Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.049187

ARTICLE

CrossFormer Embedding DeepLabv3+ for Remote Sensing Images Semantic Segmentation

Qixiang Tong, Zhipeng Zhu, Min Zhang, Kerui Cao and Haihua Xing*

School of Information Science and Technology, Hainan Normal University, Haikou, 571158, China

*Corresponding Author: Haihua Xing. Email: xinghaihua@hainnu.edu.cn

Received: 29 December 2023 Accepted: 14 March 2024 Published: 25 April 2024

ABSTRACT

High-resolution remote sensing image segmentation is a challenging task. In urban remote sensing, the presence of occlusions and shadows often results in blurred or invisible object boundaries, thereby increasing the difficulty of segmentation. In this paper, an improved network with a cross-region self-attention mechanism for multi-scale features based on DeepLabv3+ is designed to address the difficulties of small object segmentation and blurred target edge segmentation. First, we use CrossFormer as the backbone feature extraction network to achieve the interaction between large- and small-scale features, and establish self-attention associations between features at both large and small scales to capture global contextual feature information. Next, an improved atrous spatial pyramid pooling module is introduced to establish multi-scale feature maps with large- and small-scale feature associations, and attention vectors are added in the channel direction to enable adaptive adjustment of multi-scale channel features. The proposed network model is validated using the Potsdam and Vaihingen datasets. The experimental results show that, compared with existing techniques, the network model designed in this paper can extract and fuse multi-scale information, more clearly extract edge information and small-scale information, and segment boundaries more smoothly. Experimental results on public datasets demonstrate the superiority of our method compared with several state-of-the-art networks.

KEYWORDS

Semantic segmentation; remote sensing; multiscale; self-attention

1 Introduction

With the rapid development of aerospace technology and integrated earth observation systems, we are entering the era of big remote sensing data. Remote sensing technology is widely used in ecological environment monitoring [1], precision agriculture [2,3], land and resource surveys [4], and urban planning [5,6]. The explosion of remote sensing data makes the efficient processing and intelligent interpretation of massive remote sensing data a critical problem [7]. Semantic segmentation of remote sensing images, which is an important topic of remote sensing image processing, is a prerequisite for subsequent scene understanding, feature monitoring, and 3D reconfiguration, and has played a significant role in promoting the development of remote sensing technology.

In recent years, the powerful capabilities of convolutional neural networks have been extensively investigated. Because of their robust feature extraction capabilities, feature extraction modes from superficial to deep have enabled the conversion from shallow detailed features (e.g., color, location, texture) to higher abstract categories, realizing the effective capture of semantic information [8,9]. Therefore, the use of deep learning-based methods to segment feature information quickly and accurately has become a hot topic of research. In 2015, Long et al. proposed the Fully Convolutional Network (FCN) [10], which achieved end-to-end semantic segmentation for the first time and applied convolutional neural networks to the field of image semantic segmentation. Subsequently, SegNet [11] and DeconvNet [12] were proposed to reduce the loss of detail in FCNs. Ronneberger et al. further optimized FCNs in 2015 by proposing a U-Net [13] network model based on the encoder and decoder structures. U-Net was first applied to medical image semantic segmentation, which has an encoder similar to the FCN and extracts image features through operations such as concession and pooling. Yu et al. proposed the Dilated Convolution [14], which retains more contextual information by expanding the perceptual field while maintaining a constant resolution. The nature of the local receptive field of the full convolutional network restricts the pixel-level recognition to the local region. This makes it difficult to connect the contextual cues, and affects the further improvement of the segmentation accuracy.

In addition to the above, Chen et al. proposed the DeepLab v3 [15] network model based on the encoder and decoder structure, which improved on the previous DeepLab v1 [16] and DeepLab v2 [17] network models by obtaining multiscale context information with the help of several parallel dilated convolutions with different dilation rates. The segmentation accuracy of the network model is improved with the help of several parallel convolutions with different atrous rates to obtain multiscale context information. Following that, they further introduced DeepLabv3+ [18]. Zhao et al. proposed the Pyramid Scene Parsing Network (PSPNet) [19] with several global pooling operations of different step lengths to complete the aggregation of multiscale contextual information. Zhang et al. [20] achieved the fusion of high- and low-level features by constructing a feature pyramid network that propagates features from bottom to top. Chu et al. [21] enhanced the recognition accuracy of small targets by utilizing the fusion of features from multiple convolutional layers. Extracting multiscale information allows the model to better understand and process different scales and structures within images, thereby improving the performance of image analysis and processing tasks. This is particularly important in the fields of semantic segmentation, instance segmentation, and object detection. The methods of Zhang et al. [20] and Chu et al. [21] expanded the receptive field, and thus improved the coarse results of feature extraction. However, it is difficult to gain fine-grained global information from remote sensing images with complex backgrounds.

