations are alternately adopted in the PUN; as such, noisy predictions that might be induced by the one-step upscaling could be avoided. To capture both global and local contexts of multiple scales, a skip connection module called WSaG is advanced in this study and added to the PUN of the proposed TCNet. WSaG is created based on the Swin Transformer [27], with which the full flow of feature information between different scales is maximized.

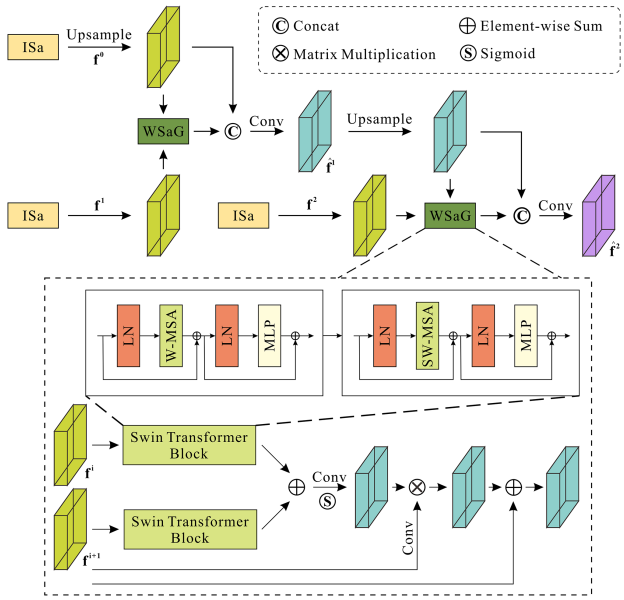

Fig. 3. Structure of the proposed WSaG-based PUN.

Within the Swin Transformer, the ordinary MSA is replaced by the window-based MSA (W-MSA) and shifted W-MSA. With the aid of these two self-attentions, self-attention within windows and cross-window connections can both be realized. The structure of the proposed WSaG-based PUN is shown in Fig. 3.

The procedures for the feature fusion with the ISa module are summarized as follows. The feature maps f^i and f^{i+1} (i = 0, 1) of different resolutions, obtained from the ISa module, are inputted to the module of WSaG, where the low-resolution feature map f^i is upsampled (Up) by a factor of 2 before the handling by the WSaG module. The output of the WSaG module is an attention map, based on which a matrix multiplication operation is conducted to refine the original high-resolution feature map f^{i+1} (i = 0, 1). Afterward, a channelwise concatenation operation and a convolution (Conv) operation are sequentially undertaken to fuse the original feature map f^i and the refined feature map f^{i+1} (i = 0, 1). Finally, feature maps $\hat{\mathrm{f}}^{i+1}$ (i = 1, 2) are generated by the WSaG-based PUN; and then, the feature map $\hat{\mathrm{f}}^2$ is processed by a simple head to generate a final segmentation map P^2.

The operation of the proposed WSaG-based PUN can be expressed as follows:

$$\hat{\mathrm{f}}^0 = \mathrm{f}^0 \tag{17}$$

$$\hat{\mathrm{f}}^{i+1} = \mathrm{Conv}([\mathrm{Up}(\hat{\mathrm{f}}^i), \mathrm{WSaG}(\mathrm{f}^{i+1}, \mathrm{Up}(\hat{\mathrm{f}}^i))]). \tag{18}$$

IV. Experiments on Three Public Datasets

In this section, experiments are carried out on three public datasets, including the Bijie Landslide Dataset, the WHU Building Dataset, and the Massachusetts Buildings Dataset, to depict the effectiveness and superiority of the proposed TCNet.

A. Introduction of the Three Public Datasets

1) Bijie Landslide Dataset: The Bijie Landslide Dataset consists of satellite optical images, shapefiles of landslides' boundaries, and digital elevation models. The dataset is constructed based on the data collected in an area in Bijie City, Guizhou Province, China [58], which covers an area of 26 853 km^2. The TripleSat satellite images taken from May to August 2018 were cropped to create 770 landslide images and 2003 nonlandslide images. The resolution of the digital elevation model is 2.0 m, compared to 0.8 m for the satellite optical images and shapefiles of landslide boundaries. In this study, 770 landslide images are studied for the performance test of the TCNet, among which, 462 images that are arbitrarily selected are taken as the training set, 154 images that are arbitrarily selected from the left 308 images are taken as the validation set, and the left 154 images are taken as the test set.

2) WHU Building Dataset: The WHU Building Dataset consists of satellite and aerial images [59]. There are 8189 aerial images of 512×512 pixels in this dataset, and the resolution of the aerial images is 0.3 m. Only aerial images are studied in this article. These aerial images cover an area of over 450 km^2 and 220 000 buildings in Christchurch, New Zealand. According to the rules provided in [59], 4736 images are taken as the training set, 1036 images are taken as the validation set, and 2416 images are taken as the test set. In the experiments conducted, the aerial images are preprocessed through cropping, each aerial image is divided into 4 smaller images of 256 × 256 pixels; and, the cropped images are then resized to larger images of 736 × 736 pixels. It is noted that the resized images, not the original aerial images, are adopted for the performance test of the TCNet.

3) Massachusetts Buildings Dataset: The Massachusetts Buildings Dataset consists of 151 aerial images collected in Boston, MA, USA; the size of each aerial image is 1500 × 1500 pixels and the related area is 2.25 km^2 (https://www.cs.toronto.edu/∼vmnih/data/). The resolution of the aerial images is 1.0 m. These aerial images include a variety of buildings, including residential, commercial, and industrial structures. According to the default rules, 137 images are taken as the training set, 4 images are taken as the validation set, and 10 images are taken as the test set. Similarly, the aerial images are preprocessed through cropping, each aerial image is divided into 36 smaller images of 250 × 250 pixels; and the cropped images are then resized to larger images of 768 × 768 pixels.

