DWin-HRFormer: A High-Resolution Transformer Model With Directional Windows for Semantic Segmentation of Urban Construction Land

Zhen Zhang, Xin Huang, *Senior Member, IEEE*, and Jiayi Li, *Senior Member, IEEE*

***Abstract*— In this article, a deep neural network for semantic segmentation of high-resolution remote sensing images is proposed for urban construction land classification. The network follows a high-resolution network (HRNet) architecture. Specifically, a directional self-attention on the paths of different resolutions is proposed, aiming to correct the directional bias caused by the attention of strip windows during the model learning, while also reducing the computational complexity, and allowing the model to improve both the accuracy and the speed. At the end of the network, a distributed alignment module with spatial information is constructed to train additional learnable parameters, to adjust the biased decision boundaries through a two-stage learning strategy, and alleviate the problem of accuracy degradation due to the unbalanced training data. We tested the proposed method and compared it with the current state-of-the-art (SOTA) semantic segmentation methods on the Luojia-fine-grained land cover (FGLC) dataset and the Wuhan Dense Labeling Dataset (WHDLD), and the proposed one obtained the best performance. We also verified the effectiveness of each component of the network through ablation experiments. The code and model will be available at https://github.com/Zhzhyd/DWin-HRFormer.**

***Index Terms*— Deep learning, remote sensing imagery, semantic segmentation, transformer, urban construction land.**

I. Introduction

With the development of aerospace and sensor technology, researchers can now quickly obtain high-quality and high-resolution remote sensing images over large areas [1], [2]. These images contain both ecological information and the footprints of human activities, such as building construction, farming, and so on. The potential information from these images provides data support for numerous applications, such as urban planning [3], [4], [5], [6], agricultural production [7], [8], and so on, among which urban construction land monitoring is a typical remote sensing application, especially in rapidly urbanizing areas, e.g., many Chinese cities [9], [10], [11].

It focuses on the intensity of the development and the use of urban construction land, mainly including monitoring of three land use categories within the urban construction sites, i.e., "buildings," "roads," and "bare soil." In the process of urban construction in China, some cities and their suburban areas have been designated by local governments as urban construction areas. Some of the projects in these areas (or parcels) are progressing well and can be completed on time. However, some projects are on hold or proceeding slowly due to a certain of complicated factors. Therefore, to monitor the progress of these urban construction projects, local governments began to implement project progress monitoring regularly (e.g., quarterly), by detecting buildings, roads, and bare soil in these designated urban construction areas to determine their construction status. In terms of this application requirement, this article aims to study the semantic segmentation of urban construction land. Specifically, according to the requirements of China's natural resources department, urban construction land should be classified into three categories, i.e., "buildings," "roads," and "bare soil." Other categories are considered nonconstruction land.

It is noteworthy that there is usually a severe imbalance between these four land-use types, as shown in Fig. 1. At the same time, urban land monitoring requires high-frequency and rapid observation of the land use, thus requiring high monitoring efficiency. Semantic segmentation, at its core, is the assignment of a semantic label to each pixel in an image. The rapid development of semantic segmentation algorithms makes it a possible solution for urban construction land monitoring.

In recent years, the deep convolutional neural networks (CNNs) have broken the accuracy bottleneck of artificial object detection. As a result, CNNs have now become the primary method for semantic segmentation. Many powerful CNN-based backbone networks, such as very deep convolutional network (VGGNet) [12], ResNet [13], and so on, have

Manuscript received 21 July 2022; revised 19 November 2022; accepted 27 January 2023. Date of publication 1 February 2023; date of current version 13 February 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 41971295 and Grant 42271328, in part by the Special Fund of Hubei Luojia Laboratory under Grant 220100031, and in part by the Wuhan 2022 Dawning under Project 2022010801020123. The work of Xin Huang was supported in part by the National Program for Support of Top-Notch Young Professionals, in part by the China National Science Fund for Excellent Young Scholars, and in part by the New Century Excellent Talents in University from the Ministry of Education of China. *(Corresponding author: Xin Huang.)*

Zhen Zhang is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: zhenzhang30134@whu.edu.cn).

Xin Huang is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China, also with the Hubei Luojia Laboratory, Wuhan 430079, China, and also with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: xhuang@whu.edu.cn).

Jiayi Li is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China, and also with the Hubei Luojia Laboratory, Wuhan 430079, China (e-mail: zjjercia@whu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2023.3241366

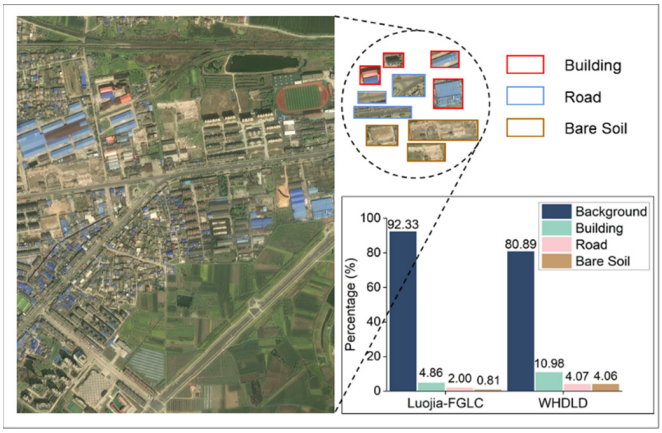


Fig. 1. Land categories of interest in urban construction land monitoring. The red boxes in the circles indicate buildings, the blue boxes indicate roads, and the brown boxes indicate SUC. The bar chart at the bottom right shows the imbalance in the proportion of the four land categories.

been designed. The subsequent development of networks, such as fully convolutional networks (FCNs) [14], feature pyramid networks (FPNs) [15], UNet [16], SegNet [17], multi-spectral, multi-view, and multi-task deep network (M^3Net) [18], and DeepLab V3+ [19], has made the encoder–decoder architecture a popular strategy for semantic segmentation networks. Furthermore, the networks, such as high-resolution network (HRNet) [20] and multiple attending path neural network (MAP-Net) [21], also demonstrate the great potential of HRNet architectures in dealing with multiscale image scenes. In addition, a large number of related studies, including the interaction of local and global information [22], [23], attention mechanisms [24], [25], [26], and multiscale feature representation [27], [28], have further improved the performance of the networks. However, CNN-based network structures have specific induction deviations, such as localization and translational invariance, which limits its ability to extract long-range spatial dependencies, leading to a reduction in the model performance [29].

The transformer model proposed by Vaswani et al. [30] can learn features through a self-attention mechanism, and it has been effectively applied to natural language processing (NLP), achieving excellent performances [31], [32]. Inspired by the success of the transformer model in NLP, researchers have applied the transformer structure to computer vision tasks. For example, Chen et al. [33] trained a sequence transformer with a comparable performance to CNNs in image classification tasks. The vision transformer (ViT) model [34], which is a pure transformer model, outperformed CNN-based models in terms of accuracy after pretraining on large-scale datasets. This demonstrates the great potential of transformer for computer vision tasks. Subsequently, transformer models have been used for various computer vision tasks, including object detection [35], [36], semantic segmentation [37], image processing [38], [39], and video understanding [40], [41].

The segmentation transformer (SETR) [37] and Segmenter [42] models were the first transformer model applied to the semantic segmentation task. Despite the powerful feature extraction capabilities of these models, they use the original image as input, which leads to a computational complexity that is proportional to the square of the image size, and a lot of computational resources and memory consumption are required. This issue can be addressed by restricting the attention area of each token from a global to a local scale (e.g., a window) and by moving windows for information interaction between them [43], [44]. However, this approach expands the receptive field slowly and requires the stacking of more transformer blocks to achieve global attention. Notably, a sufficiently large receptive field is crucial for semantic segmentation. Ho et al. [45] and Dong et al. [46], respectively, used an axial attention mechanism and a cross-shaped window self-attention mechanism to achieve global attention and effectively improved the efficiency. However, it should be noted that ignoring the directional information of natural images can lead to degradation of the model accuracy.

To fully extract the features of the ground objects in all directions and make the model have a large receptive field, we propose a directional window self-attention (DWSA) mechanism, which is further embedded into an HRNet architecture, for urban construction land monitoring. Meanwhile, a distribution alignment (DA) module with consideration of spatial contextual information (DASCI) is proposed to cope with the imbalance between the land categories. Compared with the existing two-stage imbalance learning methods [47], [48], [49], the proposed DASCI module can exploit the spatial contextual information more effectively. In summary, the main contributions of this article are as follows.

1) We propose a directional self-attention mechanism, which uses a series of windows with different directions to perform self-attention computation in parallel, in order to overcome the problem that the network tends to ignore the feature orientation bias during the feature representation. In addition, this method can promote the computational efficiency and at the same time improve the classification accuracy.

2) A DASCI is designed to deal with the problem of classifying the imbalanced land categories (i.e., with a significantly smaller sample size compared with other ones).

3) We propose a new network, namely DWin-HRFormer, by synthesizing the DWSA mechanism and DASCI module within an HRNet architecture for the semantic segmentation of urban construction regions.