In recent years, convolutional neural network-based approaches have dominated various tasks in computer vision. With the emergence of the Transformer [22] model, the self-attention mechanism has been widely used and pushed to a higher level. Initially, Transformer was introduced to the field of natural language processing. Inspired by the successful application of Transformer to natural language processing, many researchers tried to apply Transformer to computer vision tasks. In 2020, Dosovitskiy et al. proposed the Vision Transformer (ViT) [23] model, which achieved excellent results on image classification tasks. The ViT model usually requires a huge amount of data for pre-training and migrates to small tasks for classification recognition, which is more scalable than traditional convolutional neural networks. In 2021, Han et al. proposed the Transformer in Transformer (TNT) [24] model, which improves on the ViT model by better extracting the global features and local features of the image. Xie et al. [25] proposed SegFormer, which achieves efficient and high-performance segmentation. Touvron et al. proposed the DeiT (Data-efficient image Transformers) [26] model, which extends ViT in terms of knowledge distillation. Wang et al. improved Transformer in 2021 and proposed the CrossFormer [27] model based on the cross-region self-attention mechanism, which can better capture global contextual feature information. Zhao et al. [28] and Li et al. [29] presented region-level attention and frame-level attention for video captioning. Zhao et al. [30] explored the effectiveness of pairwise self-attention and patchwise self-attention in image recognition. SENet [31] expresses the relationship between channels through the global average pooling layer to understand the importance of different channels automatically. Cao et al. [32] employed a hierarchical SwinTransformer architecture and introduced Swin-Unet for medical image segmentation. These new models have made great progress in natural language processing and medical image segmentation, but their segmentation potential when applied to remote sensing images requires further analysis. Different from the above methods, we consider the self-attention associations between features at both large and small scales to capture global contextual feature information.

High-resolution remote sensing images (RSIs) contain rich feature information and complex backgrounds. The complete semantic information understanding is the basis of high-resolution RSI segmentation. Existing segmentation methods for high-resolution RSIs suffer from problems such as blurred segmentation edges and difficulty in segmenting small-scale features [4]. In particular, the presence of occlusions and shadows often leads to blurred or invisible object boundaries, which increases the difficulty of segmentation [33,34]. By learning information at different scales of the image, the complementary information capability of the network can be enhanced, and the loss of feature details and edge information can be avoided to a certain extent. Therefore, this paper describes a network model with multiscale features and the cross-region self-attention mechanism, and combines this with the advantages of the Transformer, which can solve the problems of small-object segmentation and blurred target edge segmentation in the semantic segmentation of RSIs. First, the cross-regional self-attention mechanism is incorporated into the encoder of the network, and self-attention is established to determine the correlation between features at different scales, thus capturing global contextual feature information. A modified atrous spatial pyramid pooling (ASPP) module is then introduced after the cross-regional self-attention mechanism to increase the image perceptual field and capture multiscale image feature information. The specific network architecture and design ideas are described in detail in the following sections.

The main contributions of this paper are as follows:

(1) We design a network model with multiscale features and the cross-region self-attention mechanism to address the challenges of segmenting small objects. This model improves the accuracy of object edge segmentation in high-resolution urban remote sensing imagery.

(2) We incorporate cross-scale embedding layers and cross-region self-attention mechanisms into DeepLabv3+, linking features of different scales with contextual information to enhance spatial attention.

(3) We improve the ASPP module to enhance the model's ability to extract edge information and strengthen multiscale representation. The proposed network employs five parallel convolution kernels with different receptive field sizes to extract features. We optimize the network parameters in ASPP using spatially separable convolutions in place of traditional convolutions.

2 Related Work

In this section, we review existing theories and methods related to our proposed model. This includes the self-attention mechanism and the CrossFormer.

2.1 Semantic Segmentation of Remote Sensing Images

Semantic segmentation of RSIs continues to receive widespread attention. Most current segmentation methods focus on the effective fusion of low-level spatial details and high-level semantic cues, or incorporate boundary supervision to obtain boundary guidance. Remote sensing is enriched with spatial information, but this may be masked when high-level features and lower feature layers are fused. To address this problem, Wen et al. [35] proposed the MLWNet semantic segmentation model, which uses multiscale linear self-attention modules to abstract the correlation between contexts. To enhance the contextual information and alleviate the semantic ambiguity resulting from occlusion, Xiao et al. [36] used a parallel context aggregation module in the encoder and a feature shrinkage module in the downsampling process to minimize the information loss. Wang et al. [37] investigated building extraction using high-resolution RSIs from three aspects: The extraction of multiscale information, the multilevel merging of high-level semantic information, and multilevel information fusion. To avoid any loss of spatial detail and improve the segmentation accuracy, Fan et al. [38] proposed a progressive adjacent-layer coordination symmetric cascade network for the cross-layer fusion of multimodal remote sensing data and the preservation of spatial information. For effective aggregation of remote contextual cues and combining multilevel features, Wang et al. [39] developed a boundary guided multilevel feature fusion module. This module embeds the boundary guided information into the multilevel feature fusion process and subtly facilitates spatial contextual and channel-level semantic correlation at the pixel level. Wang et al. [40] constructed an asymmetrical convolution and orientation attention module for the adaptive selection of favorable segmentation features and enhancement of the inherent geometric features of remote sensing target contours. Using the multiscale features of the image effectively improves the coarse results of feature extraction and fuses the global contextual semantic features, but tends to lose some boundary information. Therefore, the extraction of cross-scale information of RSIs, the recovery of object boundary information, and the localization of small-scale features require further study. In particular, effective auxiliary modules for deep learning models need to be developed.