B. Experimental Settings and Evaluation Indexes

The full network in this study is trained end-to-end with the weighted IoU loss and the binary cross-entropy loss, denoted as L_{IoU} and L_{bce}, respectively. Note that weighted IoU loss is mainly proposed for measuring the similarity between the segmentation map, obtained by the simple head (Pred. Head), and the ground truth [60]; and the binary cross-entropy loss is mainly employed to measure the loss of boundary pixels [61]. To improve the gradient flow, deep supervision is adopted to supervise the output segmentation map of the Transformer branch and that of the first ISa module. On the basis of the computed L_{IoU} and L_{bce}, the overall loss function L_{total} can be calculated as follows:

$$L_{\mathrm{total}} = \alpha L(\mathrm{G}, \mathrm{Pred.Head}(\hat{\mathrm{f}}^2)) + \beta L(\mathrm{G}, \mathrm{Pred.Head}(\mathrm{f}^0)) + \gamma L(\mathrm{G}, \mathrm{Pred.Head}(\mathrm{t}^2)) \tag{19}$$

$$L = L_{\mathrm{IoU}} + L_{\mathrm{bce}} \tag{20}$$

where α, β, and γ are tunable hyperparameters, the values of which are taken 0.5, 0.3, and 0.2, respectively, in this study; L is the joint loss function of L_{IoU} and L_{bce}; G is the ground truth; and Pred. Head($\hat{\mathrm{f}}^2$), Pred. Head (f^0), and Pred. Head(t^2) are the segmentation maps obtained from the WSaG-based PUN, the Transformer branch, and the first ISa module, respectively.

The results of the pixel classification can be divided into true positive (TP) (i.e., the foreground is classified as foreground), false positive (FP) (i.e., the background is classified as foreground), true negative (TN) (i.e., the background is classified as background), and false negative (FN) (i.e., the foreground is classified as background). To quantitatively assess the performance of the trained TCNet in the experiments, five indexes, in terms of the IoU, Overall Accuracy (OA), Precision, Recall, and F1 score, are evaluated, and the mathematical formulations of these indexes are given as follows:

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP}} \tag{21}$$

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{22}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{23}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{24}$$

$$\mathrm{F1score} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}. \tag{25}$$

The training and validation sets are first imported to the platform of GPU-based Pytorch, and these data are trained for 30 epochs utilizing the proposed TCNet. During the training of the model, Gaussian Blur-, Solarization-, Grayscale-, and Random Horizontal Flip-based data augmentation operations are conducted. It is noted that the training of the TCNet model is executed on a desktop equipped with 96.0 GB RAM, an Intel(R) Xeon(R) W-2145 CPU running at 3.70 GHz, and an NVIDIA Quadro P5000 64-GB GPU. Further, the Adam optimizer is adopted in this study, and the learning rate is set up as 7e-5, whereas the batch size of the training is set up as 1. The outcome of this model training is an automatic image segmentation model that can detect and map the foreground in input images.

C. Performance Evaluation of the Proposed TCNet and Comparisons With Other Models

Listed in Table I are the evaluated performance indexes of the trained TCNet models based on these three public datasets. As shown in Table I, all the IoU values obtained by the TCNet are greater than 75.00%, indicating that the segmentation maps obtained from the trained TCNet models are highly consistent with the ground truth. In addition, the other four indexes (i.e.,

TABLE I
PERFORMANCE EVALUATION OF THE TCNET ON THE THREE PUBLIC DATASETS

Datasets	IoU	OA	Precision	Recall	F1
Bijie Landslide Dataset	0.7534	0.9720	0.8419	0.8920	0.8512
WHU Building Dataset	0.9116	0.9880	0.9515	0.9555	0.9395
Massachusetts Buildings Dataset	0.7621	0.9485	0.8517	0.8682	0.8429

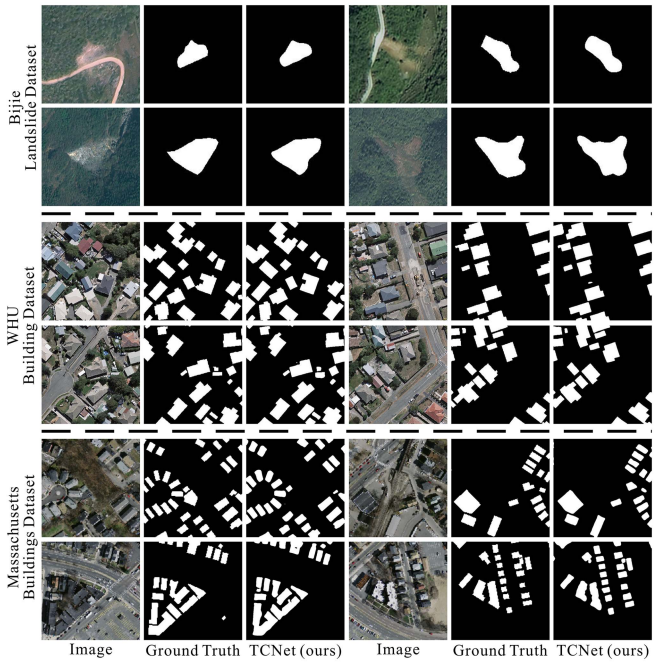

Fig. 4. Visualized results of the proposed TCNet on the three public datasets.

TABLE II
QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART METHODS ON THE BIJIE LANDSLIDE DATASET

Method	IoU	OA	Precision	Recall	F1
UNet [14]	0.5767	0.9503	0.7134	0.7361	0.6989
PSPNet [21]	0.5402	0.9396	0.7445	0.7201	0.6597
DeepLab V3+ [15]	0.6012	0.9503	0.7961	0.7604	0.7215
SegFormer [45]	0.6105	0.9538	0.7419	0.7862	0.7252
HRNetV2 [62]	0.6570	0.9614	0.7960	0.8018	0.7689
CLCFormer [50]	0.5858	0.9479	0.7415	0.7063	0.6792
SDSC-UNet [72]	0.7144	0.9679	0.8457	**0.9331**	0.8212
DSAT-Net [51]	0.7455	0.9701	**0.8749**	0.8391	0.8436
UNetFormer [5]	0.7303	0.9668	0.8421	0.8551	0.8318
TCNet (ours)	**0.7534**	**0.9720**	0.8419	0.8920	**0.8512**

The bold values indicate the highest values within the corresponding evaluation indices.