II. Related Work

A. Deep High-Resolution Structure Networks

The semantic segmentation networks can be divided into encoder–decoder structures, high-resolution structures, and other ones. Since the transformer model demonstrated its excellent performance, a lot of new models have been developed. A natural strategy is to replace the convolutional blocks with transformers or embed them into the existing network structure. For example, Cao et al. [50] proposed a U-shaped network based on the shifted window of the Swin transformer [43]; Chen et al. [51] introduced a transformer after the CNN-based encoder to extract global contextual information; Wang et al. [52] used an attention mechanism to refine the skip connection and eliminate the ambiguity between the features in different stages of the encoder

and decoder; He et al. [53] constructed a dual-encoder U-shaped network using transformer-based and CNN-based encoders in parallel. All these methods belong to the encoder–decoder architecture. On the other hand, the HRFormer [54] is based on a high-resolution structure and uses a self-attention mechanism to replace the convolution operation, in order to maintain a high-resolution representation of the network and combine the advantages of the transformer model and HRNet [20]. The high-resolution architecture has achieved great success in both pose estimation [20], [54] and semantic segmentation [20], [54] tasks. The high-resolution representation has advantages in both semantic and spatial feature expression and can cope well with the multiscale problem of remote sensing imagery.

B. Efficient Self-Attention Mechanisms

In NLP, many effective attention mechanisms have been devised to improve the efficiency of transformer processing of long sequences [55], [56], [57], [58], [59], [60]. However, these mechanisms are often difficult to apply to visual tasks. The computational complexity of a model using the original self-attention mechanism [37], [42], [61], [62], [63] is quadratic to the image size, as shown in Fig. 2(a). The efficiency of the self-attention mechanism is particularly important for high-resolution remote sensing images. To improve the efficiency of attention computation, it is reasonable to apply attention to a local window and move the window in a specific direction, so that the attention computation covers the whole feature map. Some researchers [43], [44], [50], [53] have used a local self-attention mechanism and performed the information interaction of different local windows by shifted local self-attention (SLSA), as shown in Fig. 2(b). Ho et al. [45] achieved global attention by applying a local window along the horizontal or vertical axis, as shown in Fig. 2(c). Dong et al. [46] processed horizontal and vertical windows in parallel, similar to grouped convolution, which enhanced the performance of representation learning, as shown in Fig. 2(d). However, the grouping approach tends to ignore the directional bias of the ground objects in remote sensing images.

C. Imbalance Learning

In urban construction land monitoring, construction land typically has a smaller proportion than other land-cover types, resulting in an imbalance in the foreground and background [64]. Furthermore, due to the characteristics of the urban construction land, the imbalance may exist between the each category (e.g., building and soil under construction (SUC), Fig. 1). In the field of computer vision, much research has been devoted to the problem of how to efficiently model long-tailed class distributions. In general, previous studies have primarily used one-stage imbalance learning or two-stage imbalance learning. One-stage imbalance learning mainly includes sampling [65], [66], [67], [68], loss function weighting [69], [70], [71], [72], and transfer learning [73], [74], [75], [76]. The sampling methods include upsampling the categories with a smaller size, downsampling the categories with a larger size, or other strategies of balancing the number of samples across categories. The principle of the loss function weighting methods is to control the loss weights for each category or sample. The transfer learning methods aim to transfer the knowledge learned from the head category to the tail to improve the classification accuracy. However, these methods are usually task-related or model-related and are difficult to apply to different tasks. Two-stage imbalance learning can improve the prediction performance mainly by decoupling the feature learning and classification head [47], [48], [77], [78]. Unfortunately, however, this approach requires tedious parameter tuning and is not easily applied to the downstream tasks. Recently, Zhang et al. [49] proposed a unified framework for long-tail class prediction to accommodate various downstream tasks. Nevertheless, the approach ignores spatial contextual dependencies, which are critical for dense prediction tasks. To alleviate this problem, we propose the DASCI module in this article.

III. Method

A. Overall Architecture

The overall structure of the proposed DWin-HRFormer architecture is illustrated in Fig. 3. We follow the network structure design of HRNet [20] and HRFormer [54] and leverage two convolutional layers (3×3 convolutional layers with stride 2) for token embedding (denoted as the convolutional token embedding (CTE) module) to obtain $(H/4) \times (W/4)$ patch tokens with dimension C for each. The main body of the network consists of four stages. Each stage gradually adds convolution streams from a high-to-low resolution, with adjacent convolution streams differing in resolution by a factor of 2, and multiple convolution streams of different resolutions being performed in parallel at each stage. Specifically, the xth stage contains x convolution streams corresponding to x resolutions (the ith convolution stream corresponds to a resolution of $(H/2^{i+1}) \times (W/2^{i+1})$). At each stage, the convolution streams of different resolutions are updated by a series of sequentially connected DWin transformer blocks, and multiresolution fusion is performed by repeatedly exchanging information between parallel streams. The DWin transformer block replaces the self-attention of the standard transformer block with the proposed DWSA mechanism. Finally, the learning of each category is balanced by the DASCI module and restored to the original resolution size.

B. Directional Window Self-Attention

Even though the standard full self-attention mechanism has a strong contextual modeling capability, its computational complexity is quadratic to the size of the feature map. In this study, the urban construction land monitoring mainly uses high-resolution images at the meter or submeter level. In this situation, the dense prediction tasks (e.g., semantic segmentation and target detection) require enormous computational cost [79], [80], [81]. To alleviate this problem, some researchers [43], [44] suggested executing self-attention in a local window and applying a moving window to expand

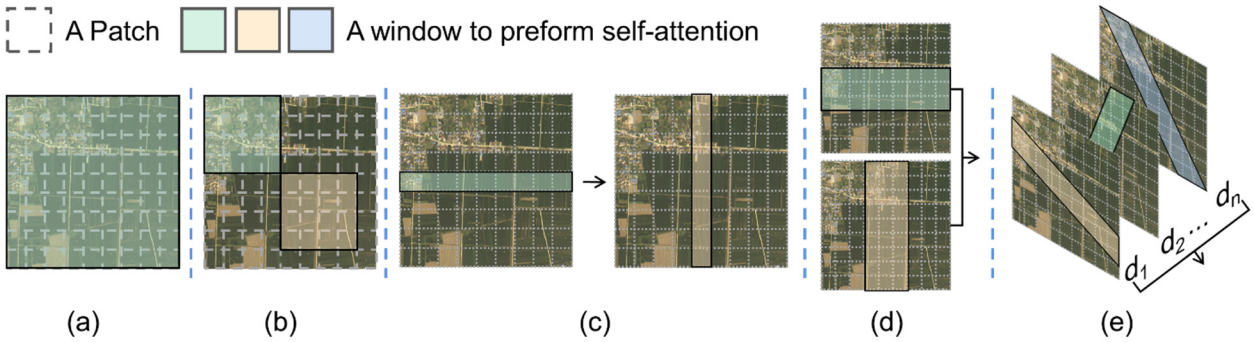


Fig. 2. (a) Full self-attention. (b) SLSA. (c) Sequential axial self-attention. (d) Cross-shaped window self-attention. (e) DWSA. Please notice that all the attention windows are divided to cover the whole picture, and the example shown in the figure is one of the windows.

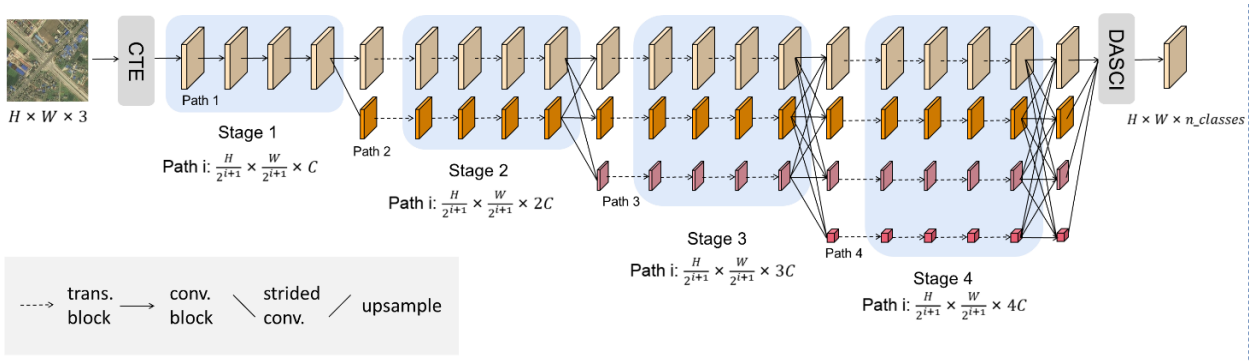


Fig. 3. Overall structure of the DWin-HRFormer architecture. The blue areas mark the different stages. The first stage uses convolutional blocks, and the other stages use DWin transformer blocks.

the receptive field. Nevertheless, in this way, the token within each transformer block still has a limited receptive field, and more blocks should be stacked to achieve global self-attention. Dong et al. [46] utilized cross-shaped windows to alleviate the problems of the limited attention area and computational complexity, but cross-shaped windows tend to ignore the directional characteristics of certain terrestrial objects (e.g., roads).