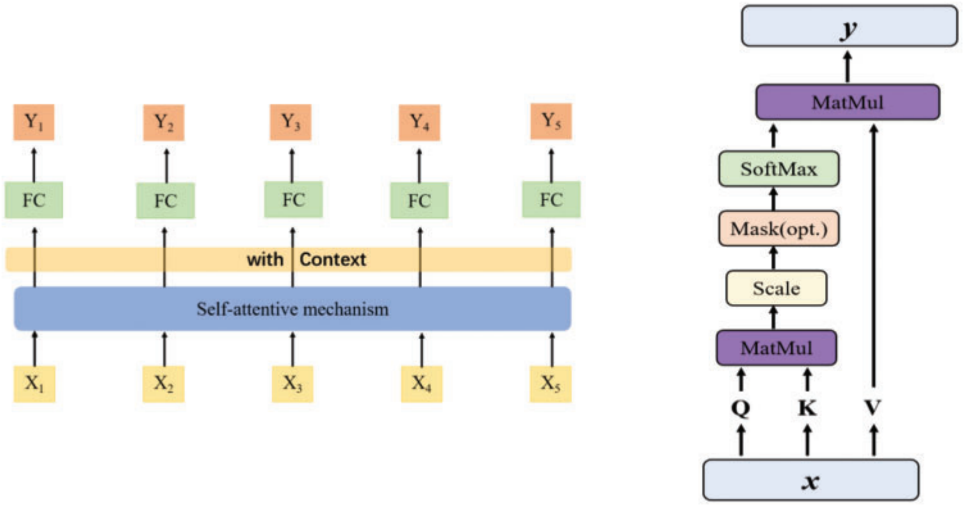
2.2 Self-Attention Mechanism

Self-attention can be considered as a network layer that can better correlate contextual information in the input vectors than fully connected neural networks. In general, the network training needs to consider the entire set of input vectors. Fully connected neural networks require the connection of all input vectors, which generates a large number of parameters and leads to overfitting, as well as increasing the computational effort. Unlike fully connected neural networks, self-attention has a larger field of perception and achieves the association between each pixel in an image through a vector group.

There is a one-to-one correspondence between the self-attention input vector and the output vector, as shown in Fig. 1a. The input sequence is X and the output sequence is Y. FC is the fully connected layer. The self-attention mechanism combines the input vector X with contextual information and then outputs Y with global relevance through FC. The specific process can be described by the following equation:

$$\mathrm{Y} = \varphi(\omega(\mathrm{X})) \tag{1}$$

where φ is the fully connected layer FC and ω is the self-attention mechanism layer.



(a) (b)

Figure 1: (a) Schematic diagram of the self-attention mechanism. (b) Operation flow of scaled dot-product for self-attention mechanism

The calculation process of the self-attention mechanism is shown in Fig. 1b. In the proposed method, a scaled dot-product self-attention mechanism is used. The dot-products of the input sequence X with the matrices U_q, U_k, U_v give the matrices Q, K, and V, respectively. The specific calculation process is expressed by the following equations:

$$Q = U_q \cdot X \tag{2}$$

$$K = U_k \cdot X \tag{3}$$

$$V = U_v \cdot X \tag{4}$$

$$A = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{5}$$

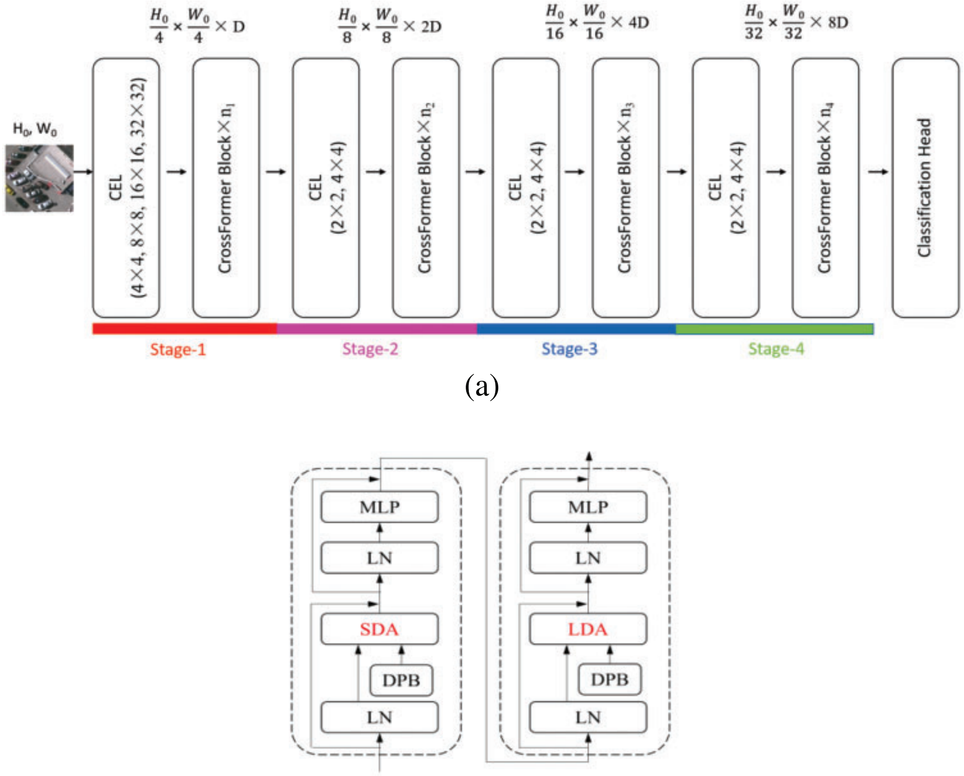
$$Y = A \cdot V \tag{6}$$

where U_q, U_k, and U_v are weight matrices that are updated by network training. Q is called the Query and K-V is the Key-Value pair. A is called the attention score, which indicates the similarity between Q and K. SoftMax is the fully connected layer. Y indicates the weighted aggregation by A and V. d_k denotes the dimension of K.