TABLE III
QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART METHODS ON THE WHU BUILDING DATASET

Method	IoU	OA	Precision	Recall	F1
UNet [14]	0.8551	0.9727	0.9186	0.9252	0.9219
DeepLab V3+ [15]	0.8578	0.9737	0.9345	0.9127	0.9235
BiSeNetv2 [64]	0.8651	-	0.9322	0.9225	-
SegNet [36]	0.8612	0.9741	0.9273	0.9235	0.9254
BOMSC-Net [69]	0.9015	0.9820	0.9514	0.9450	0.9480
MAP-Net [67]	0.9086	-	0.9562	0.9481	**0.9521**
MA-FCN [63]	0.9070	-	0.9520	0.9510	0.9515
RSR-Net [64]	0.8832	-	0.9492	0.9263	-
CLCFormer [50]	0.8420	0.9716	0.9541	0.8733	0.8954
SDSC-UNet [72]	0.9104	0.9827	0.9622	0.9521	0.9429
DSAT-Net [51]	0.9014	0.9778	0.9612	0.9446	0.9236
UNetFormer [5]	0.9018	0.9875	**0.9662**	0.9306	0.9304
TCNet (ours)	**0.9116**	**0.9880**	0.9515	**0.9555**	0.9395

The bold values indicate the highest values within the corresponding evaluation indices.

OA, Precision, Recall, and F1 score) obtained by the TCNet are all greater than 84.00%, which depicts the effectiveness of the proposed TCNet. Illustrated in Fig. 4 are the example segmentation maps obtained from the trained TCNet models, showing that the boundaries of landslides and buildings, arbitrarily selected in the datasets, can be precisely extracted.

To further demonstrate the effectiveness and superiority of the TCNet, the performance of the TCNet and that of some state-of-the-art models are compared. The models selected for the comparisons are UNet [14], PSPNet [21], SegNet [36], DeepLab V3+ [15], HRNetV2 [62], MA-FCN [63], BiSeNetv2 [64], SegFormer [45], RSR-Net [64], MANet [65], BANet [66], MAP-Net [67], BuildFormer [68], BOMSC-Net [69], DC-Swin [70], ASF-Net [71], CLCFormer [50], SDSC-UNet [72], DSAT-Net [51], and UNetFormer [5]. For ease of comparison, the models compared are retrained under an identical operating environment as that adopted by the TCNet.

The comparison is first conducted based on the Bijie Landslide Dataset. Listed in Table II are the performance indexes evaluated from the TCNet and those from the compared models (i.e., UNet, PSPNet, DeepLab V3+, SegFormer, HRNetV2, CLCFormer, SDSC-UNet, DSAT-Net, and UNetFormer). Table II shows that the IoU (75.34%) obtained from the TCNet is always much higher than those from the compared models. It is noted that there is no unified rule for the division of the training set, verification set, and test set in the Bijie Landslide Dataset. To ensure that the same training set, verification set, and test set are adopted in this comparison, the performance indexes of the compared models are derived in this study, not from the existing literature. The data in Table II also show that the performance indexes of OA and F1 score obtained from the TCNet are always higher than those from the models compared, and the performance indexes of precision and recall obtained by TCNet are close to the highest performance indexes of the models compared, implying that the proposed TCNet results in fewer pixel errors in the landslide classification. Illustrated in Fig. 5 are the example segmentation maps obtained from the TCNet model and those from the compared models, showing that the landslide pixels could be more effectively classified by the proposed TCNet.

Next, the comparison is conducted based on the WHU Building Dataset. Listed in Table III are the performance indexes evaluated from the TCNet and those from the compared models (i.e., UNet, DeepLab V3+, BiSeNetv2, SegNet, BOMSC-Net, MAP-Net, MA-FCN, RSR-Net, CLCFormer, SDSC-UNet, DSAT-Net, and UNetFormer). The data in Table III depict that the IoU (91.16%), OA (98.80%), and Recall (95.55%) obtained from the TCNet are much higher than those from the compared models, whereas the performance indexes of Precision and F1

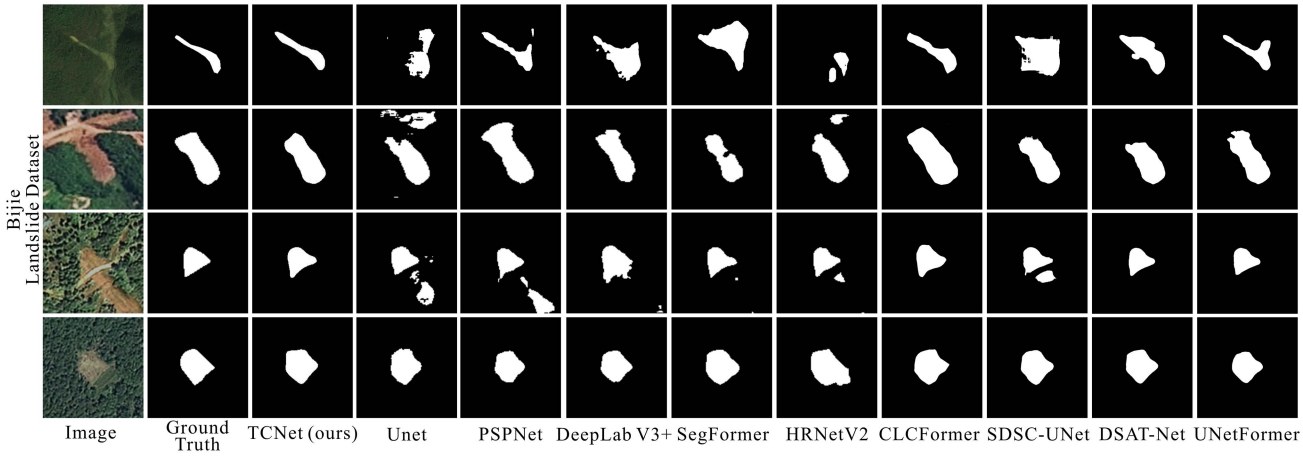

Fig. 5. Recognition results of the proposed TCNet and other models on an arbitrarily selected photograph in the Bijie Landslide Dataset.