In this research, we propose a DWSA mechanism to solve this problem, as shown in Fig. 2(e). This attention mechanism divides the feature map by using strip windows in n different directions, performs self-attention in parallel within the windows in different directions, and finally merges the results of all the directional calculations. The larger the value of n is, the more complete the feature extraction in different directions is, but the computational complexity is also enlarged. Considering the balance of model performance and complexity, and the fact that the CTE module reduces the resolution of the raw images, four different orientations of strips at 0°, 45°, 90°, and 135° were adopted in this study. The structure of the DWSA mechanism is shown in Fig. 4, where, when parameter n is 4, it adequately focuses on the features of the ground objects in various directions. It can be observed that the horizontal and vertical strips focus on the interclass attention computation of the roads in the image. In contrast, the inclined strips focus on the intraclass attention computation of the roads. sw denotes the strip width, which can balance the learning ability and computational complexity. Smaller values are used for higher resolution feature maps while larger ones

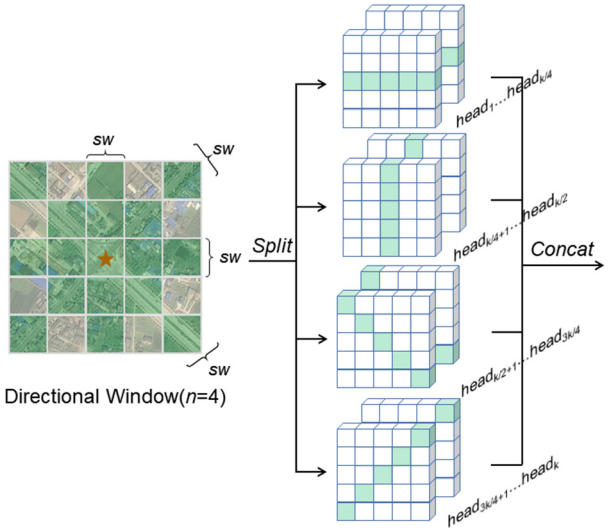


Fig. 4. DWSA mechanism.

for lower resolution maps, to speed up the model. The detailed procedure for the DWSA mechanism (with $n = 4$) is given below.

The feature map $X \in R^{(H \times W) \times C}$ is linearly projected to K heads, each of which performs a local self-attention calculation in one of the four windows.

Taking the horizontal strip window as an example, X is partitioned into P horizontal strips $[X^1, \ldots, X^P]$, each with the same width sw, and each containing $sw \times W$ tokens. The

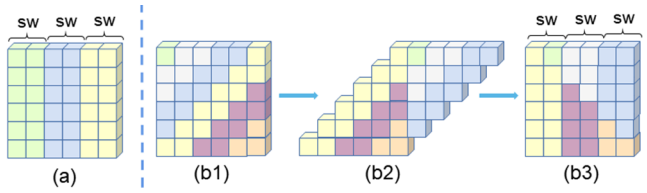


Fig. 5. (a) Horizontal strip window partitioning. (b1)–(b3) Illustration of an efficient batch computation approach for self-attention in 45° strip window partitioning.

window width can be adjusted to fit the size of the feature or to balance the computational overhead with the receptive field. The self-attention calculation is performed independently within each window. The self-attention algorithm for the kth head can be defined as follows:

$$X = \left[X^1, X^2, \ldots, X^P\right] \tag{1}$$

where $X^i \in R^{(sw\times W)\times C}$ and $P = (H/sw)$

$$Y_k^i = \text{Softmax}\left[\frac{\left(X^i W_k^Q\right)\left(X^i W_k^K\right)^{\mathrm{T}}}{\sqrt{\frac{d_k}{P}}}\right] X^i W_k^V \tag{2}$$

where $i = 1, \ldots, P$

$$\mathrm{H}-\text{Attention}_k(X) = \left[Y_k^1, Y_k^2, \ldots, Y_k^P\right] \tag{3}$$

where $W_k^Q \in R^{C\times d_k}, W_k^K \in R^{C\times d_k}$, and $W_k^V \in R^{C\times d_k}$ denote the projection matrix of the kth head queries, keys, and values, respectively, where d_k is C/K. Similarly, the output of the self-attention calculation for the vertical strip of the kth head is denoted as $V-\text{Attention}_k(X)$. Furthermore, for the skewed strip window, we propose a more efficient batch calculation method: moving the upper triangular data before the strip division and performing the mask self-attention calculation in the continuous region of the data, as shown in Fig. 5. The output of the self-attention calculation for the kth head of the tilted strip window is then denoted as $45-\text{Attention}_k(X)$ and $135-\text{Attention}_k(X)$, respectively.

The K heads are divided into four groups, each containing $K/4$ heads. These four groups correspond to four different self-attention calculations, and the final result is obtained by combining the outputs of the four groups

$$\begin{aligned}&\text{DWin}-\text{Attention}(X)\\&\quad = \text{Concat}(\text{head}_1, \ldots, \text{head}_\text{K})W^O\end{aligned} \tag{4}$$

$$\text{head}_k = \begin{cases} \mathrm{H}-\text{Attention}_k(X)k = 1, \ldots, \frac{K}{4} \\ \mathrm{V}-\text{Attention}_k(X)k = \frac{K}{4}+1, \ldots, \frac{K}{2} \\ 45-\text{Attention}_k(X)k = \frac{K}{2}+1, \ldots, \frac{3K}{4} \\ 135-\text{Attention}_k(X)k = \frac{3K}{4}+1, \ldots, K \end{cases} \tag{5}$$

where $W^O \in R^{C\times C}$ denotes the general projection matrix that projects the output dimension to the target dimension. In summary, the proposed DWSA mechanism expands the receptive field of the token within a transformer block by different head groupings. Its different directions also extend the self-attention computation within and between classes. The computational complexity of DWSA is

$$\Omega(\text{DWSA}) = \text{HWC} * (2C + sw * H + sw * W) \tag{6}$$

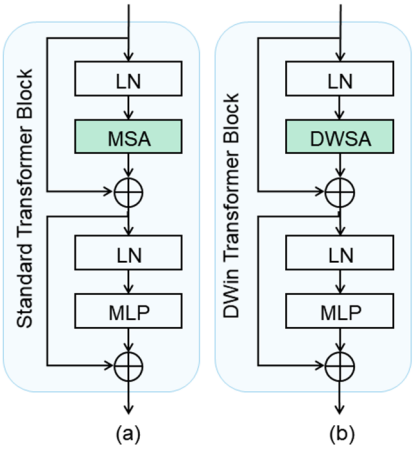


Fig. 6. (a) Structure of the standard transformer block. (b) Structure of the DWin transformer block. Layer normalization is denoted as LN. Multihead self-attention is denoted as MSA. Multilayer perceptron is denoted as MLP. Our proposed DWSA is denoted as DWSA.

where H, W, and C denote the length, the width, and the number of channels of the feature map, respectively, and sw denotes the window width. In the experimental section, we will verify the accuracy improvement brought by the different stages of the tilted strip window.

C. DWin Transformer Block

The overall structure of the DWin transformer block is shown in Fig. 6(b), which differs from the standard transformer block [Fig. 6(a)] in two aspects: 1) our proposed module uses DWSA for feature extraction and 2) it uses the locally enhanced positional encoding (LePE) method [46]. LePE uses DWConv for the position encoding, which can be efficiently applied to downstream tasks, with an arbitrary resolution as input. The LePE encoding [46] has demonstrated its superiority over absolute positional encoding (APE) [30], conditional positional encoding (CPE) [82], and relative position representation (RPE) [83]. The DWin transformer block can be expressed as follows:

$$\hat{X}^l = \text{DWin}-\text{Attention}\left(\text{LN}\left(X^{l-1}\right)\right) + X^{l-1} \tag{7}$$

$$X^l = \text{MLP}\left(\text{LN}\left(\hat{X}^l\right)\right) + \hat{X}^l \tag{8}$$

where X^l represents the output of the transformer block at layer l.

D. DA With Spatial Contextual Information

The proposed DASCI module belongs to the two-stage imbalance learning method and can flexibly adjust the correction magnitude while introducing spatial contextual information. The structure of the DASCI module is shown in Fig. 7. The Classifier Head includes only upsampling, and it can be replaced by any classification head that can improve the accuracy of the model. The black arrow indicates the first stage of the training process, the red arrow indicates the second one, and the blue arrow indicates that both stages need to be performed.

The problem with the existing long-tail methods is the biased decision boundaries [49]. Therefore, in the first stage,

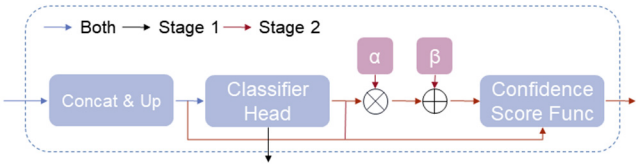


Fig. 7. Structure of the DASCI.

the network is trained using the original dataset to achieve its feature extraction ability. In this stage, the feature maps of each resolution are upsampled to the same size and combined. The prediction results at the original resolution are obtained after the classification head.

In the second stage, the focus is on correcting the classifier's output. Specifically, the training parameters of each module trained in the first stage are frozen, and the original classification predictions are obtained by inference under these parameters. We denote the category prediction after the classification head as $z = [z_1, \ldots, z_{HW}]$, where H and W are the length and the width of the feature map, respectively. z_i is a vector of length C, representing the predicted probability of each category. The original classification predictions can then be linearly mapped by α and β

$$s_i = \alpha_i \cdot z_i + \beta_i \tag{9}$$

where α and β are trainable parameters with the same size as the original image and are subsequently fed into the convolutional layer (3×3 convolutional layer with stride 1). In this way, the spatial contextual information is considered; i.e., each class at different spatial locations has different tuning parameters and can be influenced by their surrounding pixels.