2.3 Cross-Regional Self-Attention Mechanism

Transformer is a deep neural learning network based on a self-attention mechanism that processes data in parallel. It slices the input image into patches of equal size, spreads the patches into one-dimensional vectors by embedding, and labels each patch with a positional encoding. This ensures that all embeddings in the same layer of the network model have the same scale and removes cross-scale characterization. Based on the powerful global modeling capability of Transformer, Wang et al. [27] developed CrossFormer based on the cross-region self-attention mechanism, establishing the Cross-scale Embedding Layer (CEL) and the Long Short Distance Attention (LSDA) module. The LSDA module can better capture global contextual feature information.

Fig. 2a shows the overall CrossFormer structure and Fig. 2b shows two consecutive CrossFormer modules. Each CrossFormer module consists of a Short-Distance Attention (SDA) module or a Long-Distance Attention (LDA) module and a multi-layer perceptron (MLP). Moreover, the SDA and LDA modules alternate in consecutive CrossFormer modules, while the Dynamic Position Bias (DPB) module acts on the SDA and LDA to obtain the embedded position representation. Layer Normalization (LN) is applied before each module and residual connectivity is applied after each module.



(a)

(b)

Figure 2: (a) Structure of crossformer for image segmentation. (b) Two consecutive crossformer blocks. SDA and LDA represent long distance attention mechanisms and short distance attention mechanisms, respectively. DPB represents dynamic position bias

2.3.1 Cross-Scale Embedding Layer

The cross-scale embedding layer fuses multiple image blocks of different scales to provide cross-scale embedding for each stage of the input to the network model. As shown in Fig. 3, four convolutional kernels of different sizes and equal steps are set to sample the feature maps, resulting in four feature maps of different sizes. These four feature maps are projected and connected to generate the embedding layer on the right. Each embedding contains feature information of four different scales. For cross-scale embedding, setting the projection dimension of each scale is a key issue. The computational effort of a convolutional layer is proportional to K^2D^2, in which K is the size of the convolutional kernel and D is the dimension of the input or output (assuming that the input dimension is equal to the output dimension). A large convolutional kernel implies more computational effort than a small one for a given dimensionality. Therefore, the embedding layer employs different dimensional stitching for convolutional kernels of different sizes to balance the computational cost with the model performance.

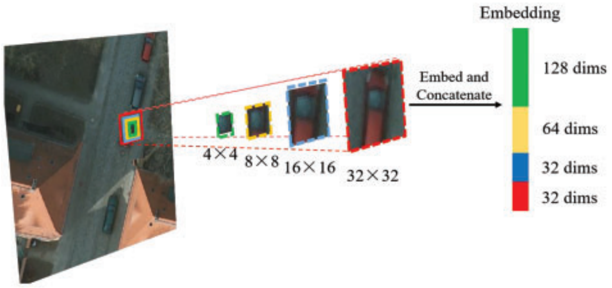


Figure 3: Cross-scale embedding layer of CrossFormer

The cross-scale embedding layer can be realized by four convolutional layers, as described in Table 1. Four convolutional kernels of different sizes (4 × 4, 8 × 8, 16 × 16, and 32 × 32) are adopted for the embedding layer. Assuming that the embedding has a total of 256 dimensions, it can be found from the table that larger convolutional kernels use fewer dimensions, while smaller ones use more dimensions.

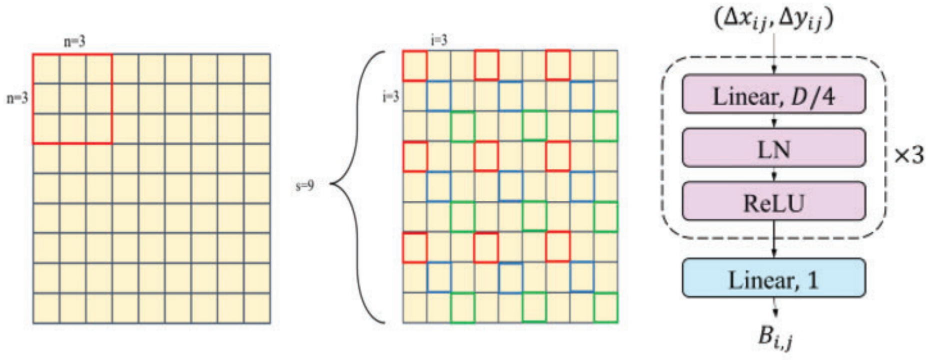
Table 1: Parameters of the cross-scale embedding layer

Type	Kernel Size	Step length	Dimension
Conv1	4 × 4	4 × 4	128
Conv2	8 × 8	4 × 4	64
Conv3	16 × 16	4 × 4	32
Conv4	32 × 32	4 × 4	32

2.3.2 LSDA Mechanism

Self-attention involves computing the relationship between each patch token and all other tokens. This intensive computation leads to an increase in the overhead of the model. To reduce the computational load, the self-attention module is divided into SDA and LDA units.