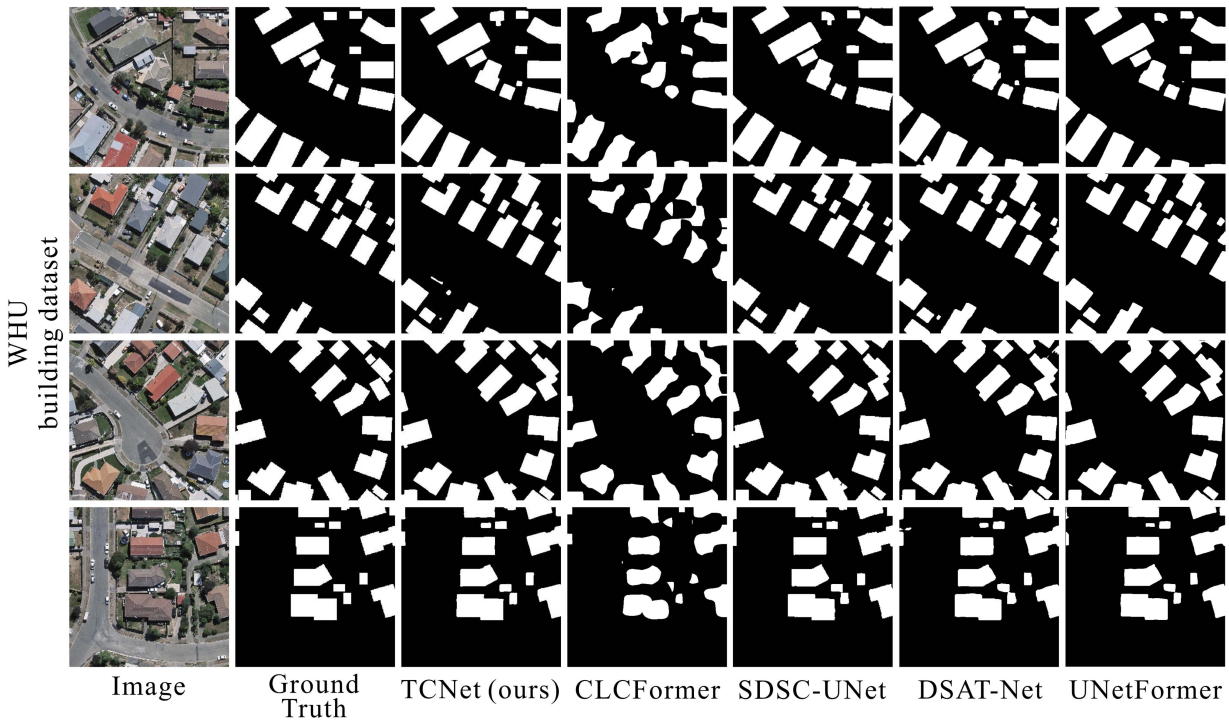

Fig. 6. Recognition results of the proposed TCNet and other models on an arbitrarily selected photograph in the WHU Building Dataset.

score obtained from the TCNet are close to the highest performance indexes of the models compared. From there, the overall performance of the TCNet could be better than the compared models. Depicted in Fig. 6 are the example segmentation maps obtained from the TCNet model and those from the model compared, showing that the building pixels can be more effectively classified by the TCNet.

Finally, the comparison is undertaken based on the Massachusetts Buildings Dataset. Listed in Table IV are the performance indexes evaluated from the TCNet and those from the compared models (i.e., UNet, DeepLab V3+, MANet, BANet, DC-Swin, BuildFormer, ASF-Net, BOMSC-Net, CLCFormer, SDSC-UNet, DSAT-Net, and UNetFormer). The data in Table IV depict that the IoU (76.21%), OA (94.85%), and Recall (86.82%) obtained from the TCNet are much higher than those from the compared models, whereas the performance indexes of Precision and F1 score obtained from the TCNet are close to the highest performance indexes of the models compared. From there, the overall performance of the TCNet is better than the compared models. Depicted in Fig. 7 are the example segmentation maps obtained from the TCNet model and those from the model compared, showing that the building pixels can be more effectively classified by the TCNet.