Finally, the original and adjusted classification predictions are fed into the confidence score function $\sigma(x)$, which is adaptively adjusted to control the magnitude of the correction

$$\hat{z}_i = \sigma(x) \cdot s_i + (1 - \sigma(x)) \cdot z_i. \tag{10}$$

The confidence score function is obtained by linearly mapping the feature map and then passing through the softmax layer.

IV. Results and Discussion

A. Datasets

1) Luojia-FGLC Dataset: The Luojia-fine-grained land cover (FGLC) dataset covers six provinces in China: Anhui, Shanxi, Hainan, Xinjiang, Gansu, and Qinghai. The detailed geographical distribution of the Luojia-FGLC dataset is provided in Table I. The images were acquired by China's high-resolution series of satellites (i.e., GF-1, GF-2, ZY-3, and so on), containing red, green, and blue bands, with a resolution ranging from 0.8 and 1 to 2 m. We merged the original categories to the categories of background as well as the three land cover classes in the urban construction regions, i.e., building, road, and bare soil, and cropped the images to 512×512, of which 41 973 images were randomly selected for the training and 4418 for the testing.

TABLE I
Geographical Distribution of the Luojia-FGLC Dataset

Province	# patches	Total area(km^2)
Shanxi	228	6840
Anhui	106	3180
Hainan	40	1200
Qinghai	90	2700
Gansu	99	2970
Xinjiang	80	2400
All	643	19290

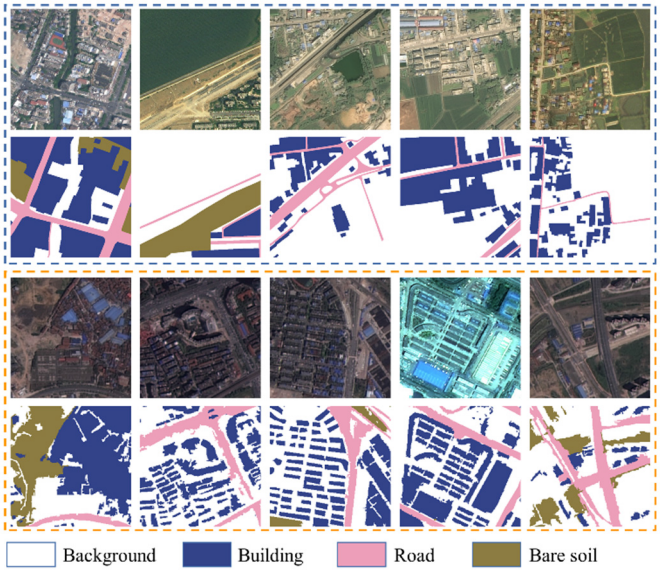


Fig. 8. Sample images and corresponding labels for the Luojia-FGLC dataset and the WHDLD. (The first and second rows are sampled from the Luojia-FGLC dataset, and the third and fourth rows are sampled from the WHDLD.)

2) Wuhan Dense Labeling Dataset [84]: The pixels of each image in this dataset are manually labeled into the following six categories: building, road, pavement, vegetation, bare soil, and water. The Wuhan Dense Labeling Dataset (WHDLD) contains 4940 red-green-blue (RGB) images with a spatial size of 256×256 and a resolution of 2 m. Similarly, we merged the original categories to the categories of background, building, road, and bare soil and randomly selected 4446 images for training and 494 for testing.

Fig. 8 shows the sample images of the two datasets.

B. Implementation Details

1) Training Settings: We used the PyTorch framework to construct the network model proposed in this article. An AdamW optimizer with a weight decay of $1e{-}4$ was applied to optimize the model. A cross-entropy loss function was employed, the initial learning rate was set to $1e{-}4$, and a cosine annealing decay strategy was adopted. The batch size was set to 6, and the maximum epoch number was 120. At the same time, the images were normalized, regularized, and enhanced using horizontal flip, vertical flip, random angle rotation, color dithering, and so on. All the experiments were completed on NVIDIA GeForce RTX 3090 GPUs $\times$ 3.

2) Evaluation Indices: In this article, the intersection-over-union (IoU), the mean IoU (mIoU), and the average $F1$-score

TABLE II

ABLATION EXPERIMENTS IN DIRECTION WINDOW SELF-ATTENTION ON THE WHDLD. ALL RESULTS IN THE TABLE ARE AVERAGED OVER MULTIPLE EXPERIMENTS. PARENTHESES IN THE MIOU COLUMN INDICATE THE STANDARD DEVIATIONS OF MULTIPLE EXPERIMENTS. THE AVERAGE ACCURACY GAINS OF THE SPECIFIC CLASSES RELATIVE TO THE BASELINE ARE ALSO PROVIDED IN THE BRACKETS

DWSA				mIoU(%)	IoU(%)			FPS
Path 1	Path 2	Path 3	Path 4		Building	Road	Bare soil	
√				57.22(0.105)	61.42(↑1.62)	65.31(↑3.28)	44.92(↑2.41)	22
	√			57.53(0.063)	61.70(↑1.90)	65.91(↑3.88)	44.97(↑2.46)	20
		√		57.67(0.063)	62.18(↑2.38)	**65.97(↑3.94)**	44.87(↑2.36)	19
			√	57.19(0.055)	61.80(↑2.00)	65.78(↑3.75)	43.98(↑1.47)	18
	√	√		**57.73**(0.055)	**62.27(↑2.47)**	65.93(↑3.90)	**44.99(↑2.48)**	22
√	√	√	√	56.95(0.055)	61.74(↑1.94)	65.13(↑3.10)	43.97(↑1.46)	**25**
HRFormer(baseline)				54.78(0.077)	59.80	62.03	42.51	17

TABLE III

ABLATION EXPERIMENTS WITH THE DASCI MODULE ON THE LUOJIA-FGLC DATASET AND THE WHDLD. THE BASELINE HERE REFERS TO THE HRNET USING DWSA. DA INDICATES THE DISTRIBUTION ALIGNMENT MODULE, AND DASCI INDICATES THE PROPOSED DA MODULE. ALL RESULTS IN THE TABLE ARE AVERAGED OVER MULTIPLE EXPERIMENTS. PARENTHESES IN THE MIOU COLUMN INDICATE THE STANDARD DEVIATIONS OF MULTIPLE EXPERIMENTS. THE AVERAGE ACCURACY GAINS OF THE SPECIFIC CLASSES RELATIVE TO THE BASELINE ARE ALSO PROVIDED IN THE BRACKETS

Dataset	Method	IoU(%)			Evaluation index	
		Building	Road	Bare soil	mIoU(%)	Average F1-score
Luojia-FGLC	Baseline	69.70	**56.52**	28.77	51.66(0.055)	66.36
	Baseline(w/ DA)	69.97(↑0.27)	55.44(↓1.08)	29.99(↑1.22)	51.80(0.063)	66.62(↑0.26)
	Baseline(w/ DASCI)	**70.16(↑0.46)**	56.50(↓0.02)	**30.15(↑1.38)**	**52.27**(0.055)	**66.77(↑0.31)**
WHDLD	Baseline	61.74	65.13	43.97	56.95(0.055)	72.12
	Baseline(w/ DA)	62.19(↑0.45)	65.84(↑0.71)	44.05(↑0.10)	57.36(0.063)	72.42(↑0.30)
	Baseline(w/ DASCI)	**62.28(↑0.54)**	**66.25(↑1.12)**	**45.43(↑1.46)**	**57.99**(0.055)	**72.96(↑0.83)**

are used to evaluate the model performance

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{11}$$

$$\text{mIoU} = \frac{1}{k+1} \sum_{i=0}^{k} \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{12}$$

$$\text{AverageF1} = \frac{1}{k+1} \sum_{i=0}^{k} 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{13}$$

For each category, TP denotes true prediction on a positive sample, FP denotes false prediction on a positive sample, and FN denotes false prediction on a negative sample. Precision indicates the proportion of TP in the total positive prediction, and recall indicates the proportion of TP in the total positive samples.

We evaluated the model speed in terms of the number of images processed per second, measured in frames per second (FPS). All the FPS values in this article are model inference speed.

C. Ablation Study

To evaluate the performance of the DWSA mechanism and DASCI module in urban construction land monitoring, we used HRFormer [54] as a baseline to perform ablation experiments on the WHDLD [84].

1) Effect of DWSA: The DWSA mechanism can capture the features of ground objects in different directions. In the proposed network structure, the different paths have different resolutions. We applied the DWSA mechanism to varying combinations of paths and analyzed its effect on the classification accuracy of each category. The results are listed in Table II.

