



Construction of Improved Semantic Segmentation Model and Application to Extraction of Anthropogenically Disturbed Parcels With Soil Erosion From Remote Sensing Images

Jialin Li , Li Hua , Lu Li, Zijing Zhang, and Chongfa Cai

Abstract—With the rapid socioeconomic development in China, increasing soil erosion caused by anthropogenic production and construction activities is taking place, which is characterized by short duration, high frequency, and great damage to its surrounding environment. Therefore, the regulation and control of soil erosion of anthropogenically disturbed parcels is an urgent task. This study proposes an improved model that combines the boundary constraint and jagged hybrid dilated convolution channel shuffling module (BCJHDC) and the polarized self-attention (PSA) module for extracting anthropogenically disturbed parcels with soil erosion from high-resolution remote sensing images in Hubei Province. First, the PSA module is added to the encoder to better extract the feature information of the target object. Second, the BCJHDC module is used to extract multiscale semantic information from images and improve the boundary segmentation quality. Precision, recall, intersection over union (IOU), and F1 score (F1) are calculated to evaluate the model accuracy. The results indicate that our improved model performs well on the human-perturbed parcel extraction task, with an F1 of 87.92% and an IOU of 78.44%. Ablation experiments and application experiments suggest the validity of the applicability and the portability of our proposed improved model, respectively. Compared with the other seven advanced semantic segmentation models, our improved model has significant advantages. Overall, this study provides a valuable reference for policy formulation of water and soil conservation.

Index Terms—Anthropogenically disturbed parcels with soil erosion (ADPSE), boundary constraints and jagged hybrid dilated convolution channel shuffling module (BCJHDC), deep learning, remote sensing, self-attention.

I. INTRODUCTION

THE unreasonable development and exploitation of natural resources by human beings have led to the aggravation

of soil erosion, and environmental problems have become increasingly severe. This hinders the sustainable development of social economy and ecological environment to a great degree. The causes of soil erosion mainly include natural factors and human factors. The latter is primarily responsible for serious soil erosion [1]. With the rapid development of urbanization and industrialization in China, anthropogenic disturbances, such as the development of mountains, railways, highways, electric power, economic areas, and agroforestry, as well as town construction, result in increasingly serious soil erosion problems. The problem of soil erosion is still serious in many regions such as the Yangtze River Economic Zone, Loess Plateau, and Northeastern Black Earth Region in China. Therefore, it is urgent to strengthen the supervision of anthropogenically disturbed parcels with soil erosion (ADPSE). However, few studies have been conducted on ADPSE monitoring. At present, specific target identification from remote sensing images is mostly focused on watersheds, municipalities, and counties and relatively less on areas at large scale, such as provinces.

Traditionally, ADPSE supervision usually depends on the visual interpretation method, which directly determines the feature properties and scope of the targets based on some visual interpretation marks of remote sensing images, but this method relies heavily on the professional knowledge of the interpreters and has a high demand on the experience in the field survey. In addition, this visual interpretation method is easily affected by external factors. In recent years, with the continuous development of satellite technology, high-resolution remote sensing images have become the main means of earth observation. Remote sensing technology has shown great advantages in the soil and water conservation supervision of ADPSE, and high-resolution remote sensing images have increasingly become an important source of soil and water conservation information.

Using the object-oriented method, Kang et al. extracted the changes of new bare land and impervious layer from multitemporal high-resolution remote sensing images and removed the pseudochanges to obtain the disturbed areas, achieving more than 90% accuracy [2]. Similarly, using the object-oriented classification method, Tan et al. extracted the disturbed areas by setting thresholds for features such as image texture and spectrum based on the high-resolution remote sensing imagery,

Manuscript received 5 November 2022; revised 5 January 2023; accepted 4 February 2023. Date of publication 13 February 2023; date of current version 13 March 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFD1500703 and in part by the National Natural Science Foundation of China under Grant 41601280. (Corresponding author: Li Hua.)

Jialin Li, Li Hua, Zijing Zhang, and Chongfa Cai are with the College of Resources and Environment, Huazhong Agricultural University, Wuhan 43007, China (e-mail: 1535420728@qq.com; huali@mail.hzau.edu.cn; 3344236235@qq.com; cfcai@mail.hzau.edu.cn).

Lu Li is with the Hubei Water Resources Research Institute and the Hubei Water Conservancy & Hydropower Research Institute, Wuhan 430070, China (e-mail: 420246066@qq.com).

Digital Object Identifier 10.1109/JSTARS.2023.3244209

and in their study, disturbed areas of an underconstruction mine project in Tibet were extracted and analyzed with an extraction accuracy reaching more than 83% [3]. Based on multitemporal high-resolution remote sensing imagery, Nascimento et al. used spectral bands, the normalized difference vegetation index, the normalized difference water index, the light detection and ranging digital terrain model, and a slope map to establish thresholds for distinguishing open-cast mining complexes with an overall accuracy reaching up to 90% [4]. Ali et al. used the coma cap transform and the K-means algorithm to implement image classification in the GIS program and thus identified the surface coal mine area in the study area [5]. He et al. proposed a tree-root algorithm coupled with an extreme learning machine to construct an open-pit mine classification model and achieved a desirable extraction effect, and their method could monitor the changes in the mine areas in real time [6]. Dai et al. proposed a semiautomatic method for extracting rural roads from remote sensing images based on a combination of multiple features, and their method can identify rural roads, especially the rural dirt roads [7]. Based on GF-1 high-resolution remote sensing images, Liu et al. used five remote sensing image fusion algorithms, including high-pass filter transform, principal component transform (PC), Gram–Schmidt (GS, orthogonalization transform), and other two algorithms, and found that of the five algorithms, PC and GS algorithms exhibited higher accuracy, with the coal area extraction accuracy of PC reaching up to 100% and the extraction accuracy of GS reaching over 80% in the mixed areas of bare surface and construction land [8].

Although the above-mentioned methods can achieve good recognition performance, the feature extraction of these methods relies on prior knowledge and parameter settings, which limits their wide application. When these methods are used for remote sensing image classification, the input features undergo relatively less linear or nonlinear transformation, and thus the rich semantic information of ADPSE may not be well captured by these methods. These limitations will affect the robustness and discriminative power of feature extraction, especially for high-resolution remote sensing images with large differences in spectral features and highly complex texture information. The recently emerging deep learning has demonstrated great potential in feature extraction based on end-to-end architecture, and the performance of deep learning has far exceeded that of traditional machine learning in many applications. In fact, deep learning has penetrated various aspects of remote sensing applications, such as road recognition [9], building recognition [10], crop yield estimation [11], and pest and disease identification [12]. Multiple advanced semantic segmentation models, such as U-Net [13], fully convolutional network (FCN) [14], SEGNet [15], PSPNet [16], and Deeplabs [17], [18], have been proposed for the recognition of specific targets from remote sensing images. In the semantic segmentation models described above, a convolutional neural network composed of multiple convolutional layers and pooling layers is used to learn the underlying and high-level features of the target objects and identify different types of features from remote sensing images, and these semantic segmentation models have exhibited desirable performance in many remote sensing image analysis tasks. For

example, Yan et al. used a U-Net network framework based on multifeature fusion receptive and successfully extracted the objects from remote sensing images with a mean intersection over union (IOU) of 87.34% [19]. He et al. introduced edge information as priori knowledge into FCN and used the edge information obtained by the holistic nested edge detection network to correct the results of FCN, and after improvement, the recognition accuracy of forest land was improved from 95.66% to 97.16% [20]. Based on SegNet, Jing et al. selected a random walk classification seed region using a sliding window strategy and optimized the edge weight network by fusing the gradient map and probability map of the original image, thus achieving high-performance semantic segmentation of remote sensing images [21]. Fan et al. proposed a DeepLab-V3 network model in combination with improved U-Net fusion shallow features and solved the sample imbalance problem by constructing a combined loss function integrating dice loss and BCE loss, thus increasing the accuracy of road extraction from remote sensing images [22]. Based on FCN, Mou et al. proposed the RiFCN model, which used forward propagation to generate multilevel feature maps and backward propagation to generate fusion feature maps as the basis for classification, and this model exhibited a good segmentation effect on remote sensing images, but with the continuous replacement of parameters in the model, the number of parameters increased substantially [23]. Du et al. effectively optimized the segmentation results by combining the DeepLabv3+ model with an object-based image analysis, but the network structure of their method was very complex, thus increasing the training cost [24]. In order to acquire target features accurately, attention mechanisms were employed to improve the model. Guo et al. designed a multitask parallel attention convolutional network by integrating CBAM and DA into a semantic segmentation model, which effectively reduced misclassification and improved classification accuracy [25]. Li et al. designed a model combining CNN and CBAM, which also achieved better segmentation results [26].