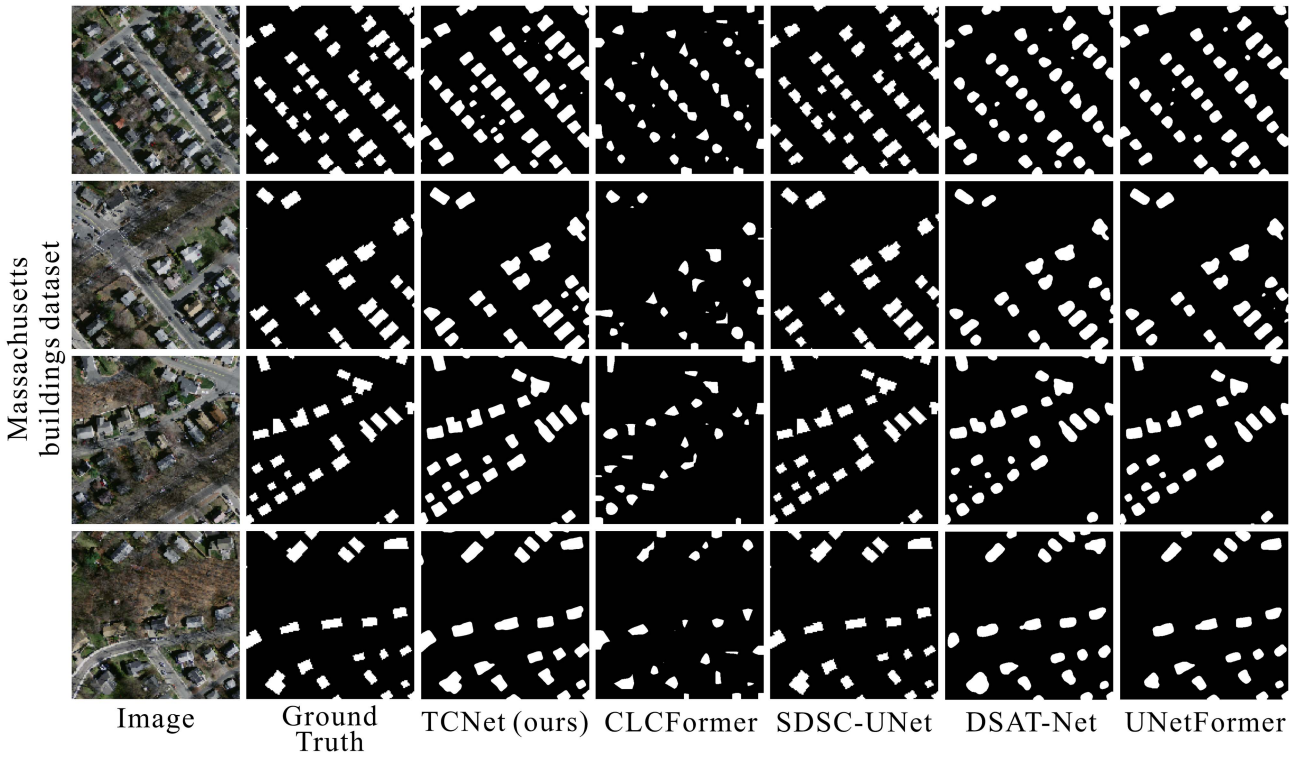

Fig. 7. Recognition results of the proposed TCNet and other models on an arbitrarily selected photograph in the Massachusetts Buildings Dataset.

TABLE IV
QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART METHODS ON THE MASSACHUSETTS BUILDINGS DATASET

Method	IoU	OA	Precision	Recall	F1
UNet [14]	0.6761	-	0.7913	0.8229	0.8068
DeepLab V3+ [15]	0.6923	-	0.8473	0.7910	0.8182
MANet [65]	0.7076	-	0.8200	0.8377	0.8288
BANet [66]	0.7220	-	0.8307	0.8466	0.8386
DC-Swin [70]	0.7259	-	0.8307	0.8519	0.8412
BuildFormer [68]	0.7574	-	0.8752	0.8490	0.8619
ASF-Net [71]	0.7420	-	-	-	**0.9460**
BOMSC-Net [69]	0.7471	0.9471	0.8664	0.8368	0.8513
CLCFormer [50]	0.4911	0.8781	0.8787	0.5187	0.6176
SDSC-UNet [72]	0.7550	0.9412	0.8856	0.8364	0.8419
DSAT-Net [51]	0.7251	0.9239	0.8476	0.8420	0.8055
UNetFormer [5]	0.7271	0.9349	**0.8945**	0.7956	0.8238
TCNet (ours)	**0.7621**	**0.9485**	0.8517	**0.8682**	0.8429

The bold values indicate the highest values within the corresponding evaluation indices.

TABLE V
ABLATION TEST RESULTS OF EACH BRANCH STRUCTURE AND KEY MODULE

Method	F1	IoU
TCNet without CNN branch	0.9343	0.9038
TCNet with CNN branch	0.9395	0.9116
TCNet without Transformer branch	0.9214	0.8899
TCNet with Transformer branch	0.9395	0.9116
TCNet without ISa	0.9316	0.8990
TCNet with ISa	0.9395	0.9116
TCNet without WSaG	0.9311	0.9002
TCNet with WSaG	0.9395	0.9116

D. Ablation Experiments

To illustrate the significance of each branch structure and key module of the proposed TCNet, ablation experiments are conducted based on the WHU Building Dataset.

1) Significance of the CNN Branch: The CNN branch, in the proposed TCNet, is mainly adopted to encode the local context information. To test the significance of the CNN branch, the CNN branch is removed from the TCNet in this ablation test, and the test results are listed in Table V. Here, the removement of the CNN branch decreases the F1 score and IoU by 0.52% and 0.78%, respectively. Thus, the significance of the CNN branch can be demonstrated.

2) Significance of the Transformer Branch: The Transformer branch, in the TCNet, is mainly adopted to encode the global context information. To test the significance of the Transformer branch, the Transformer branch is removed from the TCNet in this ablation test; and the test results are listed in Table V. Here, the removement of the Transformer branch decreases the F1 score and IoU by 1.81% and 2.71%, respectively. Thus, the significance of the Transformer branch is depicted.