First, we replaced the original attention mechanism with the DWSA mechanism on each path. Note that without using the DWSA mechanism on all the paths, the network degenerates to the original HRFormer. Results show that when the DWSA mechanism is used for Path2, the most significant gain is obtained for bare soil, with a 2.46% increment in IoU compared with the baseline. The DWSA mechanism used for Path3 exhibits the largest gains for buildings and roads, with the IoU increasing by 2.38% and 3.94%, respectively, compared with the baseline. It is apparent that the performances of the DWSA mechanism in the higher resolution paths (Path1) and lower resolution paths (Path4) are not as significant as its use in the medium-resolution paths (Path2 and Path3), in terms of the classification accuracy. In the case of lower resolution paths, a possible reason for this phenomenon is the small size of the feature maps on the lower resolution paths; e.g., the size of the feature map for Path4 in the proposed network is 16 × 16. At this size, the attention window (7 × 7 in HRFormer and 8 × 16 in the proposed network) is close to one-half of the global attention, resulting in the features' orientation bias being insignificant within each window. Hence, the improvement in the network performance is limited. On the other hand, the features in the higher resolution paths contain rich spatial information, and the model focuses on the edges of the features and the information about the smaller objects. Moreover, the smaller window in this path makes the features more fragmented by each window, so that some land cover classes cannot be entirely included in one window. Hence, in this case, the improvement in the classification effect is also limited.

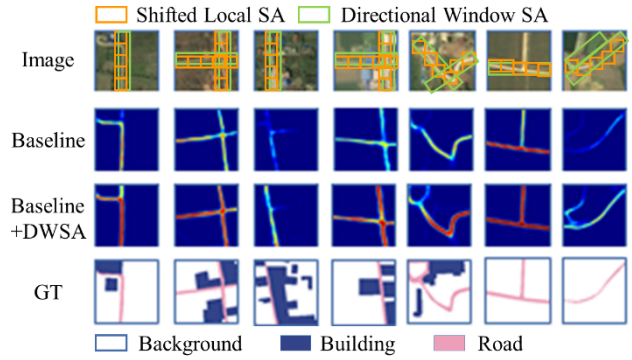


Fig. 9. Difference between DWSA and SLSA in feature extraction. The orange boxes indicate the windows of SLSA, and the green boxes indicate the windows of DWSA. Baseline refers to HRformer. Baseline (w/ DWSA) refers to the HRformer with the DWSA module.

Second, we adopted the proposed DWSA mechanism in all the paths. In this situation, the mIoU and the IoU for buildings, roads, and bare soil increased by 2.17%, 1.94%, 3.10%, and 1.46%, respectively, compared with the baseline network. However, this is less effective than just performing attention mechanism in a certain path. A possible reason is that multiple paths with different resolutions are parallel in the HRNet architecture, and the different paths process simultaneously and interact with the information after different stages. This mechanism makes the output of multiple paths have similar information, hence leading to information redundancy, and affects the final result of the network. Therefore, performing a directional self-attention calculation on every parallel path may be unnecessary and redundant. Our results show that, actually, it is more reasonable to correct the directional bias of the network feature extraction on one or two paths, and then, the effects can be applied to other paths during the multipath information fusion. To verify this conclusion, we adopted the DWSA mechanism in Path2 and Path3, resulting in the highest mIoU, and class-specific IoU than other paths.

To better illustrate the orientation information of the features and the prediction bias of the model as well as the motivation for proposing DWSA, we take the roads as the examples of SLSA and DWSA. As shown in Fig. 9, the orange boxes indicate the windows of SLSA, and the green boxes indicate the windows of DWSA. Note that all the windows of SLSA and DWSA cover the whole image, but here we only draw the windows that cover the roads in the image. It can be observed that SLSA requires more windows to cover the roads in the image, thus leading to the segmentation fragmentation of the roads. Notably, the attention computation is performed independently between different windows. Therefore, in this way, it is difficult to capture the complete road information (including directional information) within a window or through the information interaction between multiple windows. In contrast, DWSA can contain and describe a complete road within one or two windows, thus enabling more effective road feature extraction. In addition, the overlapping area of the windows in different directions (e.g., the windows of four directions in this article) of DWSA is larger than that of SLSA, which is more conducive to the information interaction between the windows with different directions. In this way, DWSA is more effective in extracting directional features of ground objects (e.g., roads). Fig. 9 shows the difference in feature extraction between DWSA and SLSA, where the second and third rows demonstrate the intermediate feature maps of the models with SLSA and DWSA, respectively.

In addition, we evaluated the influence of model speed when replacing the original attention mechanism by the proposed DWSA in different paths. The results (see Table II) show that from the high-resolution path to the low-resolution one, the feature map size decreases, the attention window size increases, and the model speed slows down. This phenomenon is in line with our expectations. The model is fastest when replacing the attention mechanism by our proposed DWSA in all the paths.

2) Effect of the DASCI Module: As shown in Table III, DW-Baseline (w/ DA) and Baseline (w/ DASCI) indicate the DA module [49], and the proposed DA module DASCI is embedded to the proposed DW-HRFormer network, respectively. When tested on the Luojia-FGLC dataset, the Baseline (w/ DASCI) achieves an increment of 0.46, 1.38, 0.61, and 0.31 over the Baseline in building IoU, bare soil IoU, mIoU, and average $F1$-score, respectively. It gives the best accuracy for the bare soil that has the smallest sample size, suggesting that the proposed DASCI module can correct the biased decision boundaries in the second training stage. Furthermore, compared with the DA module, the DASCI module shows an increase of 0.19, 0.16, 0.47, and 0.05 in building IoU, bare soil IoU, mIoU, and average $F1$-score, respectively, verifying the importance of spatial information for dense prediction tasks.

In the experiment on the WHDLD, similar results were obtained. On this dataset, Baseline (w/ DASCI) shows the increases of 0.54, 1.12, 1.46, 1.04, and 0.83 over Baseline for building IoU, road IoU, bare soil IoU, mIoU, and average $F1$-score, respectively. Compared with the DA module, it shows an increase of 0.09, 0.41, 1.36, 0.63, and 0.53 in building IoU, road IoU, bare soil IoU, mIoU, and average $F1$-score, respectively.

3) Visualization of Ablation Experiment Results: As mentioned before, the DWSA and DASCI modules can improve the accuracy of the semantic segmentation. To better demonstrate their gains on the results, a visual comparison is provided (Fig. 10). The baseline refers to HRformer. Baseline (w/ DWSA) refers to the HRformer with the DWSA module. Baseline (w/ DWSA and DASCI) is the proposed network, which uses both DWSA and DASCI modules. The DWSA module effectively takes the directional information of the features into account, and significant gains are obtained for the objects with obvious directional information, e.g., roads, as shown in the red circles in Fig. 10(b). The DASCI module can alleviate the accuracy degradation caused by the unbalanced number of the land categories, and its segmentation results for the categories with smaller numbers (e.g., bare soil) are significantly improved, while the results for other categories (e.g., buildings) can also be improved to varying degrees, as shown in the red circles in Fig. 10(c).

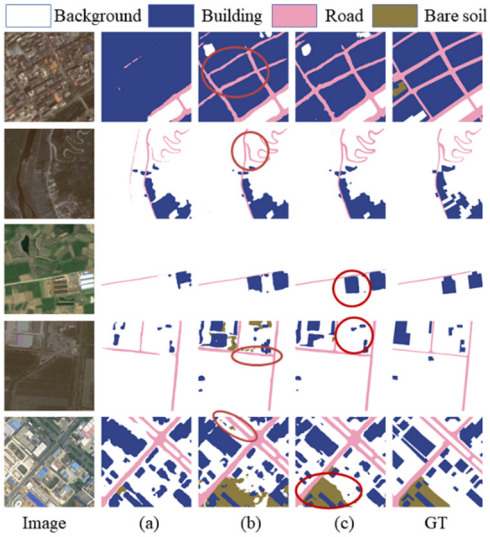


Fig. 10. Visual comparison of the results of ablation experiments on the Luojia-FGLC dataset. (a) Baseline. (b) Baseline (w/ DWSA). (c) Baseline (w/ DWSA and DASCI). Baseline refers to HRformer. Baseline (w/ DWSA) refers to the HRformer with the DWSA module. Baseline (w/ DWSA and DASCI) is the proposed network, which uses both DWSA and DASCI modules.

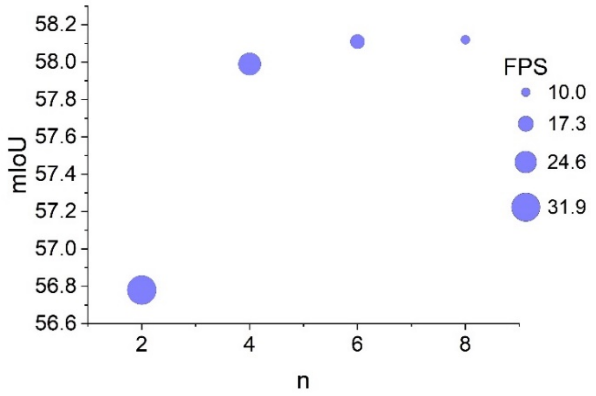


Fig. 11. Performance of DWin-HRFormer with different values of n. The horizontal axis indicates the number of strip directions, and the vertical axis indicates the mIoU. The size of the circle indicates FPS during inference. A larger FPS value represents faster model inference.

D. Discussion of the Hyperparameter n

We have discussed the number of directions of the window. As shown in Fig. 11, the increase in the number of directions can improve the prediction accuracy of the model. However, after n is greater than 4, the gain becomes marginal. A possible reason is that when the number of directions increases, the overlap of windows in different directions becomes larger, resulting in a decrease of the individual contribution from each direction. At the same time, the increase in strip types causes the model to be slower, since the dimensionality of the feature map has to be changed frequently when computing attention at different directions. Therefore, balancing the accuracy and efficiency of the model, in this article, n is set to 4.