However, the above-mentioned methods fail to integrate global pixels well, potentially resulting in information loss. In addition, these methods cannot establish a long-distance dependency on high-resolution image input or output features, thus failing to acquire fine features of the destination object well. Large-target segmentation from high-resolution remote sensing images is susceptible to the effects of a scene, illumination, imaging angle, and imaging time [27], [28], thus causing the following possible defects in segmenting large targets. First, the features of the same type of targets vary greatly in different places. For example, the size, shape, and spectral features of ADPSE differ significantly between urban and rural areas. Second, different types of features can easily be confused. For example, a building site being excavated can easily be confused with an already harvested agricultural field. Currently, these existing methods tend to have salt and pepper noise, making it difficult to obtain target segmentation boundaries accurately.

In order to solve the above-mentioned problems, in this study, an improved large-target recognition model combining the boundary constraint and jagged hybrid dilated convolution channel shuffling module (BCJHDC) module and the polarized

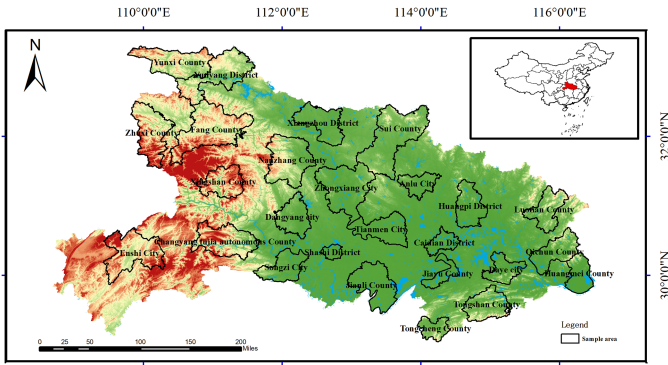


Fig. 1. Study area: Hubei Province.

self-attention (PSA) module is designed for high-resolution remote sensing images. The BCJHDC module is designed to reduce information loss while acquiring a large sensory field so as to improve segmentation boundary accuracy. In addition, the newly proposed PSA module is introduced into the encoder of the model to mine finer and higher quality features, and PSA can maintain relatively high resolution in spectral and spatial dimensions, compared with other attention mechanisms, thus reducing the loss of features. The main contributions of this study are as follows.

- 1) The improved model is proposed to solve the problems of blurred boundaries and low accuracy of large-target extraction from high-resolution remote sensing images, and this model replaces the standard convolution in the encoder with the residual convolution.
- 2) The PSA module is introduced into semantic segmentation models to guide the model to learn pixel-level semantic information so as to obtain segmentation maps with more accurate boundaries.
- 3) To incorporate global context information, a BCJHDC is constructed. This module not only addresses the local information loss caused by dilated convolution but also improves the segmentation accuracy and edge effect of the model for objects at different scales.
- 4) Using this model, we obtain the distribution map of ADPSE in Hubei Province, China, which provides a reference for the policy formulation of soil and water conservation.

The rest of this article is organized as follows. Section II presents data and their preprocessing. Section III elaborates on the methodology. Section IV focuses on the experimental design and the results. Finally, Sections V and VI present conclusions and future works.

II. DATA AND PREPROCESSING

A. Study Area Overview

As shown in Fig. 1, Hubei Province is located in central China. It is between $29^{\circ}01'53''$ — $33^{\circ}6'47''$ N and $108^{\circ}21'42''$ — $116^{\circ}07'50''$ E. Its total area is $185\,900\text{ km}^2$, accounting for 1.94% of China's total area. Hubei Province is located in the

transition zone from the second to third steps of China's terrain, surrounded by mountains on its north, west, and east sides with a low level in the middle. Hubei Province has a variety of landform types, such as mountains, hills, and plains. In recent years, with the economic development of Hubei Province, urbanization development, engineering construction, and mineral resource development have gradually increased, thus aggravating soil erosion caused by anthropogenic disturbance in Hubei Province. Particularly, the area of production and construction-project-induced erosion has gradually increased, thus inevitably damaging the surface vegetation and greatly reducing the soil erosion resistance. Development zone construction, quarrying, road construction, and other urbanization construction are mainly responsible for soil erosion in Hubei Province. The results of the first China Water Census show that Hubei Province has a total area of $36\,900\text{ km}^2$ with soil erosion, which accounts for 19.85% of the total area of the province, making it one of the provinces with serious soil erosion in China.

B. High-Resolution Remote Sensing Image Preprocessing


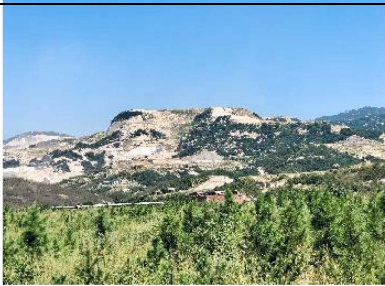




In order to investigate the distribution of ADPSE in Hubei Province, a remote sensing field survey was carried out. The relevant information on different land use types, vegetation coverage, soil and water conservation biological measures, engineering measures, and farming measures was collected at the county scale to establish remote sensing interpretation marks. On the basis of the interpretation marks established by the field survey, the human-computer interaction method was used to plot the anthropogenic disturbance parcels with soil erosion (data obtained from Hubei Water Resources and Hydropower Research Institute). The types of ADPSEs are given in Table I. In this study, remote sensing images of Hubei Province were stitched from 2-m resolution RGB images obtained from GF-6 and Resource 3 satellites from January to May 2021. Image preprocessing, such as alignment, radiometric calibration, atmospheric correction, cropping, and filtering, were performed to obtain more accurate image information. The data on anthropogenically disturbed parcel distribution in Hubei Province were obtained by combining indoor supervision classification and field verification. In this study, 27 counties evenly distributed in Hubei were selected as the sample areas, and the image and label data were cropped to 512×512 size. After data cleaning and removing the images with the ADPSE area accounting for less than 10% in a single image, a total of 8892 images were obtained, of which 60% were used for training, 20% for validation, and 20% for testing.

III. METHOD DESCRIPTION

A. U-Net++ Model

The U-Net++ model consists of three parts: an encoder, a skip connection, and a decoder. The encoder extracts high-resolution image features to provide precise target localization. The U-Net++ model achieves spatial information fusion from the shallow layer to the deep layer by adjusting the long connection of U-Net to a nested dense short connection, namely, connecting

TABLE I
ADPSE REMOTE SENSING INTERPRETATION MARKS

Types of ADPSE	Images	Image Features	Photo
Open-Cast Mining Complexes		Irregular shape, large area, strong spectral reflectivity, bright white in the image, and mixed with mostly hardened ground and bare ground	
Railroad and Highway Engineering		Long bands with neat edges, more spectral texture, homogeneous internal spectrum, and off-white color in the image	
Urban construction projects		Regular shape, mostly composed of bare ground, hardened ground and structures, yellowish brown or bright white, connected to the road, artificial objects inside the construction site	

the high-resolution features with the decoder upsampling output features. Additionally, in the semantic information fusion process, the dense skip connections between the modules at the same layer reduce the information loss in feature extraction and shorten the semantic gap between the encoder and decoder, so that the network model can fully and effectively capture detailed features of the target.