3) Significance of the ISa Module: The ISa module, in the TCNet, is mainly adopted to encode the information of various scales. To test the significance of the ISa module, the ISa module is removed from the TCNet in this ablation test, and the test results are listed in Table V. Here, the removement of the ISa module decreases the F1 score and IoU by 2.69% and

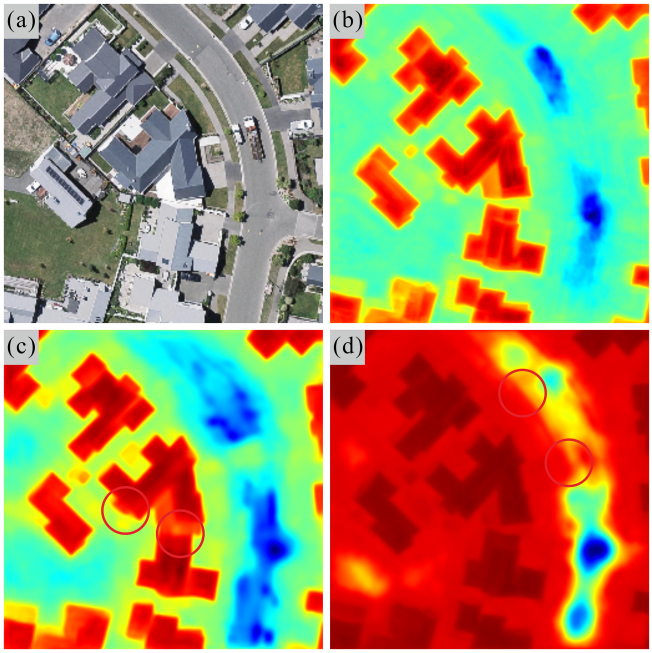

Fig. 8. Feature visualization results obtained from an arbitrarily selected photograph in the WHU Building Dataset. (a) Original photograph. (b) Recognition result with both local information and global context information. (c) Recognition result with global context information. (d) Recognition result with local information.

4.26%, respectively. Thus, the significance of the ISa module is demonstrated.

4) Significance of the WSaG Module: The WSaG module, in the TCNet, is mainly adopted to maximize the flow of feature information between different scales. To test the significance of the WSaG module, the WSaG module is removed from the TCNet in this ablation test, and the test results are listed in Table V. Here, the removement of the WSaG module decreases the F1 score and IoU by 1.24% and 1.64%, respectively. Thus, the significance of the WSaG module is depicted.

V. Discussion

To further illustrate the significance of the parallel-in-branch architecture adopted in the TCNet, feature visualization results are studied; and, comparative studies on the efficiency of the model are conducted.

A. Two Branches Analysis

To study the influences of the CNN branch and Transformer branch on the feature visualization result in the segmentation, these two branches are removed from the TCNet separately, similar to the ablation experiments presented above, and some of the feature visualization results on an arbitrarily selected photograph in the WHU Building Dataset are shown in Fig. 8.

The feature visualization results, shown in Fig. 8, indicate that accurate recognition of the boundaries of the building could be challenging when the local information is removed [see the red circles in Fig. 8(c)], whereas irrelevant backgrounds such as roads might be recognized as buildings when the global context information is removed [see the red circles in Fig. 8(d)]. When both local information and global context information are combined, accurate recognition of the building boundaries can be realized [see Fig. 8(b)]. It can be seen that the local modeling capabilities of the CNN help to recognize small detailed areas in the segmentation, whereas the global modeling capabilities of the Transformer help to establish global dependencies between the building and the background. Thus, the significance of each branch of the TCNet (i.e., CNN branch and Transformer branch) is further demonstrated.

TABLE VI
Comparison of Model Parameters and FLOPs of the Proposed TCNet and Some Existing Models

Method	Parameters (M)	FLOPs (G)
UNet [14]	24.891	255.836
PSPNet [21]	2.376	3.016
HRNetV2 [62]	29.538	45.463
DeepLab V3+ [15]	5.813	26.434
SegFormer [45]	3.715	66.739
UNetFormer [5]	11.70	11.738
CLCFormer [50]	54.110	47.560
SDSC-UNet [72]	21.320	29.820
DSAT-Net [51]	48.371	57.750
TCNet (ours)	34.875	91.875

B. Model Efficiency Analysis

It should be noted that the computational cost of a semantic segmentation model is often assessed by the amount of model parameters and that of floating point operations (FLOPs). Table VI lists the amount of parameters and FLOPs of the proposed TCNet and those of some existing models, in terms of the UNet [14], PSPNet [21], HRNetV2 [62], DeepLab V3+ [15], SegFormer [45], UNetFormer [5], CLCFormer [50], SDSC-UNet [72], and DSAT-Net [51]. Similarly, these models are trained under an identical operating environment. The data in Table VI show that the TCNet does not incur excessive memory and computational overhead. For example, the TCNet yields a similar amount of parameters and FLOPs, in comparison to the UNetFormer, DSAT-Ne, and CLCFormer. However, the proposed TCNet could improve the segmentation accuracy significantly (see Tables II–IV and Figs. 5–7).

VI. Conclusion

A novel network called TCNet was proposed in this study for the semantic segmentation of remote sensing images. A parallel-in-branch architecture of the Transformer and the CNN is adopted in the TCNet; as such, both global context and low-level spatial details can be captured. In addition, an ISa module was developed and adopted in the TCNet to fuse the multilevel features extracted from the two branches; and, to bridge the semantic gap between regions, a WSaG-based PUN was developed and adopted in the TCNet. To demonstrate the effectiveness and versatility of the TCNet, experiments on three public datasets, including the Bijie Landslide Dataset, the WHU

Building Dataset, and the Massachusetts Buildings Dataset, were conducted.

The results of the experiments showed that both foreground and background could be precisely extracted by the trained TCNet model, as indicated by the relatively high values of IoU, OA, Precision, Recall, and F1 score. Further, comparisons between the proposed TCNet and some state-of-the-art models were undertaken; and, the results of the comparison showed the superiority of the proposed TCNet. For example, compared to the existing models, the proposed TCNet almost always yields higher performance indexes. Meanwhile, the significance of each branch structure and key module of the proposed TCNet was verified through ablation experiments based on the WHU Building Dataset. Finally, the significance of the parallel-in-branch architecture adopted and the efficiency of the TCNet were discussed. Note that although the proposed TCNet was deemed effective in the analyses conducted, the following limitations warrant further investigation: the TCNet was only applied to semantic segmentation of building and landslide in remote sensing images of urban and mountain scenes, whereas other tasks such as road segmentation and plot segmentation have not been tested.