E. Comparison With SOTA Methods

To validate the performance of the proposed network, we performed comparison experiments with the existing state-of-the-art (SOTA) networks, e.g., UNet [16], HRNetV2-W48 [20], MAP-Net [21], swin transformer embedded in a U-shaped network (ST-UNet) [53], HRFormer [54], UNet-like fully transformer-based network (FT-UNetFormer) [85], and densely connected swin transformer (DC-Swin) [86]. Among these networks, UNet [16], HRNetV2-W48 [20], and MAP-Net [21] are CNN-based methods, while ST-UNet [53], HRFormer [54], FT-UNetFormer [85], DC-Swin [86], and the proposed network are transformer-based methods. UNet [16], ST-UNet [53], FT-UNetFormer [85], and DC-Swin [86] are encoder–decoder structures, and all the other networks belong to high-resolution structures.

TABLE IV

COMPARISON OF THE SEGMENTATION RESULTS OBTAINED ON THE LUOJIA-FGLCDATASET AND MCNEMAR'S TEST BETWEEN THE PROPOSED METHOD AND OTHER ONES. THE SIGNIFICANTLY DIFFERENT METHODS ARE INDICATED AS ** WITH $\gamma > 3.84$ AT 95% CONFIDENCE LEVEL AND * FOR $\gamma > 2.71$ AT 90% LEVEL, RESPECTIVELY

Method	IoU(%)			Evaluation index	
	Building	Road	Bare soil	mIoU(%)	Average F1-score
UNet	68.56**	53.13**	23.42**	48.37**	62.91**
HRNetV2-W48	69.54**	55.75**	27.55**	50.95**	65.61**
MAP-Net	69.14**	54.92**	28.11**	50.72**	65.51**
ST-UNet	65.27**	49.04**	24.21**	46.17**	61.25**
HRFormer	63.89**	48.47**	21.51**	44.62**	59.55**
FT-UNetFormer	69.43**	56.34**	22.69**	49.49**	63.62**
DC-Swin	67.96**	53.20**	23.89**	48.35**	62.98**
DWin_HRFormer	**70.16**	**56.50**	**30.15**	**52.27**	**66.77**

1) Results on the Luojia-FGLC Dataset: The experimental results of each method obtained on the Luojia-FGLC dataset are listed in Table IV. In general, the proposed network shows the highest accuracy compared with the other ones. Building IoU, road IoU, bare soil IoU, mIoU, and the average $F1$-score are increased by 0.62, 0.16, 2.04, 1.32, and 1.16, respectively, compared with the best results of the other methods.

HRNetV2-W48 and MAP-Net show an improvement of 2.58 and 2.35 over the mIoU of UNet, respectively, indicating that the HRNet architecture can better deal with the multiscale issues. It can be seen that the mIoU of UNet is 2.20 higher than that of ST-UNet, and the mIoU of HRNetV2-W48 is 6.33 higher than that of HRFormer. The transformer-based models can model long-range dependencies effectively, but they rely on a considerable amount of data. Moreover, transformer-based models tend to require more computation and memory resources to maintain a high batch size, due to the numerous parameters. Notably, in this study, we used the same hardware and software configurations to fairly compare all the networks. In this way, however, transformer-based models were implemented with smaller batch sizes, leading to their performance degradation to some extent. Therefore, the performance of the proposed network can potentially be further improved with a larger batch size.

Fig. 12 shows the visualization results of the related semantic segmentation methods. From the first row, it can be seen that some parts of the road are obscured by buildings or vegetation in the crowded building areas, which increases the difficulty of the road extraction. Compared with the other ones, the proposed method can more accurately distinguish between roads and buildings. The situation is similar in the second row. Our classification results are more accurate in identifying the roads whose features are not apparent (see the red circle

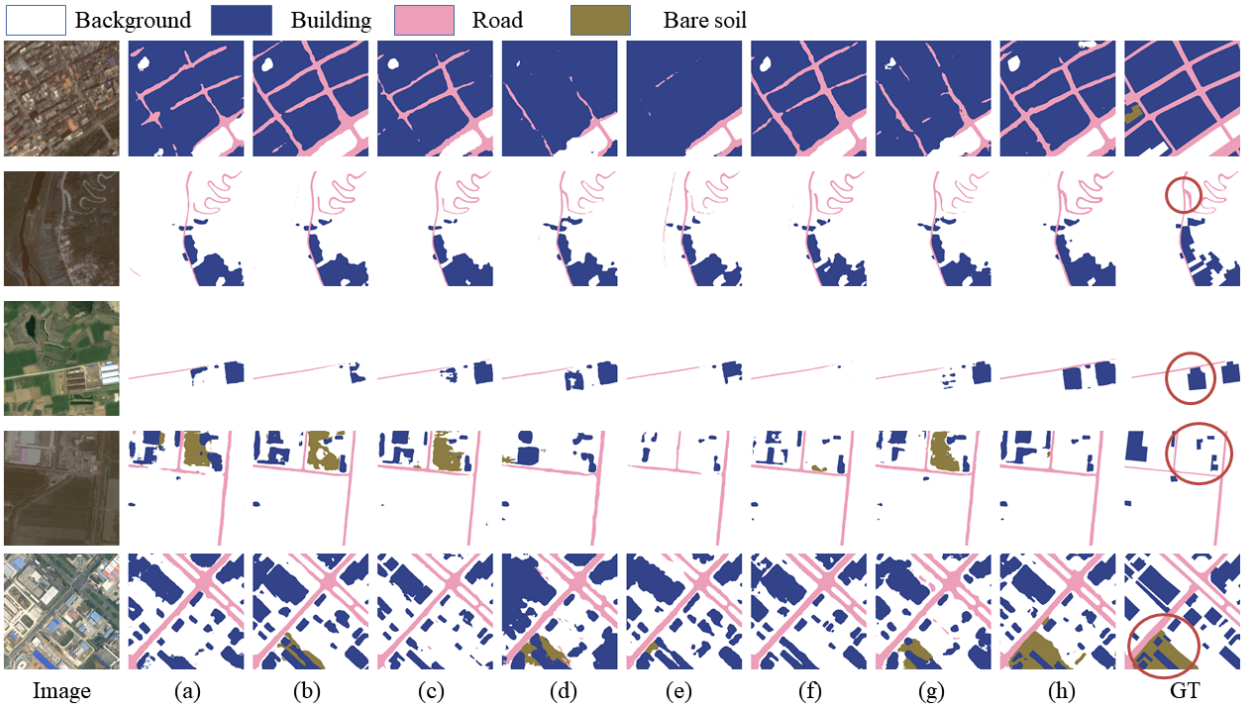


Fig. 12. Examples of semantic segmentation results obtained on the Luojia-FGLC dataset. (a) UNet. (b) HRNetV2-W48. (c) MAP-Net. (d) ST-UNet. (e) HRFormer. (f) FT-UNetFormer. (g) DC-Swin. (h) DWin-HRFormer.

in the second row), which can be attributed to the fact that the attention computation from different directions is more capable of describing the road spatial contextual information. In the third row, the proposed method correctly distinguishes different styles of buildings, also benefiting from the proposed attention mechanism. Specifically, in the proposed method, the window width increases with the decreasing feature map size on different paths, which not only improves the extraction of elongated and directional features, but also the extraction of buildings. The main challenge at the fourth and fifth rows lies in the identification of bare soil, as marked by the red circles. It can be clearly seen that the proposed method is more effective in extracting the bare soil from the urban scenes. The DASCI module introduces contextual information around the ground objects, making its classification results not only dependent on the features of the ground objects themselves, but also influenced by their surroundings.

2) Results on the WHDLD: Table V lists the experimental results of different semantic segmentation methods on the WHDLD. The proposed method achieves an mIoU of 58.0% and an average $F1$-score of 73.0%, outperforming the other ones in all the metrics. This further demonstrates the superiority of the proposed network for the semantic segmentation task of urban construction land. The accuracies obtained on the WHDLD are generally higher than those obtained on the Luojia-FGLC dataset, due to the differences between the datasets. However, their trend and conclusions are similar.

Fig. 13 shows the prediction results of the different semantic segmentation methods. Similarly, at the first and second rows, the effectiveness of the proposed model in identifying roads from the dense buildings is again demonstrated. At the third row, the integrity of the single building extracted by the proposed method is better. The extraction results for the building areas are more precise at the fourth row. At the last row, the proposed model identifies a small piece of bare soil, which is missed by other methods. The proposed model performs consistently on different datasets, further demonstrating the model's robustness.