SDA divides the adjacent image blocks into one group, as shown in Fig. 4a. All image blocks in the red border are divided into one group, which uses a window of size (in Fig. 4a, n = 3). Each group has nine image blocks, so there are a total of nine groups. SDA restricts the self-attention calculation to within each group, thus obtaining the relevance of the local information. LDA performs interval sampling, as shown in Fig. 4b, using a fixed interval of i = 3 for sampling. Therefore, the image blocks represented by red, blue, and green borders are divided into separate groups. LDA acquires cross-regional information because its self-attention computation is cross-scale, unlike SDA. LDA reduces the computational volume and retains feature information from both small and large scales. The calculation process of DPB is shown in Fig. 4c.



(a) (b) (c)

Figure 4: (a) Short Distance Attention (SDA). Embeddings (yellow squares) are grouped by red boxes. (b) Long Distance Attention (LDA). Embeddings with the same borders belong to the same group. (c) Dynamic Position Bias (DPB), with its output being a scalar

3 Proposed Method

DeepLab was launched in 2014 and has shown strong potential in semantic segmentation. The rise of ViT has taken visual recognition into a new era. Hierarchical Transformer structures (e.g., CrossFormer, SwinTransformer [41]) have enabled the success of ViT as a generalized visual task backbone, and have shown strong performance in major visual tasks. The powerful global modeling ability of Transformer is what CNNs lack. Therefore, we combine CrossFormer with a CNN. CrossFormer solves the problem of insufficient interaction between different scale features, and achieves the correlation of information at different scales, establishing contextual connections. In revisiting the development of computer vision, ConvNeXt [42] shows us that ConvNets are highly dynamic. We believe that DeepLab still has research value and have explored it in conjunction with Transformer.

3.1 Multiscale Feature Cross-Regional Self-Attention Mechanism Network

The proposed network with multiscale features and the cross-region self-attention mechanism is shown in Fig. 5. The overall structure follows the design of DeepLabv3+. It inherits the encoding–decoding structure of DeepLabv3+ and realizes the fusion of shallow features and multiscale features.

Unlike ViT, CrossFormer compensates for the interaction of different-scale features. The extraction of equal-scale features in each layer and the adjacent embedding of self-attention modules not only uses cross-scale interaction, but also sacrifices small-scale features. The core of CrossFormer, i.e., the Cross-scale Embedding Layer (CEL), and LSDA generate a cross-attention module that realizes the associations of different-scale information in Transformer. This is in accordance with the idea of the DeepLabv3+ backbone using atrous convolution to extract different-scale features. Thus, we consider the CrossFormer Block as the core for feature extraction.

The backbone structure used in the proposed method is shown in Fig. 6. First, the backbone network extracts features from the input map, which is a cross-region self-attention network consisting of four CrossFormer-Stages connected sequentially. Each CrossFormer-Stage consists of a CEL and a CrossFormer Block (see Fig. 2a). The CEL mixes each embedding with multiple patches of different scales (see Fig. 3). In the backbone, the feature map (Feature-a) output from the first stage is used as a low-level feature input to the decoder for skip connection. In the cross-regional self-attention mechanism network, the feature map output from the first stage is input into the decoder as a low-level feature for hopping connections. Feature-a implements feature fusion with the high-level semantics of another branch in the decoder. Feature-b, which is output from the last stage of the backbone network, is input into the depthwise-separable ASPP as high-level semantic features. We use bilinear interpolation upsampling throughout the network.