The network topology of the U-Net++ model is shown in Fig. 2, which consists of an encoding structure, a decoding structure, and a dense skip connection. Feature maps of the same size are defined as the same layer, and there are five layers, namely, Layers 0–4 (L0–L4, top-down). Each node represents a feature extraction module, and each feature module consists of two 3×3 convolutional layers; the two convolutional operations are followed by one batch normalization and one rectified linear unit.

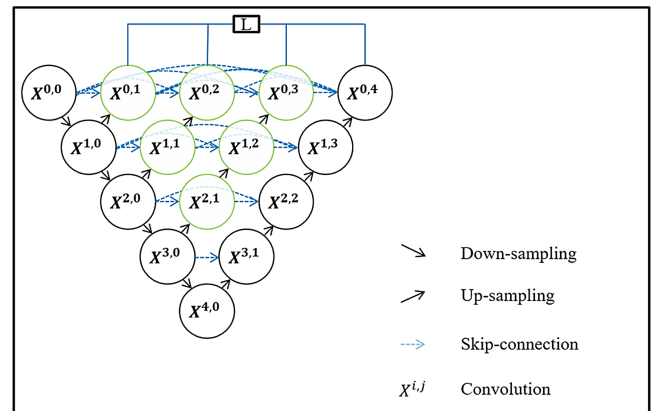


Fig. 2. Network structure of the U-Net++ model.

B. Boundary Constraints and Jagged Hybrid Dilated Convolution Channel Shuffling Module

1) *Hybrid Dilated Convolution (HDC)*: The early deep-learning-based image segmentation method mainly includes three steps. First, the convolution operation is used to extract image features. Then, the resultant image is downsampled to reduce the size of the feature map. Finally, the obtained image is upsampled to the size of the original image for prediction. The steps of first decreasing and then increasing the image size cause the loss of the internal data information, and thus the small target information cannot be reconstructed. To solve this problem, Yu et al. proposed a dilated convolution structure to increase the perceptual field without using a pooling operation [29], thus allowing each convolution output to contain a larger range of information without sacrificing feature space resolution. However, there exists a grid problem in Yu's dilated convolution method. Considering this, Wang et al. proposed the HDC strategy to mitigate the impact of the grid problem by using multiple dilated convolutions with different dilation rates consecutively and alternately [30]. Suppose there are N convolution kernels with the size of $ksize \times ksize$ in the dilated convolution layer with a dilation rate of $\{d_1, \dots, d_i, \dots, d_n\}$, the maximum distance between two non-zero points nodes is defined as follows:

$$L_i = \text{MAX} [L_{i+1} - 2d_i, L_{i+1} - 2(L_{i+1} - d_i), d_i] \quad (1)$$

where L_n indicates the maximum distance (d_n) between two non-zero points. In the HDC design, $L_i \leq ksize$ constant.

a) *PointRender algorithm*: Large-target extraction from remote sensing images focuses more on the boundary information of the target and less on the smoothness of the boundary. Recently, Kirillov et al. proposed a PointRender technique to optimize image segmentation and improve accuracy performance by identifying object edges [31]. PointRender introduces image rendering into semantic segmentation. PointRender uses a bilinear bracket to upsample the fine-grained features from the CNN network and selects multiple uncertain points from the upsampling results as the boundary segmentation difficult points. Furthermore, based on these difficult points obtained directly from the fine-grained features of the CNN, PointRender performs label prediction using a multilayer perceptron (MLP). The process of upsampling, difficult point selection, and label prediction is repeated until the specified resolution is achieved. By implementing adaptive subdivision in the upsampling process, the contour information can be preserved as much as possible, and the boundary points can be identified in the segmentation to improve the segmentation accuracy. Fig. 3 demonstrates the upsampling adaptive subdivision process.

In the PointRender model, the MLP needs to be trained through difficult points selected by the point head. We used a random sampling selection strategy to accommodate the backward propagation of the neural network. The k was the upper limit parameter ($k < 1$). The β was the lower limit parameter ($0 \leq \beta \leq 1$). The selection strategy included the following three steps.

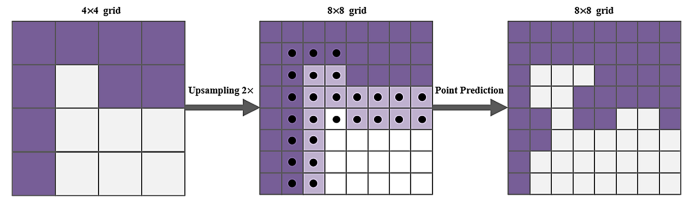


Fig. 3. Upsampling adaptive subdivision process in the PointRender model.

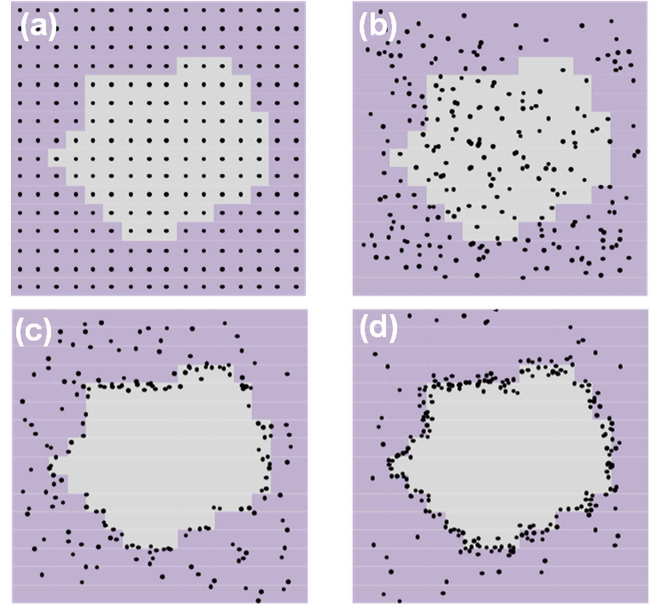


Fig. 4. Training of the PointRender model. (a) Regular grid. (b) Uniform, $k = 1$, $\beta = 0$. (c) Mildly biased, $k = 5$, $\beta = 0.75$. (d) Heavily biased, $k = 10$, $\beta = 0.85$.

- 1) *Random sampling*: We randomly selected kN points from the uniformly distributed points in order to generate more candidate points.
- 2) *Boundary sampling*: For uncertain regions, we interpolated the rough predicted values of all kN points. βN points with the highest uncertainty were selected from the kN candidate points and used to calculate the uncertainty of the task.
- 3) *Residual point sampling*: In the same sampling way as in step 2), the remaining $(1-\beta)N$ points were selected from the uniformly distributed unbounded region, and this three-step selection process is displayed in Fig. 4. With k and β increasing, the distribution of points changes from even distribution in the image to mildly biased and heavily biased toward the boundaries in uncertainty regions.

2) *BCJHDC Module Structure*: The structure of the BCJHDC module is shown in Fig. 5, and the input features are $X \in R^{C \times H \times W}$. The input X is divided into four groups X_1 – X_4 along the channel: $X = [X_1, X_2, X_3, X_4]$, $R^{C/G \times H \times W}$. For each group, feature X_k is divided into two branches X_{k1} and X_{k2} . Different importance coefficients are generated, respectively, by the HDC module and the PointRender module. For the first branch X_{k1} , the dilation rate in the HDC layer was set as $\{1$,

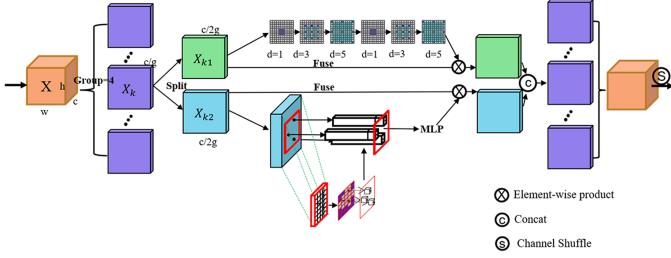


Fig. 5. Structure of the BCJHDC module.