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, doi: 10.1007/s13735-017-0141-z.

[2] Z. Cheng et al., "UAV photogrammetry-based remote sensing and preliminary assessment of the behavior of a landslide in Guizhou, China," *Eng. Geol.*, vol. 289, Aug. 2021, Art. no. 106172, doi: 10.1016/j.enggeo.2021.106172.

[3] T. Ren, W. Gong, L. Gao, F. Zhao, and Z. Cheng, "An interpretation approach of ascending–descending SAR data for landslide identification," *Remote Sens.*, vol. 14, no. 5, Mar. 2022, Art. no. 1299, doi: 10.3390/rs14051299.

[4] F. Zhao, W. Gong, H. Tang, S. P. Pudasaini, T. Ren, and Z. Cheng, "An integrated approach for risk assessment of land subsidence in Xi'an, China using optical and radar satellite images," *Eng. Geol.*, vol. 314, Mar. 2023, Art. no. 106983, doi: 10.1016/j.enggeo.2022.106983.

[5] L. Wang et al., "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 190, pp. 196–214, 2022, doi: 10.1016/j.isprsjprs.2022.06.008.

[6] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding UNet for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 4408715, doi: 10.1109/TGRS.2022.3144165.

[7] W. Wang, Y. Yang, J. Li, Y. Hu, Y. Luo, and X. Wang, "Woodland labeling in Chenzhou, China, via deep learning approach," *Int. J. Comput. Intell. Syst.*, vol. 13, no. 1, pp. 1393–1403, Sep. 2020, doi: 10.2991/ijcis.d.200910.001.

[8] P. Dey, S. K. Chaulya, and S. Kumar, "Hybrid CNN-LSTM and IoT-based coal mine hazards monitoring and prediction system," *Process Saf. Environ. Protection*, vol. 152, pp. 249–263, Aug. 2021, doi: 10.1016/j.psep.2021.06.005.

[9] Z. Zhang, G. Vosselman, M. Gerke, D. Tuia, and M. Y. Yang, "Change detection between multimodal remote sensing data using Siamese CNN," 2018, doi: 550/arXiv.1807.09562.

[10] X. Gao, T. Chen, R. Niu, and A. Plaza, "Recognition and mapping of landslide using a fully convolutional DenseNet and influencing factors," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7881–7894, Aug. 2021, doi: 10.1109/JSTARS.2021.3101203.

[11] 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, Jan. 2022, Art. no. 4408820, doi: 10.1109/TGRS.2022.3144894.

[12] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440, doi: 10.1109/TPAMI.2016.2572683.

[13] 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, doi: 10.1007/s10462-020-09854-1.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[15] 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. 833–851, doi: 10.48550/arXiv.1802.02611.

[16] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, doi: 10.48550/arXiv.1706.05587.

[17] P. Song, J. Li, Z. An, H. Fan, and L. Fan, "CTMFNet: CNN and transformer multiscale fusion network of remote sensing urban scene imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Dec. 2023, Art. no. 5900314, doi: 10.1109/TGRS.2022.3232143.

[18] 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.

[19] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

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

[21] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6230–6239, doi: 10.1109/CVPR.2017.660.

[22] P. Wang et al., "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 1451–1460.

[23] L. Mou, Y. Hua, and X. X. Zhu, "Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7557–7569, Nov. 2020.

[24] 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, doi: 10.48550/arXiv.2010.11929.

[25] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, vol. 12346, pp. 213–229, doi: 10.1007/978-3-030-58452-8_13.

[26] C. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Mar. 2021, pp. 1–12.

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

[28] R. Azad, M. Heidari, Y. Wu, and D. Merhof, "Contextual attention network: Transformer meets U-Net," in *Proc. Mach. Learn. Med. Imag.*, 2022, pp. 377–386, doi: 10.1007/978-3-031-21014-3_39.

[29] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022.

[30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[31] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 60–77, Nov. 2018, doi: 10.1016/j.isprsjprs.2018.04.014.

[32] H. Zhang, Y. Liao, H. Yang, G. Yang, and L. Zhang, "A local–global dual-stream network for building extraction from very-high-resolution remote sensing images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 3, pp. 1269–1283, Mar. 2020.

[33] H. Hosseinpour, F. Samadzadegan, and F. D. Javan, "CMGFNet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 184, pp. 96–115, Feb. 2022, doi: 10.1016/j.isprsjprs.2021.12.007.