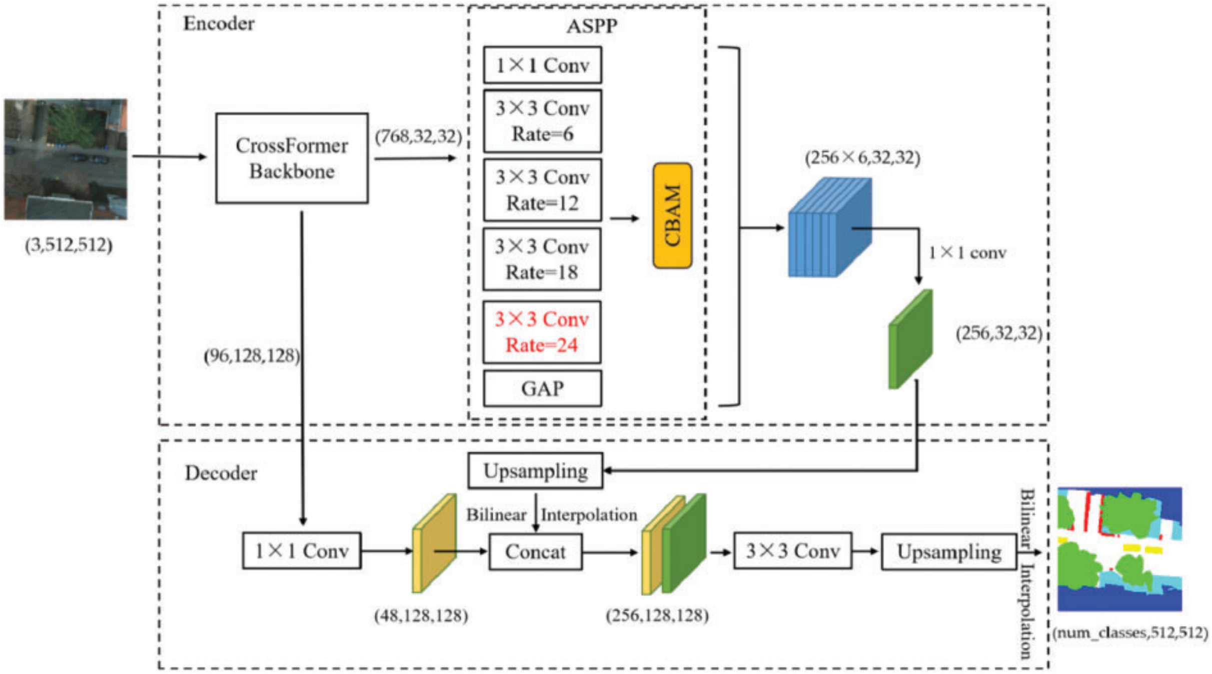


Figure 5: Modified model based on DeepLabv3+ designed in this paper

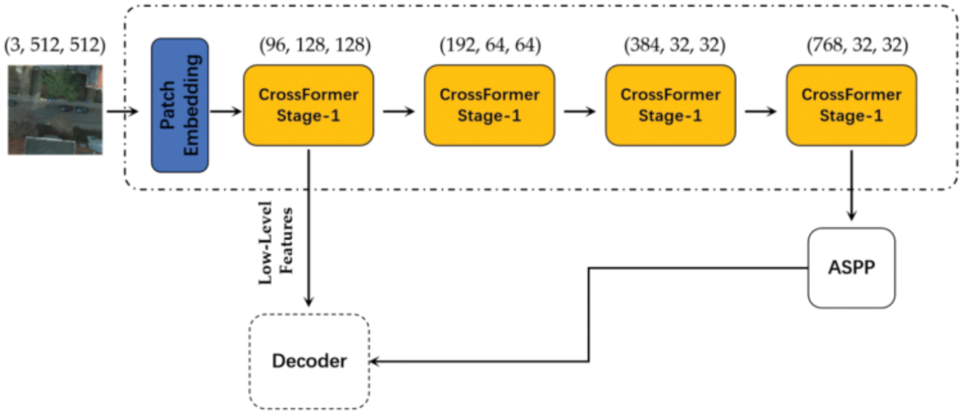
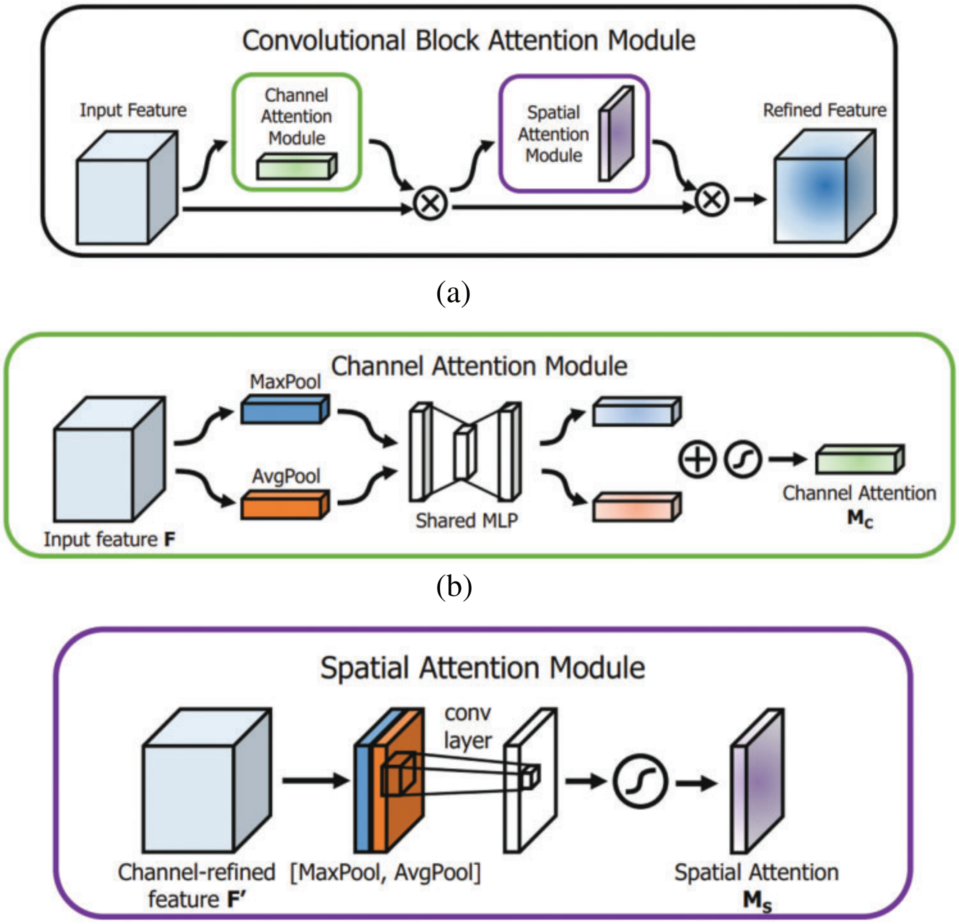


Figure 6: CrossFormer backbone network described in this paper

3.2 Depthwise-Separable ASPP Module with CBAM

SeNet, ECA (Efficient Channel Attention) [43], and SE Block (Squeeze-and-excitation) [44] construct interdependencies between channels, allowing the network to adaptively recalibrate feature responses in the channel direction. ConvNets extract feature information by mixing cross-channel and spatial information, and the Convolutional Block Attention Module (CBAM) [45] emphasizes the features along these two dimensions: Channels and spatial axes. CBAM applies the channel and spatial attention modules sequentially (see Fig. 7), enabling each branch to learn "what" and "where" in the channels and spatial axes, respectively. The attention is inferred along these two dimensions to refine the image features adaptively. Thereby, the network can concentrate on essential features and suppress the unwanted features.