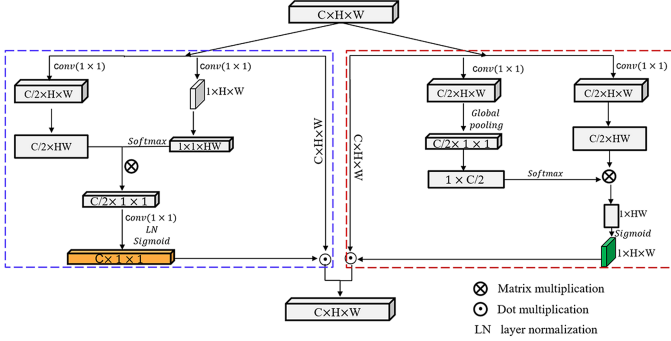


Fig. 6. Parallel PSA mechanism.

2, 5, 1, 2, 5}, and finally two groups of HDC(1, 2, 5) are formed. For the second branch X_{K2} , with k and β increasing, the distribution of points changes from even distribution in the image to severely biased toward the boundary, but the computational burden also increases dramatically. To balance the accuracy and training complexity, we set $k = 5$ and $\beta = 0.75$. After finishing the two-branch calculation, we fuse their calculation results through simple concatenation to obtain X'_k , $X'_k = [X'_{k1}, X'_{k2} \in R^{C/G \times H \times W}]$. Finally, we performed intergroup communication by a channel permutation operation. The final output of the module has the same dimensions as the input, which allows the module to be easily embedded into existing CNN architectures.

C. Self-Attention Module

For the self-attention module, we chose the PSA mechanism due to its following two advantages.

- 1) *Filtering advantage*: The PSA module can maintain high resolution in channel and spatial attention calculations while fully folding the input tensor along the corresponding dimension.
- 2) *Enhancement advantage*: The PSA module can output nonlinear functions of distribution.

The PSA mechanism can be instantiated as the following module (see Fig. 6), which consists of channel-only self-attention (in the blue dotted box) and spatial-only self-attention (in the red dotted box).

Channel-only self-attention O^{ch} (the orange block in Fig. 6) can be calculated according to the following equation:

$$O^{\text{ch}}(x) = \text{Sigmoid}[\text{LN}(\text{Conv}(R_1(\text{conv}(x)))$$

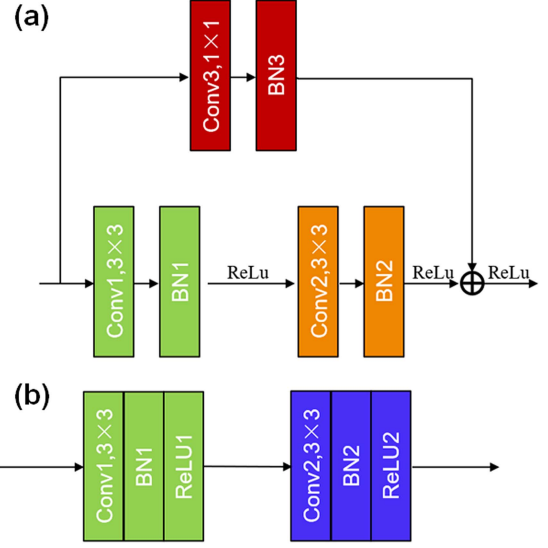


Fig. 7. (a) Structure of the residual block. (b) Structure of the VGG block.

$$\times \text{Softmax}(R_2(\text{conv}(x))))]. \quad (2)$$

Spatial-only self-attention G^{sp} (the green block in Fig. 6) can be calculated according to the following equation:

$$G^{\text{sp}}(x) = \text{Sigmoid}[R_3(\text{softmax}(R_1(\text{GP}(\text{conv}(x))) \times R_2(\text{conv}(x))))] \quad (3)$$

where Conv is 1×1 convolution operator; R_1 , R_2 , and R_3 are three tensor reshape operators; GP is a global pooling operator; LN is a layer normalization; and x is the matrix multiplication.

The final output equation of the parallel PSA is as follows:

$$\text{PSA}_p(x) = O^{\text{ch}} \odot^{\text{ch}} x + G^{\text{sp}} x \quad (4)$$

where \odot^{ch} is a channel-wise multiplication operator, and \odot^{sp} is a spatial-wise multiplication operator.

D. Improved Model Structure

The residual structure is shown in Fig. 7(a). In the improved model, the encoder uses multiple convolution operations to extract features. With the number of layers of the network increasing, the network training becomes difficult, and the feature extraction ability is insufficient. The improved model has a deeper network depth due to the addition of PSA and BCJHDC modules, and thus a residual block is added to the encoder for extracting more semantic information from the image so as to improve the feature extraction ability and prevent the gradient from disappearing. As shown in Fig. 7(b), dense skip connection and decoders use VGG16-based standard convolution and the max pooling layer.

As shown in Fig. 8, there are residual convolution blocks at encoder Layers 1–4 (X_{00} , X_{10} , X_{20} , and X_{30}) in the improved model network. Each of layer X_{00} , X_{10} , X_{20} , and X_{30} is connected to a PSA module whose function is to focus on the key information in the image and filter irrelevant information. At the end of the encoder (layer X_{40}) is a BCJHDC for multiscale

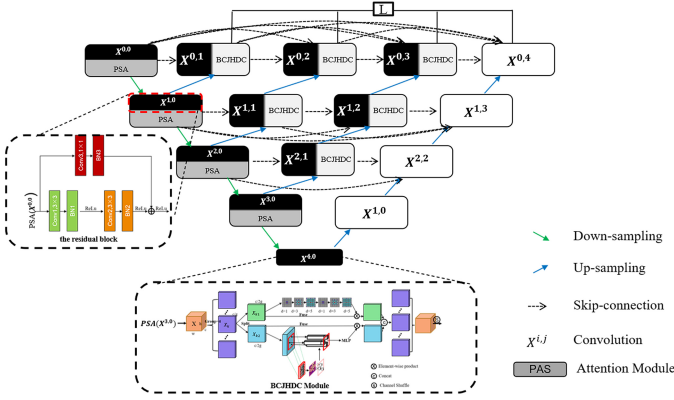


Fig. 8. Structure of the improved model. The green arrow indicates down-sampling, and the blue arrow indicates upsampling. The black dashed arrow represents the concatenation operation. The U-shaped network is mainly used for deep feature extraction and upsampling, with the encoder on the left border and the decoder on the right border.

extraction of image features and encoding of global contextual information. The feature images at each layer obtained from the encoder of the U-Net++ network need to pass through dense convolution blocks before being received by the decoder. The skip connection platform fuses the feature images from different layers and feeds them into a 3×3 standard convolution module before inputting them into the decoder. In order to better guide the model to learn pixel-level semantic information, the global contextual information is integrated, classification boundaries are constrained, and BCJHDC modules are added to all skip connection layers. The improved model combines the advantages of U-Net++, the attention mechanism, and the BCJHDC module to enhance the feature extraction capability of the network, thus enabling the network to acquire high-level semantic information while focusing on detailed information, eventually improving the target extraction capability and optimizing the edge effect of segmentation.

The structure of the encoder and decoder is shown in Table II.

IV. EXPERIMENTS AND RESULTS

The experimental environment is windows 10. The running framework is PyTorch 1.7.1, and the processor is Intel(R) Core(TM) i9-12900k CPU@3.20 GHz. The graphics card is NVIDIA GeForce GTX 3090, and the programming environment is Python 3.7.