[34] D. Eigen and R. Fergus, "Predicting depth surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2650–2658.
[35] Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew, "A review of semantic segmentation using deep neural networks," *Int. J. Multimedia Inf. Retrieval*, vol. 7, no. 2, pp. 87–93, 2018, doi: 10.1007/s13735-017-0141-z.
[36] 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.
[37] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1520–1528.
[38] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets atrous convolution and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[39] R. Li, S. Zheng, C. Duan, J. Su, and C. Zhang, "Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Mar. 2022, Art. no. 8009205, doi: 10.1109/LGRS.2021.3063381.
[40] R. Li, C. Duan, S. Zheng, C. Zhang, and P. M. Atkinson, "MACU-Net for semantic segmentation of fine-resolution remotely sensed images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jan. 2022, Art. no. 8007205, doi: 10.1109/LGRS.2021.3052886.
[41] R. Li, S. Zheng, C. Zhang, C. Duan, L. Wang, and P. M. Atkinson, "ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 181, pp. 84–98, Nov. 2021, doi: 10.1016/j.isprsjprs.2021.09.005.
[42] M. Y. Yang, S. Kumaar, Y. Lyu, and F. Nex, "Real-time semantic segmentation with context aggregation network," *ISPRS J. Photogramm. Remote Sens.*, vol. 178, pp. 124–134, Aug. 2021.
[43] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[44] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 7262–7272.
[45] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 12077–12090.
[46] S. Xiao, K. Shang, K. Lin, Q. Wu, H. Gu, and Z. Zhang, "Pavement crack detection with hybrid-window attentive vision transformers," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 116, Dec. 2022, Art. no. 103172, doi: 10.1016/j.jag.2022.103172.
[47] H. Cao et al., "Swin-Unet: Unet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 205–218, doi: 10.1007/978-3-031-25066-8_9.
[48] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, doi: 10.48550/arXiv.2102.04306.
[49] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing transformers and CNNs for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2021, pp. 1–11, doi: 10.1007/978-3-030-87193-2_2.
[50] J. Long, M. Li, and X. Wang, "Integrating spatial details with long-range contexts for semantic segmentation of very high-resolution remote-sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Mar. 2023, Art. no. 2501605, doi: 10.1109/LGRS.2023.3262586.
[51] R. Zhang, Z. Wan, Q. Zhang, and G. Zhang, "DSAT-Net: Dual spatial attention transformer for building extraction from aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Aug. 2023, Art. no. 6008405, doi: 10.1109/LGRS.2023.3304377.
[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[53] H. Touvron et al., "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 10347–10357, doi: 10.48550/arXiv.2012.12877.
[54] R. Zhang et al., "Lesion-aware dynamic kernel for polyp segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, Sep. 2022, pp. 99–109, doi: 10.1007/978-3-031-16437-8_10.
[55] D. P. Fan et al., "PraNet: Parallel reverse attention network for polyp segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2020, pp. 263–273, doi: 10.1007/978-3-030-59725-2_26.
[56] O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," 2018, doi: 10.48550/arXiv.1804.03999.
[57] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6881–6890.
[58] S. Ji, D. Yu, C. Shen, W. Li, and Q. Xu, "Landslide detection from an open satellite imagery and digital elevation model dataset using attention boosted convolutional neural networks," *Landslides*, vol. 17, no. 6, pp. 1337–1352, Jun. 2020, doi: 10.1007/s10346-020-01353-2.
[59] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.
[60] G. Máttyus, W. Luo, and R. Urtasun, "DeepRoadMapper: Extracting road topology from aerial images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3438–3446.
[61] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7479–7489.
[62] K. Sun et al., "High-resolution representations for labeling pixels and regions," 2019, *arXiv: 1904.04514*.
[63] S. Wei, S. Ji, and M. Lu, "Toward automatic building footprint delineation from aerial images using CNN and regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2178–2189, Mar. 2020.
[64] H. Huang, Y. Chen, and R. Wang, "A lightweight network for building extraction from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2022, Art. no. 5614812, doi: 10.1109/TGRS.2021.3131331.
[65] R. Li et al., "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2022, Art. no. 5607713, doi: 10.1109/TGRS.2021.3093977.
[66] L. Wang, R. Li, D. Wang, C. Duan, T. Wang, and X. Meng, "Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images," *Remote Sens.*, vol. 13, no. 16, Aug. 2021, Art. no. 3065, doi: 10.3390/rs13163065.
[67] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, "MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6169–6181, Jul. 2021.
[68] L. Wang, S. Fang, X. Meng, and R. Li, "Building extraction with vision transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jun. 2022, Art. no. 5625711, doi: 10.1109/TGRS.2022.3186634.
[69] Y. Zhou et al., "BOMSC-Net: Boundary optimization and multi-scale context awareness based building extraction from high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 2022, Art. no. 5618617, doi: 10.1109/TGRS.2022.3152575.
[70] 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, Jan. 2022, Art. no. 6506105, doi: 10.1109/LGRS.2022.3143368.
[71] J. Chen, Y. Jiang, L. Luo, and W. Gong, "ASF-Net: Adaptive screening feature network for building footprint extraction from remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 4706413, doi: 10.1109/TGRS.2022.3165204.
[72] R. Zhang, Q. Zhang, and G. Zhang, "SDSC-UNet: Dual skip connection ViT-based U-shaped model for building extraction," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Apr. 2023, Art. no. 6005005, doi: 10.1109/LGRS.2023.3270303.

Xuyang Xiang received the B.S. degree in geological engineering from the College of Earth Sciences, Guilin University of Technology, Guilin, China, in 2021. He is currently working toward the M.S. degree in geological engineering with the Faculty of Engineering, China University of Geosciences, Wuhan, China.

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

Wenping Gong received the B.S. degree in civil engineering from Tongji University, Shanghai, China, in 2011, and the Ph.D. degree in civil engineering from Clemson University, Clemson, SC, USA, in 2014.

He is currently a Full Professor with the Faculty of Engineering, China University of Geosciences, Wuhan, China. His research interests include engineering geology, geohazards, risk and reliability, uncertainty modeling, and remote sensing.

Dr. Gong is the Editor-in-Chief for *Engineering Geology* (Elsevier).

Shuailong Li received the B.S. degree in civil engineering from the School of Civil Engineering, Zhengzhou University, Zhengzhou, China, in 2021. He is currently working toward the M.S. degree in civil engineering with the Faculty of Engineering, China University of Geosciences, Wuhan, China.

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

Jun Chen (Member, IEEE) received the B.S. degree in electronic and information engineering and M.S. degree in communication and information system from China University of Geosciences, Wuhan, China, in 2002 and 2004, respectively, and the Ph.D. degree in communication and information system from Huazhong University of Technology, Wuhan, in 2014.

From 2004 to 2008, she was an Assistant Professor with China University of Geosciences, where she is currently an Associate Professor with the School of Automation. Her research interests include computer vision, pattern recognition, geoscience, and remote sensing.

Tianhe Ren received the B.E. and M.S. degrees in civil engineering, in 2019 and 2022, respectively, from the Faculty of Engineering, China University of Geosciences, Wuhan, China, where he is currently working toward the Ph.D. degree in geological engineering.

His research interests include remote sensing image processing and geoscience.