TABLE V

COMPARISON OF THE SEGMENTATION RESULTS OBTAINED ON THE WHDLD AND MCNEMAR'S TEST BETWEEN THE PROPOSED METHOD AND OTHER ONES. THE SIGNIFICANTLY DIFFERENT METHODS ARE INDICATED AS ** WITH $\gamma > 3.84$ AT 95% CONFIDENCE LEVEL AND * FOR $\gamma > 2.71$ AT 90% LEVEL, RESPECTIVELY

Method	IoU(%)			Evaluation index	
	Building	Road	Bare soil	mIoU(%)	Average F1-score
UNet	60.57**	63.73**	43.92**	56.07**	71.45**
HRNetV2-W48	61.64**	65.20**	43.88**	56.91**	72.07**
MAP-Net	61.28**	65.13**	44.22**	56.88**	72.07**
ST-UNet	58.72**	61.83**	44.34**	54.96**	70.61**
HRFormer	59.74**	62.01**	42.61**	54.78**	70.37**
FT-UNetFormer	58.38**	61.23**	45.35**	54.99**	70.71**
DC-Swin	54.85**	57.30**	41.95**	51.37**	67.60**
DWin_HRFormer	**62.28**	**66.25**	**45.43**	**57.99**	**72.96**

F. Efficiency Analysis

Table VI compares the speed and volume of model parameters for all the models in the same implementation environment. A larger FPS indicates that the model has a faster processing speed. Although the transformer-based models outperform the CNN structure in capturing long-range dependencies, the computational efficiency of the transformer-based models is generally lower than that of the CNN-based

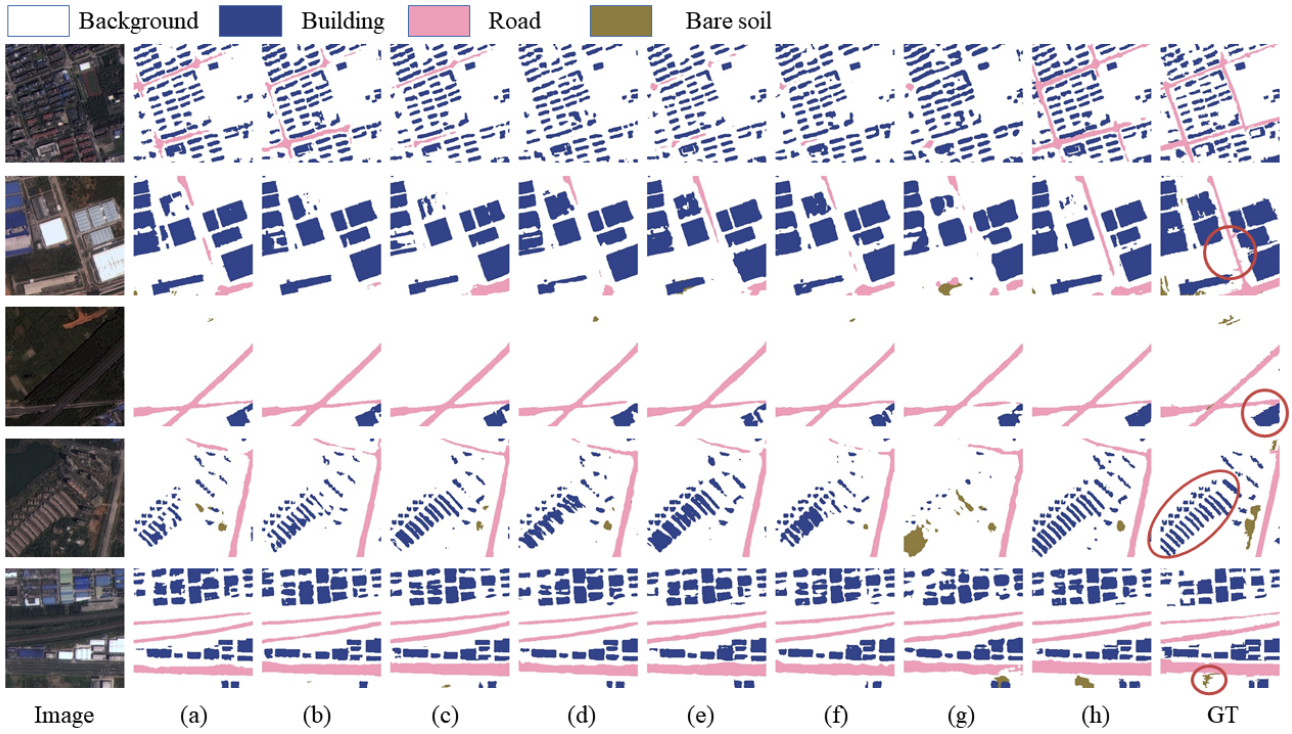


Fig. 13. Examples of semantic segmentation results obtained on the WHDLD. (a) UNet. (b) HRNetV2-W48. (c) MAP-Net. (d) ST-UNet. (e) HRFormer. (f) FT-UNetFormer. (g) DC-Swin. (h) DWin-HRFormer.

TABLE VI
COMPARISON OF THE MODEL PARAMETERS, SPEED, AND ACCURACY. FPSIS THE SPEED OF INFERENCE. TIME REFERS TO THE TIME CONSUMED FOR TRAINING

Method	Parameters	FPS	Luojia-FGLC		WHDLD		Basis
			Time(h)	mIoU(%)	Time(h)	mIoU(%)	
UNet	25.13MB	78	41.1	48.37	4.4	56.07	CNN
HRNetV2-W48	63.60MB	36	77.7	50.95	8.2	56.91	CNN
MAP-Net	24.00MB	51	77.7	50.72	8.2	56.88	CNN
ST-UNet	160.97MB	10	233.2	46.17	24.7	54.96	Transformer
HRFormer	43.20MB	18	199.9	44.62	21.2	54.78	Transformer
FT-UNetFormer	165.61MB	61	87.4	49.49	9.3	54.99	Transformer
DC-Swin	253.20MB	60	82.3	48.35	8.7	51.37	Transformer
DWin-HRFormer	93.18MB	25	88.4	52.27	10.0	57.99	Transformer

models. HRNetV2-W48 maintains the high resolution of the feature map, and therefore, its parameters are more than those of the other CNN methods. MAP-Net benefits from some of its lightweight structure [21]. ST-UNet has a larger volume of parameters, since it uses both the Swin transformer and UNet encoder. The proposed model simultaneously parallels multiple directional attention windows and includes the DASCI module, so its parameters are greater than those of HRFormer. However, its unique attention mechanism makes its speed competitive among the transformer-based models.

V. CONCLUSION

In practical use, urban construction land monitoring involves the detection of buildings, bare soil, roads, and urban background. It represents an actual application of urban remote sensing and is of great significance for urban planning and illegal land-use monitoring in rapidly urbanizing areas, such as in many Chinese cities. In the existing literature, the relevant studies concerning the urban construction land are lacking. In order to meet the demand of urban construction land monitoring in terms of efficiency and accuracy, and to cope with the problem caused by the imbalance between the categories in the urban construction land, the DWin-HRFormer neural network has been proposed in this article. Specifically, the proposed directional attention mechanism can effectively extract the features of the geographical objects in each direction, depict road information more completely, reduce road breakage, and detect building information more effectively. Meanwhile, the proposed DASCI module can effectively cope with the data imbalance problem in semantic segmentation of urban construction land and improves the extraction accuracy for the categories with a small sample size.

We verified the effectiveness of the various network modules through extensive ablation experiments and explored the effect of the directional attention mechanism on the network at different resolutions. Experiments on two benchmark datasets showed that the proposed DWin-HRFormer network

outperformed the other semantic segmentation algorithms, with a higher accuracy and lower complexity.

Overall, this research provides a new approach for accurate and effective extraction of construction land information. In our future work, we plan to further advance the urban construction land monitoring algorithms by improving the algorithm speed and the accuracy of the boundary extraction.

ACKNOWLEDGMENT

The authors would also like to thank the editors and anonymous reviewers for the insightful suggestions, which significantly improved the quality of this article.

REFERENCES

[1] S. Jiang, C. Jiang, and W. Jiang, "Efficient structure from motion for large-scale UAV images: A review and a comparison of SfM tools," *ISPRS J. Photogramm. Remote Sens.*, vol. 167, pp. 230–251, Sep. 2020.

[2] T. Hoeser and C. Kuenzer, "Object detection and image segmentation with deep learning on Earth observation data: A review—Part I: Evolution and recent trends," *Remote Sens.*, vol. 12, no. 10, p. 1667, May 2020.

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

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

[5] X. Huang et al., "High-resolution urban land-cover mapping and landscape analysis of the 42 major cities in China using ZY-3 satellite images," *Sci. Bull.*, vol. 65, no. 12, pp. 1039–1048, 2020.

[6] X. Huang, J. Yang, J. Li, and D. Wen, "Urban functional zone mapping by integrating high spatial resolution nighttime light and daytime multi-view imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 403–415, May 2021.

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

[8] R. Sheikh, A. Milioto, P. Lottes, C. Stachniss, M. Bennewitz, and T. Schultz, "Gradient and log-based active learning for semantic segmentation of crop and weed for agricultural robots," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 1350–1356.

[9] M. Aboelnour and B. A. Engel, "Application of remote sensing techniques and geographic information systems to analyze land surface temperature in response to land use/land cover change in greater Cairo region, Egypt," *J. Geographic Inf. Syst.*, vol. 10, no. 1, pp. 57–88, 2018.

[10] S. Chen, S. Zeng, and C. Xie, "Remote sensing and GIS for urban growth analysis in China," *Photogramm. Eng. Remote Sens.*, vol. 66, no. 5, pp. 593–598, 2000.

[11] M. Sabet Sarvestani, A. L. Ibrahim, and P. Kanaroglou, "Three decades of urban growth in the city of Shiraz, Iran: A remote sensing and geographic information systems application," *Cities*, vol. 28, no. 4, pp. 320–329, Aug. 2011.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

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

[15] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[17] 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.