A. Experimental Design

1) *Loss Function*: The loss function is used to minimize the error between the output of the segmentation network and the labels so as to obtain an optimal model. In remotely sensed images, the sizes of ground elements vary from large to small, and these elements are not evenly distributed, thus causing the learning process to fall into the local minima of the loss function. Consequently, the prediction results of the network tend to be biased toward the elements with a larger proportion. Due to the large proportion of background in the dataset in this study,

TABLE II
ENCODER AND DECODER STRUCTURES

Encoder Structure		Decoder Structure	
$X_{0,0}$	$\begin{cases} \text{Conv3-32} \\ \text{Conv3-32} \\ \text{Conv1-32} \end{cases}$ Stride=1 512× 512	$X_{0,4}$	$\begin{cases} \text{Conv3-32} \\ \text{Conv3-32} \end{cases}$ 512× 512
$X_{1,0}$	$\begin{cases} \text{Conv3-64} \\ \text{Conv3-64} \\ \text{Conv1-64} \end{cases}$ Stride=1 256× 256	$X_{0,3}$	$\begin{cases} \text{Conv3-64} \\ \text{Conv3-64} \end{cases}$ 256× 256
$X_{2,0}$	$\begin{cases} \text{Conv3-128} \\ \text{Conv3-128} \\ \text{Conv1-128} \end{cases}$ Stride=1 128× 128	$X_{0,2}$	$\begin{cases} \text{Conv3-128} \\ \text{Conv3-128} \end{cases}$ 128× 128
$X_{3,0}$	$\begin{cases} \text{Conv3-256} \\ \text{Conv3-256} \\ \text{Conv1-256} \end{cases}$ Stride=1 64× 64	$X_{0,1}$	$\begin{cases} \text{Conv3-256} \\ \text{Conv3-256} \end{cases}$ 64× 64
$X_{4,0}$	BCJHDC-256		Conv1-256

we used a mixture of the cross-entropy loss [32] and the dice loss [33] as loss function to address the problem of unbalanced category shares. This hybrid function [34] is defined as follows:

$$L(Y, P) = -\frac{1}{N} \sum_{c=1}^C \sum_{n=1}^N \left(y_{n,c} \log p_{n,c} + \frac{2y_{n,c}p_{n,c}}{y_{n,c}^2 + p_{n,c}^2} \right) \quad (5)$$

where $y_{n,c} \in Y$ and $p_{n,c} \in P$ are the target labels and predicted probabilities of the c th type and n th pixel in one batch, respectively; Y and P are the actual values and predicted values of ADPSE, respectively, and C and N denote the number of types and pixels in the dataset in one batch, respectively.

2) *Implementation details*: At the beginning of training, the weights of the model are initialized randomly. If a large learning rate is chosen, the model may be unstable. To stabilize the model, a small warm-up learning rate will be chosen in the first few epochs or some steps. In this study, we selected the warm-up strategy in the first five epochs and then used the ‘‘poly’’ learning rate strategy. The learning rate is calculated according to the following formula:

$$\text{lr} = \text{base}_{\text{lr}} \times \left(1 - \frac{\text{epoch}}{\text{number}_{\text{epoch}}} \right)^{\text{power}}. \quad (6)$$

The parameter setting of model training was as follows. The base learning rate is set as 0.01, and power is set as 0.9. The number of epochs is set as 200. Momentum and weight decay are set as 0.9 and 0.0003, respectively. The training batch size is set as 8.

3) *Evaluation Index*: Four indicators, including IOU, recall, precision, and F1-score (F1), are used for evaluating the prediction results of the network. IOU is the rate of the intersection and union between the prediction values of each type and the actual value. Recall is the rate of correctly predicted target features for

TABLE III
IOU AND F1 OF DIFFERENT BATCH SIZES

Batch size of our improved model	IOU (%)	F1 (%)
4	77.91	87.58
8	78.44	87.92
12	78.26	87.81

all actual target features, and precision is the rate of correctly predicted target features for all predicted target features. These four indicators were calculated using the following formulae:

$$\text{IOU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \quad (7)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (8)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (9)$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

where TP is the pixel number of correctly predicted ADPSE; TN is the pixel number of correctly predicted background; FP is the pixel number of incorrectly predicted ADPSE; and FN is the pixel number of incorrectly predicted background.

B. Experiments Results

BatchSize is the size of the batch for each training epoch. We designed experiments to verify the effect of batch size on the model. Thus, to optimize the design of parameters for the model, two indicators, including IOU and F1, are used for evaluating the prediction results of the model. Table III presents a comparison of the segmentation evaluation metrics of different batch sizes. Within a reasonable range, the bigger the batch size is, the higher the IOU and F1 are. But when the batch size is 12, the IOU and F1 drop slightly. Although a batch size of 8 is longer than 4 for the iteration time, IOU and F1 improved by 0.53% and 0.34%, respectively. Because the goal of this study is to accurately identify the ADPSE, it was worth it to take longer to train.

To validate the effectiveness of the improved model, we tested the data from the study area. In addition, we compared our improved model with other advanced semantic segmentation models, including DeepLabv3+ [22], PSPNet [20], U-Net++ [35], DANet [36], DenseASPP [37], TransUNet [38], and BiSENetV2 [39]. These models provided references for evaluating the performance of our improved models. Table IV shows a comparison of the segmentation evaluation metrics of these models.

The U-Net++ (VGG-block) model has a recall of 84.61%, precision of 80.46%, F1 of 82.48%, and IOU of 70.19%. Its precision is lower than other models, but its IOU, recall, and F1 are higher than those of DeepLabv3+ and PSPNet. Both

TABLE IV
IOU, RECALL, PRECISION, AND F1 OF DIFFERENT MODELS

Models	IOU (%)	Recall (%)	Precision (%)	F1 (%)
DeepLabv3+	68.13	79.54	82.61	81.05
PSPNet	69.33	82.65	81.14	81.89
U-Net++ (VGG-block)	70.19	84.61	80.46	82.48
DANet	73.84	83.73	86.21	84.94
DenseASPP	72.01	85.24	82.26	83.72
TransUNet	74.20	85.73	84.65	85.19
BiSENetV2	73.61	86.73	82.94	84.79
Improved model (in this study)	78.44	87.47	88.38	87.92

DANet and DenseASPP are end-to-end models, and they are added respectively with dual attention and densely connected atrous spatial pyramid pools. TransUNet achieves high accuracy by combining transformer and UNet networks, but its four evaluation metric values are slightly lower than those of our improved model. Of the eight models, BiSENetV2 achieves the highest recall value, but its precision is relatively lower. Our improved model outperforms the classical U-Net++ model with the precision increased by 6.64%, recall increased by 2.86%, F1 increased by 5.42%, and IOU increased by 8.25%. Overall, our improved model raises the recognition accuracy of ADPSE and achieves better results in all four metrics than the other seven models.

Fig. 9 shows the results of five randomly selected scenes in the test datasets, which are predicted by eight different segmentation models.

For scene 1, all the models segmented the target features basically completely, but our improved model identified the boundaries of the target features more accurately, and the predicted results were closer to the real scene. In addition, the edge effect segmented by our improved model was optimized by the BCJHDC module, thus expanding the receptive field, enriching semantic information, and improving the segmentation accuracy of objects. There were fewer small holes and adhesions in the results predicted by the improved model, thus resulting in more complete segmentation of large targets.

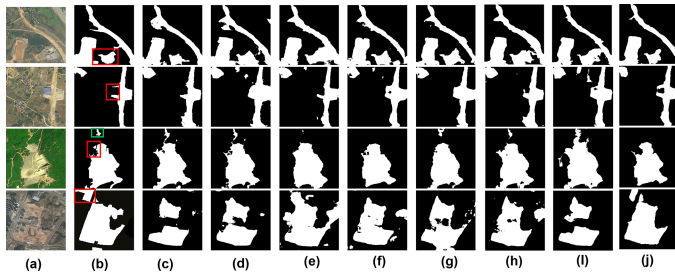


Fig. 9. Segmentation results by different models on the test dataset. (a) Remote sensing images. (b) Ground truth. (c) Results predicted by DeepLabv3+. (d) Results predicted by PSPNet. (e) Results predicted by U-Net++. (f) Results predicted by DANet. (g) Results predicted by DenseASPP. (h) Results predicted by TransUNet. (i) Results predicted by BiSENetV2. (j) Results predicted by our improved model.

For scene 2, only DeepLabv3+, U-Net++, TransUNet, and our improved model identified the buildings (in the red box in Fig. 9), whereas the other four models did not. In addition, our improved model segmented the boundaries more accurately. For scene 3, among eight models, only our improved model segmented the vegetation (in the red box), but it did not identify the targets in the green box.