(a)

(b)

(c)

Figure 7: (a) Overall structure of CBAM. (b) Channel attention mechanism module in CBAM. (c) Spatial attention mechanism module in CBAM

We introduce CBAM to further refine and supervise the multiscale feature layer. The spatial attention module can supervise the spatial features of the extracted multiscale feature layers under different expansion rates, which enables the model to focus its attention on specific regions of the image, thus improving the model perception and adapting to inputs of different scales. The channel attention module enables the model to concentrate on channels containing specific features of greater importance, and dynamically adjusts the attention to each channel, thus reducing the influence of redundant channels and improving the expression ability of features. Therefore, we use CBAM after ASPP to supervise the multiscale feature map in two dimensions and further enhance the connection between large and small scales.

We apply Depthwise-Separable Convolution (DSC) to the ASPP module to constitute a depth-separable ASPP. DSC includes both Depthwise Convolution and Pointwise Convolution, as shown in Fig. 8.

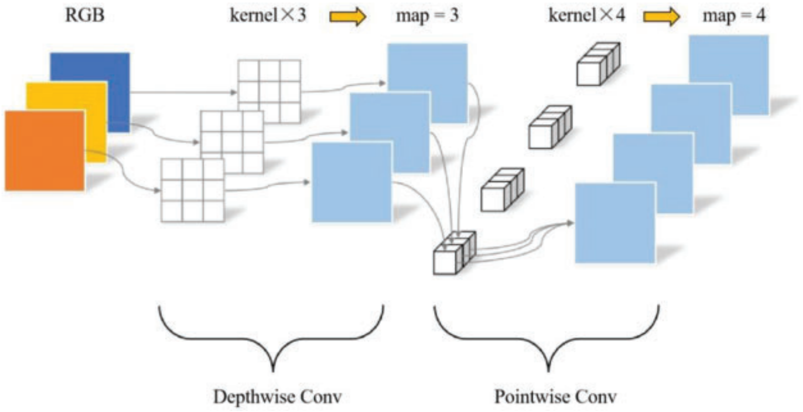


Figure 8: Schematic diagram of Depthwise-Separable Convolution (DSC)

The input image or feature map is subjected to Depthwise Convolution to obtain new feature maps with the same number of channels as the input layer. To add the number of feature maps, a new feature map is generated by splicing the feature map dimensionally by Pointwise Convolution (1 × 1 Conv). DSC integrates both the spatial information and the information from different channels; compared with standard convolution, it has fewer parameters and lower computational cost.

Compared with the original DeepLabv3+ backbone network, the CrossFormer backbone has cross-scale representation, but does not possess the strengths of atrous convolution for edge information extraction. Thus, we use five different 3 × 3 DSCs (with rates of 1, 6, 12, 18, and 24) in the depthwise-separable ASPP module to enhance the multiscale representation. ASPP can encode multiscale contextual information with filtering or pooling. The loss of boundary information during the downsampling of the backbone network can be alleviated by extracting denser feature mappings with atrous convolution. The ASPP module designed for the experiments in this paper is shown in Fig. 9.

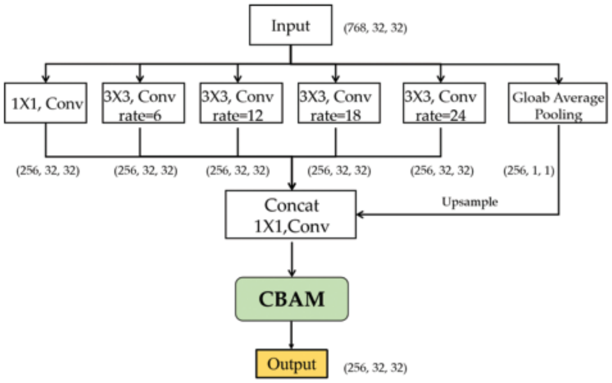


Figure 9: Diagram of depthwise-separable ASPP module

The main improvements of this ASPP module are described below. First, we add an atrous convolution with a rate of 24 to the original ASPP module. The purpose of this improvement is to obtain a larger receptive field and provide a more effective extraction of the fine-edge information that is easily lost in the downsampling process. Second, to improve the efficiency of model training, depthwise-separable atrous convolution is introduced by replacing the convolution in the original ASPP module. The new ASPP module has a 1 × 1 DSC in the first branch, which is designed to preserve the original receptive field. The second to fifth branches apply depthwise-separable atrous convolution with different rates to obtain different receptive fields for multiscale feature extraction. Third, global average pooling is used to obtain global feature information. Finally, the feature maps are stacked in the channel dimension and passed through a standard (1 × 1) convolution to fuse the information of different scales. The final deep information feature map of the encoder is output through the CBAM.

4 Experiments and Results Analysis

4.1 Datasets

The experiments are conducted using two open datasets, Potsdam and Vaihingen, provided by the International Society for Photogrammetry and Remote Sensing (ISPRS) [46]. Potsdam is a high-resolution RSI dataset containing 38 aerial high-resolution RSIs taken by an unmanned aerial vehicle over the German city of Potsdam. Each image is 6000 × 6000 pixels in size. Vaihingen is an aerial image set of a German village, containing 33 aerial images of different sizes. The numbers of images used for training, validation, and testing are listed in Table 2. Sample images of the Potsdam and Vaihingen datasets are shown in Fig. 10.