[18] Y. Cao and X. Huang, "A deep learning method for building height estimation using high-resolution multi-view imagery over urban areas: A case study of 42 Chinese cities," *Remote Sens. Environ.*, vol. 264, Oct. 2021, Art. no. 112590.

[19] 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. (ECCV)*, 2018, pp. 801–818.

[20] J. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[21] 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.

[22] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[23] W. Kang, Y. Xiang, F. Wang, and H. You, "EU-Net: An efficient fully convolutional network for building extraction from optical remote sensing images," *Remote Sens.*, vol. 11, no. 23, p. 2813, Nov. 2019.

[24] A. Howard et al., "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 1314–1324.

[25] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[26] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[27] L. Li, J. Liang, M. Weng, and H. Zhu, "A multiple-feature reuse network to extract buildings from remote sensing imagery," *Remote Sens.*, vol. 10, no. 9, p. 1350, Sep. 2018.

[28] G. Wu et al., "Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks," *Remote Sens.*, vol. 10, no. 3, p. 407, 2018.

[29] 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.

[30] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[32] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1877–1901.

[33] M. Chen et al., "Generative pretraining from pixels," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1691–1703.

[34] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[35] 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.*, 2020, pp. 213–229.

[36] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.

[37] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 6881–6890.

[38] H. Chen et al., "Pre-trained image processing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 12299–12310.

[39] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 568–578.

[40] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8739–8748.

[41] Y. Wang et al., "End-to-end video instance segmentation with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 8741–8750.

[42] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7262–7272.

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

[44] A. Vaswani, P. Ramachandran, A. Srinivas, N. Parmar, B. Hechtman, and J. Shlens, "Scaling local self-attention for parameter efficient visual backbones," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12894–12904.

[45] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans, "Axial attention in multidimensional transformers," 2019, *arXiv:1912.12180*.

[46] X. Dong et al., "CSWin transformer: A general vision transformer backbone with cross-shaped windows," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 12124–12134.

[47] K. Tang, J. Huang, and H. Zhang, "Long-tailed classification by keeping the good and removing the bad momentum causal effect," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1513–1524.

[48] A. Krishna Menon, S. Jayasumana, A. Singh Rawat, H. Jain, A. Veit, and S. Kumar, "Long-tail learning via logit adjustment," 2020, *arXiv:2007.07314*.

[49] S. Zhang, Z. Li, S. Yan, X. He, and J. Sun, "Distribution alignment: A unified framework for long-tail visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2361–2370.

[50] H. Cao et al., "Swin-UNet: UNet-like pure transformer for medical image segmentation," 2021, *arXiv:2105.05537*.

[51] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[52] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "UCTransNet: Rethinking the skip connections in U-Net from a channel-wise perspective with transformer," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 36, no. 3, pp. 2441–2449.

[53] 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.

[54] Y. Yuan et al., "HRFormer: High-resolution vision transformer for dense predict," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 7281–7293.

[55] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating long sequences with sparse transformers," 2019, *arXiv:1904.10509*.

[56] K. Choromanski et al., "Rethinking attention with performers," 2020, *arXiv:2009.14794*.

[57] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are RNNs: Fast autoregressive transformers with linear attention," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 5156–5165.

[58] N. Kitaev, Ł. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," 2020, *arXiv:2001.04451*.

[59] J. W. Rae, A. Potapenko, S. M. Jayakumar, and T. P. Lillicrap, "Compressive transformers for long-range sequence modelling," 2019, *arXiv:1911.05507*.

[60] A. Roy, M. Saffar, A. Vaswani, and D. Grangier, "Efficient content-based sparse attention with routing transformers," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 53–68, Feb. 2021.

[61] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers & distillation through attention," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[62] H. Wu et al., "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 22–31.

[63] L. Yuan et al., "Tokens-to-token ViT: Training vision transformers from scratch on ImageNet," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 558–567.

[64] Z. Zheng, Y. Zhong, J. Wang, and A. Ma, "Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4096–4105.

[65] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 28, pp. 321–357, Jun. 2006.

[66] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in *Proc. Int. Conf. Intell. Comput.*, 2005, pp. 878–887.

[67] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Large-scale long-tailed recognition in an open world," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2537–2546.

[68] Y. Gao et al., "Solution for large-scale hierarchical object detection datasets with incomplete annotation and data imbalance," 2018, *arXiv:1810.06208*.

[69] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[70] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3573–3587, Feb. 2017.

[71] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[72] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 4334–4343.

[73] Y.-X. Wang, D. Ramanan, and M. Hebert, "Learning to model the tail," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[74] Y.-X. Wang, R. Girshick, M. Hebert, and B. Hariharan, "Low-shot learning from imaginary data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7278–7286.

[75] T.-Y. Wu, P. Morgado, P. Wang, C.-H. Ho, and N. Vasconcelos, "Solving long-tailed recognition with deep realistic taxonomic classifier," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 171–189.

[76] J. Wu, C. Zhou, Q. Zhang, M. Yang, and J. Yuan, "Self-mimic learning for small-scale pedestrian detection," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2012–2020.

[77] T. Wang et al., "The devil is in classification: A simple framework for long-tail instance segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 728–744.

[78] X. Wang, T. E. Huang, T. Darrell, J. E. Gonzalez, and F. Yu, "Frustratingly simple few-shot object detection," 2020, *arXiv:2003.06957*.

[79] D. Wen et al., "Change detection from very-high-spatial-resolution optical remote sensing images: Methods, applications, and future directions," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 4, pp. 68–101, Dec. 2021.

[80] X. Huang et al., "A multispectral and multiangle 3-D convolutional neural network for the classification of ZY-3 satellite images over urban areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10266–10285, Dec. 2021.

[81] X. Huang, Y. Cao, and J. Li, "An automatic change detection method for monitoring newly constructed building areas using time-series multi-view high-resolution optical satellite images," *Remote Sens. Environ.*, vol. 244, Jul. 2020, Art. no. 111802.

[82] X. Chu et al., "Conditional positional encodings for vision transformers," 2021, *arXiv:2102.10882*.

[83] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," 2018, *arXiv:1803.02155*.

[84] Z. Shao, W. Zhou, X. Deng, M. Zhang, and Q. Cheng, "Multilabel remote sensing image retrieval based on fully convolutional network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 318–328, 2020.

[85] 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.

[86] 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," 2021, *arXiv:2104.12137*.

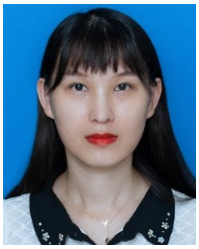
Zhen Zhang received the B.S. degree in earth information science and technology from Henan Polytechnic University, He'nan, China, in 2019. He is currently pursuing the Ph.D. degree in remote sensing with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China.

His research interests include building information extraction, land cover classification, high-resolution image processing, and deep learning.

Xin Huang (Senior Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China, in 2009.

He is currently a Full Professor with Wuhan University and he also with Hubei Luojia Laboratory, where he teaches remote sensing and image interpretation. He is also the Head of the Institute of Remote Sensing Information Processing (IRSIP), School of Remote Sensing and Information Engineering, Wuhan University. He has authored or coauthored more than 200 peer-reviewed articles (Science Citation Index (SCI) papers) in the international journals. His research interests include remote sensing image processing methods and applications.

Prof. Huang has been an Editorial Board Member of the Remote Sensing of Environment since 2019. He was a recipient of the Boeing Award for the Best Paper in Image Analysis and Interpretation from the American Society for Photogrammetry and Remote Sensing (ASPRS) in 2010, the John I. Davidson President's Award from ASPRS in 2018, and the National Excellent Doctoral Dissertation Award of China in 2012. In 2011, he was recognized by the IEEE Geoscience and Remote Sensing Society (GRSS) as the Best Reviewer of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS. He was the winner of the IEEE GRSS Data Fusion Contest in 2014 and 2021. He was a Lead Guest Editor of the special issue for the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, the *Journal of Applied Remote Sensing Photogrammetric Engineering and Remote Sensing*, and *Remote Sensing*. He was an Associate Editor of the *Photogrammetric Engineering and Remote Sensing* from 2016 to 2019, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS from 2014 to 2020, and the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING from 2018 to 2022. He has been serving as an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING since 2022.



Jiayi Li (Senior Member, IEEE) received the B.S. degree from Central South University, Changsha, China, in 2011, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2016.

She is currently an Associate Professor with the School of Remote Sensing and Information Engineering, Wuhan University, and she also with Hubei Luojia Laboratory. She has authored more than 60 peer-reviewed articles (Science Citation Index (SCI) articles) in international journals. Her research interests include hyperspectral imagery, sparse representation, computation vision and pattern recognition, and remote sensing images.

Dr. Li is a young Editorial Board Member of Geospatial-Information Science (GSIS). She is a Reviewer of more than 30 international journals, including the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS), the IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), the IEEE TRANSACTIONS ON CYBERNETICS (TCYB), *Remote Sensing of Environment (RSE)*, and *ISPRS Journal of Photogrammetry and Remote Sensing (ISPRS-J)*. She is a Guest Editor of the *Remote Sensing* (an open access journal from Multidisciplinary Digital Publishing Institute (MDPI)) and *Sustainability* (an open access journal from MDPI).