For scene 4, it is difficult to identify the targets in the red box, but the improved model identified them accurately. The possible reason may be that the improved model combines the attention module and the BCJHDC module. The attention module can extract the target feature more finely, and the BCJHDC module can employ multiscale and contextual information to obtain more accurate segmentation results.

Appropriately increasing the model depth enables the model to extract deep semantic features, thus improving the segmentation accuracy. However, the parameter number and the inference time of the model will keep increasing with increasing model depth. Therefore, there needs to be a tradeoff between accuracy and complexity. Table V shows the comparison of parameter number and inference time between the improved model and other models.

As can be seen from Table V, the parameter number of the improved model was only 1.09 MB higher than the smallest one, and the inference time of our model was only 2.84 s longer than the shortest one. The inference time of our improved model was only longer than that of UNet++ and BiSENetV2, but it was much shorter than that of the other five semantic segmentation models. From the results presented in Tables IV and V, it is clear that our improved model increased segmentation efficiency while improving segmentation accuracy.

C. Ablation Experiment

An ablation experiment was conducted to investigate the effect of the BCJHDC module, the PSA module, and residual convolution on the improved model and the action mechanism of the individual modules in the network. Experiments with test dataset were conducted, and Res-UNet++ was used as a baseline model with each of the above-mentioned three modules added gradually to form different combinations. Table VI

TABLE V
COMPARISON OF MODEL PARAMETER NUMBER AND INFERENCE TIME

Models	Parameters (MB)	Inference time(s)
DeepLabv3+	59.34	37.95
PSPNet	62.97	35.15
U-Net++ (VGG-block)	9.04	32.42
DANet	49.49	42.78
DenseASPP	10.20	56.24
TransUNet	105.32	62.13
BiSENetV2	17.04	31.76
Improved model (in this study)	10.13	34.60

shows the detailed values of precision, recall, F1, and IOU under different combinations of improvement strategies. As a baseline model, the original U-Net++ provides reference values of precision, recall, F1, and IOU.

When residual convolution was added, the recall reflecting extraction capability increased by 0.21%, but precision decreased by 0.09%. This result suggested that this combination Res-UNet++ was likely to identify uncertain regions as ADPSE. The addition of the PSA module to Res-UNet++ increased IOU, precision, and F1 by 1.86%, 3.07%, and 1.25%, respectively, indicating that the PSA module contributed to an increase in the accuracy of the obtained target feature information. The addition of the BCJHDC module to Res-UNet++ increased IOU and F1, which were 1.46% and 1.25% higher than those of the PSA module added to Res-UNet++, respectively. The addition of both the PSA module and the BCJHDC module to Res-UNet++ achieved the best performance.

Figs. 10 and 11 show the segmentation results of roads and construction sites predicted by different combination strategies, respectively. As shown in Fig. 10(c), there were some missing ADPSE regions in the results predicted by the original U-Net++, indicating the weak extraction capability of the U-Net++ model. The U-Net++ model tended to identify uncertain elements as background. The addition of residual convolution increased the extraction capability of the model but decreased the extraction precision. As shown in Fig. 10(d), the missing ADPSE regions in Fig. 10(c) were identified, but

TABLE VI
IOU, RECALL, PRECISION, AND F1 OF DIFFERENT COMBINATIONS

UNet++ (Baseline)	Residual block	PSA	BCJHDC	IOU	Recall	Precision	F1
√				70.19	84.61	80.46	82.48
√	√			70.25	84.82	80.37	82.54
√	√	√		72.11	84.16	83.44	83.79
√	√		√	73.57	85.93	83.65	84.78
√	√	√	√	78.44	87.47	88.38	87.92

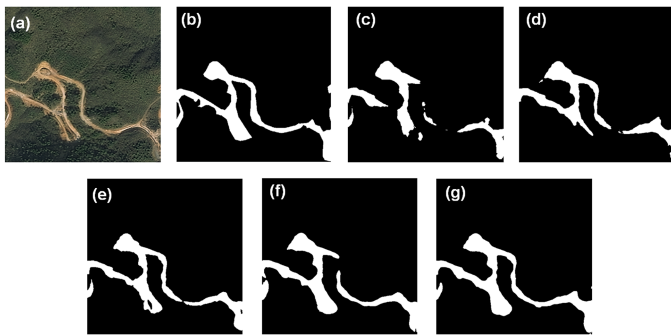


Fig. 10. Ablation experiment segmentation results of roads under construction. (a) Remote sensing image. (b) Ground truth. (c) Results predicted by U-Net++. (d) Results predicted by Res-UNet++. (e) Results predicted by the PSA module added to Res-UNet++. (f) Results predicted by the BCJHDC module added to Res-UNet++. (g) Results predicted by our improved model combining the PSA module and the BCJHDC module.

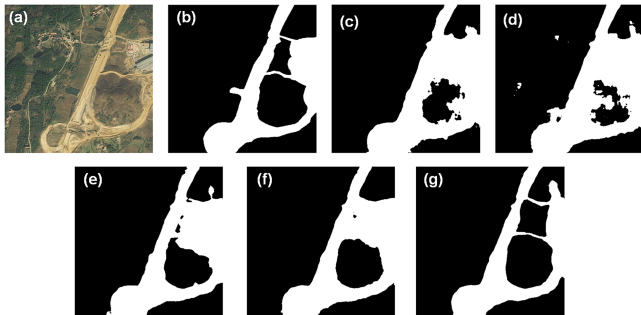


Fig. 11. Ablation experiment segmentation results of construction sites. (a) Remote sensing image. (b) Ground truth. (c) Results predicted by U-Net++. (d) Results predicted by Res-UNet++. (e) Results predicted by the PSA module added to Res-UNet++. (f) Results predicted by the BCJHDC module added to Res-UNet++. (g) Results predicted by our improved model combining the PSA module and the BCJHDC module.

the identification of ADPSE regions was still incomplete with a large number of misidentified pixels. After the PSA module was added, the identification of ADPSE was more complete, but the edge identification was less accurate, and some edges were jagged, indicating that the PSA module addition enhanced the

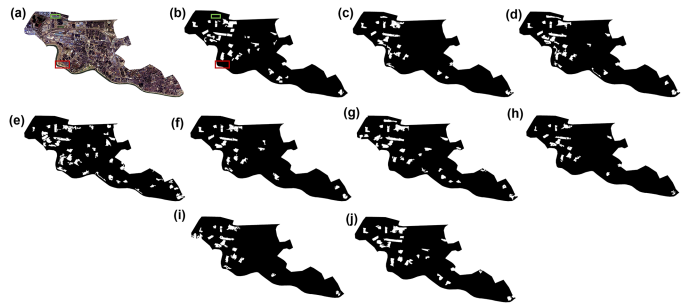


Fig. 12. Application experiment results of another dataset predicted by different models. (a) Remote sensing image. (b) Ground truth. (c) Results predicted by DeepLabv3+. (d) Results predicted by PSPNet. (e) Results predicted by U-Net++. (f) Results predicted by DANet. (g) Results predicted by DenseASPP. (h) Results predicted by TransUNet. (i) Results predicted by BiSENetV2. (j) Results predicted by our improved model.

target extraction capability of the model. When the BCJHDC module was added, some small ADPSE regions were unrecognized, but the fitting of boundaries improved and the jaggedness effect was significantly reduced. When the residual convolution, PSA module, and BCJHDC module were added simultaneously, the optimal completeness and accuracy of the extraction were achieved. Similarly, as shown in Fig. 11, the simultaneous addition of residual convolution, PSA, and BCJHDC resulted in the optimal extraction effect of ADPSE.

D. Application Experiment

In order to verify the applicability and portability of the model, an off-site application test was carried out in Hubei Province. As shown in Fig. 12, among eight models, U-Net++ exhibited the lowest ADPSE recognition precision and the roughest boundaries, whereas the DeepLabv3+ and PSPNet models displayed poorer identification completeness than other models. In the green rectangle of Fig. 12, the actual target object was an impermeable layer, but it was misidentified by DeepLabv3+, PSPNet, and DenseASPP. In the red rectangle of Fig. 12, the actual target object was the exposed riverbank, but it was misidentified by PSPNet, U-Net++, DANet, and BiSENetV2. Although TransUNet correctly identified the target objects in

TABLE VII
IOU, RECALL, PRECISION, AND F1 OF DIFFERENT MODELS

Models	IOU (%)	Recall (%)	Precision (%)	F1 (%)
DeepLabv3+	53.46	68.91	70.46	69.68
PSPNet	54.50	71.71	69.47	70.57
UNet++ (VGG-block)	55.00	73.09	68.97	70.97
DANet	59.59	73.25	76.17	74.68
DenseASPP	57.65	75.12	71.26	73.13
TransUNet	59.43	74.39	74.72	74.55
BiSENetV2	58.79	74.72	73.39	74.05
Improved model (This work)	62.64	75.20	78.95	77.03

red and green rectangles, it failed to identify more other target objects, compared with the above-mentioned models. Taken together, in the application experiments, our improved model exhibited better extraction performance with higher recognition accuracy and better boundary fitting.

As shown in Table VII, the IOU of our improved model was 3%–9% higher than that of the seven other models. The F1 of our improved model was about 2%–7% higher than that of seven other models, indicating the applicability and portability of our improved model.

E. Map of ADPSE in Hubei Province

The distribution map of ADPSE in Hubei Province was plotted. There are a total of 103 county-level administrative regions in Hubei Province. However, in this map, the data of nine counties and cities were missing, including Lichuan City, Jiansi County, Badong County, Xuanen County, Laifeng County, Xianfeng County, Hefeng County, Yiling District, and Zigui County, due to the lack of remote sensing images. Generally, the distribution of ADPSE showed a convergence or dispersion pattern.

As shown in Fig. 13, ADPSE is mainly distributed in the eastern and central regions of Hubei Province, which might be mainly due to the sharp increase in construction land in urbanization progression in these regions and the rapid development of some cities and areas such as Wuhan City, Ezhou City, Huangshi City, and Xianning City. The distribution of ADPSE is more

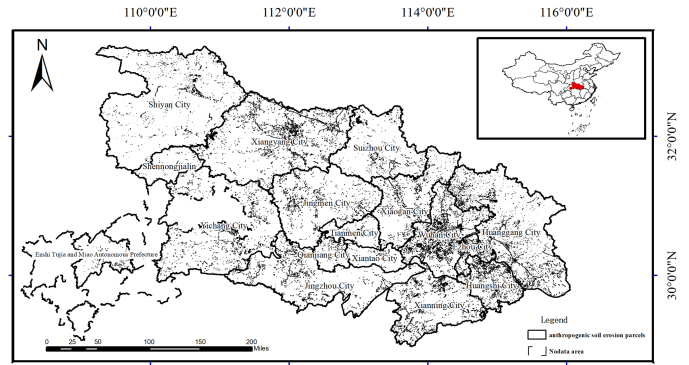


Fig. 13. Spatial distribution of ADPSE.

concentrated in the northern part of Hubei Province, which might be attributed to the construction of provincial highway 316 and the expansion of mining areas in some regions such as Laohekou County, Zaoyang City, Sui County, Guangshui City, and Dawu County. In mountainous western regions with relatively greater development difficulty, such as Shennongjia forest area and Shiyan City, ADPSEs are less distributed. Overall, with the accelerated urbanization in Hubei Province, the significant increase in production and construction projects has led to a much larger area of anthropogenically disturbed land in and around cities than in rural regions. An improved model is proposed, and an ADPSE distribution map is plotted to facilitate the macroregulation and control of ADPSE.

V. DISCUSSION

Compared with other semantic segmentation models, our proposed improved model has two obvious advantages in identifying ADPSE. First, our improved model has the advantage of excluding noise. This might be due to the addition of the PSA module, which enhances feature extraction by modeling the channels and space. Additionally, the PSA module can mitigate the effects of the same spectrum of different targets or the same target with different spectra in the ADPSE identification process. Second, the improved model can segment targets more completely, which might be because the BCJHDC module implements the weighted feature combination of global contextual information and image edge segmentation, thus improving target localization in high-resolution remote sensing images and refinement recovery of high-resolution features.

Although the proposed improved model outperforms the other seven advanced models, there are some problems with our model. For example, buildings within the ADPSE are not well identified, and some types of the ADPSE cannot be identified accurately either. This might be due to the time-sensitive nature of remote sensing imagery. For example, vegetation growth in the ADPSE area in spring and summer adversely affects target identification.

The missing detection in the semantic segmentation task can be minimized by using multispectra for target feature segmentation and optimizing the sample dataset [40], [41]. A high-quality sample dataset plays a crucial role in model training of the

semantic segmentation task [42], [43]. Although the advantages of the proposed improved model have been confirmed in various experiments, the accuracy of the application experiments is relatively lower than that of the test experiments. The possible reasons for this might be as follows. 1) ADPSEs in different regions have different physical structures and exhibit different characteristics in high-resolution remote sensing images. The current training dataset is not abundant enough to cover all types of ADPSE. 2) For the sake of data integrity, data of each county and city in Hubei Province are fused with different data sources, resulting in differences in the spectra between different regions.

VI. CONCLUSION

In this study, an improved semantic segmentation model is proposed by combining BCJHDC and PSA modules for ADPSE recognition from high-resolution remote sensing images. The PSA module helps the model obtain more fine-grained and accurate semantic information on targets and contexts, and the BCJHDC module helps the model learn multiscale information and optimize segmentation edges. In addition, the addition of residual convolution to the improved model effectively avoided gradient loss. The experiments confirmed the validity of the improved model. We compared our improved model with seven other advanced semantic segmentation models, reaching the following conclusions.

- 1) Our improved model outperformed the other mainstream models in identifying ADPSE from remote sensing images with IOU and F1 values increased by 3%–10%.
- 2) The results of the ablation experiments showed that the BCJHDC model could expand the receptive field, refine segmentation results, and optimize boundary constraints.
- 3) The PSA module can obtain finer feature information, and the BCJHDC module can capture global contextual information. Thus, the improved model added with these two modules possesses a better segmentation effect.

However, ADPSE is easily confused with other targets in the process of target extraction, and the structure of our model is relatively complex with a large parameter number. Therefore, future work is suggested to combine multispectrum remote sensing images for target feature extraction, trim the model, and replace residual convolution with deep separable convolution to reduce the network parameters.

REFERENCES

- [1] P. V. G. Batista, J. Davies, M. L. N. Silva, and J. N. Quinton, "On the evaluation of soil erosion models: Are we doing enough?," vol. 197, 2019, Art. no. 102898.
- [2] R. Kang, M. Shi, and Y. Zhao, "Extraction of information on the distribution of production construction projects during the construction period based on multi-temporal GF-1 images," *Bull. Soil Water Conservation*, vol. 36, pp. 253–257, 2016.
- [3] J. J. Tan, Y. Bi, and C. Zhou, "Remote sensing monitoring method and supervision application for disturbed surface of production and construction projects in Tibet," *Soil Water Conservation China*, vol. 2, pp. 63–66, 2019.
- [4] F. S. Nascimento et al., "Land cover changes in open-cast mining complexes based on high-resolution remote sensing data," *Remote Sens.*, vol. 12, 2020, Art. no. 611.
- [5] N. Ali et al., "Remote sensing for surface coal mining and reclamation monitoring in the central salt range, Punjab, Pakistan," *Sustainability*, vol. 14, 2022, Art. no. 9835.
- [6] D. He et al., "Coal mine area monitoring method by machine learning and multispectral remote sensing images," *Infrared Phys. Technol.*, vol. 103, 2019, Art. no. 103070.
- [7] J. Dai, R. Ma, and H. Ai, "Semi-automatic extraction of rural roads from high-resolution remote sensing images based on a multifeature combination," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Oct. 2022, Art. no. 3000605.
- [8] E. J. Liu et al., "Extraction of disturbed plots by different fusion methods for GF-1 data," *J. Soil Water Conservation*, vol. 32, pp. 358–363, 2018.
- [9] S. G. Kanakaraddi, A. K. Chikaraddi, B. L. Pooja, and T. Preeti, "Detection of roads in satellite images using deep learning technique analysis and applications," in *ICT Analysis and Applications*. Singapore: Springer, 2021, pp. 441–451.
- [10] G. Prathap and I. Afanasyev, "Deep learning approach for building detection in satellite multispectral imagery," in *Proc. Int. Conf. Intell. Syst.*, 2018, pp. 461–465.
- [11] D. Elavarasan and P. M. Durairaj Vincent, "Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications," *IEEE Access*, vol. 8, pp. 86886–86901, May 2020.
- [12] N. Ullah, J. A. Khan, L. A. Alharbi, A. Raza, W. Khan, and I. Ahmad, "An efficient approach for crops pests recognition and classification based on novel DeepPestNet deep learning model," *IEEE Access*, vol. 10, pp. 73019–73032, Jul. 2022.
- [13] L. Wang et al., "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogrammetry*, vol. 190, pp. 196–214, 2022.
- [14] G. Chen et al., "SDFCNv2: An improved FCN framework for remote sensing images semantic segmentation," *Remote Sens.*, vol. 13, 2021, Art. no. 4902.
- [15] L. Weng et al., "Water areas segmentation from remote sensing images using a separable residual SegNet network," *ISPRS Int. J. Geo-Inf.*, vol. 9, 2020, Art. no. 256.
- [16] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.
- [17] 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.
- [18] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.
- [19] C. Yan, X. Fan, J. Fan, and N. Wang, "Improved U-Net remote sensing classification algorithm based on multi-feature fusion perception," *Remote Sens.*, vol. 14, 2022, Art. no. 1118.
- [20] C. He, S. Li, D. Xiong, P. Fang, and M. Liao, "Remote sensing image semantic segmentation based on edge information guidance," *Remote Sens.*, vol. 12, 2020, Art. no. 1501.
- [21] J. Jiang, C. Lyu, S. Liu, Y. He, and X. Hao, "RWSNet: A semantic segmentation network based on SegNet combined with random walk for remote sensing," *Int. J. Remote Sens.*, vol. 41, pp. 487–505, 2020.
- [22] H. Fan, Q. Wei, D. Q. Shu, Y. Li, L. Zhang, and C. D. Yang, "An improved Deeplab based model for extracting cultivated land information from high definition remote sensing images," in *Proc. IEEE Int. Conf. Signal Inf. Data Process.*, 2019, pp. 1–6.
- [23] L. Mou and X. X. Zhu, "RiFCN: Recurrent network in fully convolutional network for semantic segmentation of high resolution remote sensing images," 2018, *arXiv:1805.02091*.
- [24] S. Du, B. Liu, and X. Zhang, "Incorporating DeepLabv3+ and object-based image analysis for semantic segmentation of very high resolution remote sensing images," *Int. J. Digit. Earth*, vol. 14, pp. 357–378, 2021.
- [25] H. Guo, Q. Shi, B. Du, L. Zhang, D. Wang, and H. Ding, "Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4287–4306, May 2021.
- [26] H. Li, K. Qiu, L. Chen, X. Mei, L. Hong, and C. Tao, "SCAttNet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 905–909, May 2021.
- [27] W. Boullila, "A top-down approach for semantic segmentation of big remote sensing images," *Earth Sci. Inform.*, vol. 12, pp. 295–306, 2019.
- [28] B. Chen, M. Xia, and J. Huang, "MFANet: A multi-level feature aggregation network for semantic segmentation of land cover," *Remote Sens.*, vol. 13, no. 4, pp. 731–751, 2021.

- [29] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.
- [30] P. Wang et al., "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2015, pp. 1451–1460.
- [31] A. Kirillov, K. He, Y. Wu, and R. Girshick, "PointRend: Image segmentation as rendering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9799–9808.
- [32] P. T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinfeld, "A tutorial on the cross-entropy method," *Ann. Operations Res.*, vol. 134, pp. 19–67, 2005.
- [33] N. Anuar and A. B. M. Sultan, "Validate conference paper using dice coefficient," *Comput. Inf. Sci.*, vol. 3, 2010, Art. no. 139.
- [34] X. Li, M. He, H. Li, and H. Shen, "A combined loss-based multiscale fully convolutional network for high-resolution remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jul. 2022, Art. no. 8017505.
- [35] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput., Comput. Assist. Intervention*, 2018, pp. 3–11.
- [36] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.
- [37] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3684–3692.
- [38] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.
- [39] C. Yu et al., "BiSeNet V2: Bilateral network with guided aggregation for real-time semantic segmentation," *Int. J. Comput. Vis.*, vol. 11, pp. 3051–3068, 2021.
- [40] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 60–77, 2018.
- [41] L. P. Osco et al., "Semantic segmentation of citrus-orchard using deep neural networks and multispectral UAV-based imagery," *Precis. Agriculture*, vol. 22, pp. 1171–1188, 2021.
- [42] J. E. Ball, D. T. Anderson, and C. S. Chan Sr., "Comprehensive survey of deep learning in remote sensing: Theories, tools, and challenges for the community," *J. Appl. Remote Sens.*, vol. 11, pp. 52–68, 2017.
- [43] Q. Zhang, Y. Guang, and G. X. Zhang, "Collaborative network for super-resolution and semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Aug. 2022, Art. no. 4404512.



Jialin Li received the B.S. degree in geographic information science in 2019 from Huazhong Agricultural University, Wuhan, China, where he is currently working toward the M.S. degree in agricultural engineering and information technology.

His research interests include remote sensing and deep learning.



Li Hua received the B.S. and M.S. degrees in surveying and mapping engineering from Wuhan University, Wuhan, China, in 1999 and 2003, respectively, and the Ph.D. degree in soil and water conservation from Huazhong Agricultural University, Wuhan, China, in 2013.

She is currently a Professor with Huazhong Agricultural University. In recent years, she has presided over many provincial and national projects, including the National Natural Science Foundation of China, National Science and Technology Major Project, Yangtze River Basin Soil Erosion Dynamic Monitoring Project, etc. Her research interests include soil erosion dynamic monitoring and driving mechanism, crop monitoring, environmental monitoring, etc.



Lu Li received the B.S. degree in geographic information system and the Ph.D. degree from Huazhong Agricultural University, Wuhan in 2005 and the second Ph.D. degree in information engineering of resources & environment from Huazhong Agricultural University, Wuhan in 2010.

She is the Director of the Institute of Water Ecology and Soil and Water Conservation, Hubei Provincial Research Institute of Water Resources and Hydropower, Wuhan, China, and also a Senior Engineer.



Zijiang Zhang received the B.S. degree in geographic information science in 2020 from Huazhong Agricultural University, Wuhan, China, where she is currently working toward the M.S. degree in resource and environmental information engineering.

Her research interests include agricultural remote sensing and GIS.



Chongfa Cai received the B.S., M.S., and Ph.D. degrees in pedology from Huazhong Agricultural University, Wuhan, China in 1983, 1986, and 1998, respectively.

He is a Professor and Doctoral Supervisor. He has done a lot of research on soil erosion and soil conservation, mainly on the erosion mechanism, process, and prediction of major soils in the south, soil structure and erosion interrelationship, soil erosion prediction and regional hazard evaluation based on GIS and remote sensing, regional development and soil erosion prevention and control, arable land quality evaluation and management, agricultural surface source pollution control, the structure and function of a slope agroforestry complex system, etc. He has conducted more than 20 national research projects, including key projects of the National Natural Science Foundation of China and special projects of science and technology. He has published more than 230 papers, including 78 SCI-indexed papers, in *Soil & Tillage Research*, *Geoderma*, *Agriculture, Ecosystems & Environment*, *Soil Science Society of America Journal*, *Journal of Hydrology*, and other important academic journals at home and abroad.