Table 2: Division of the Potsdam and Vaihingen datasets

Datasets	Train	Val	Test
Potsdam	3853	1284	1284
Vaihingen	3265	1030	859

Both the Potsdam and Vaihingen datasets were classified into six common landcover categories, namely Impervious surfaces, Buildings, Low vegetation, Trees, Cars, and Background, as described in Table 3.

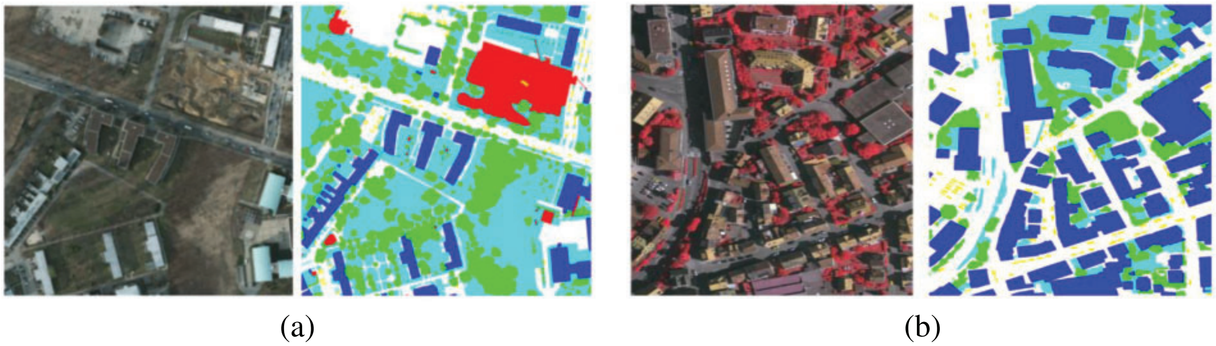


Figure 10: (a) Potsdam dataset image with corresponding labels. (b) Vaihingen dataset image with corresponding labels

Table 3: Potsdam and Vaihingen datasets category label color comparison table

Category	Category name	Color (RGB)
0	Impervious surfaces	White (255, 255, 255)
1	Buildings	Blue (0, 0, 255)
2	Low vegetation	Cyan (0, 255, 255)
3	Trees	Green (0, 255, 0)
4	Cars	Yellow (255, 255, 0)
5	Background	Red (255, 0, 0)

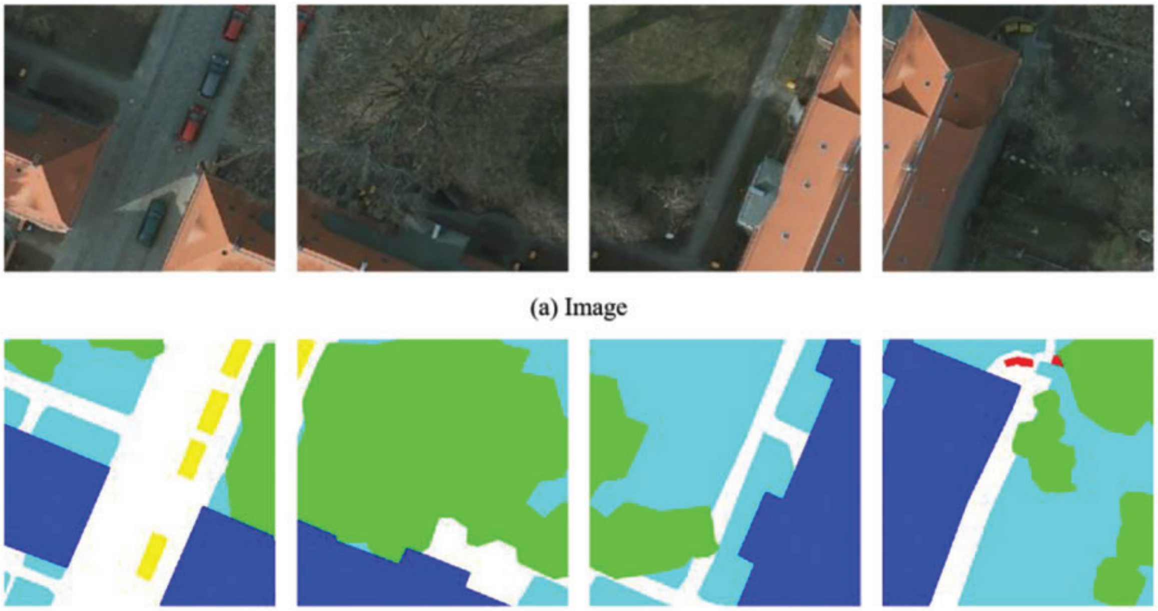
4.2 Data Preprocessing

In this experiment, the original large images of the Potsdam and Vaihingen datasets and the corresponding segmentation labels were cropped into sub-images of size 512 × 512 pixels. The cropping selection was performed in sliding window mode, with the image edge repetition rate set to 15%, as shown in Fig. 11.

To further enhance the model's generalization ability, we expanded the number of samples using data augmentation. This increases the number of samples by adding Gaussian noise or pepper noise, or by performing random rotation, vertical flip, and horizontal flip operations (Fig. 12). Data augmentation improves the generalization ability of the model.

4.3 Evaluation Indicators

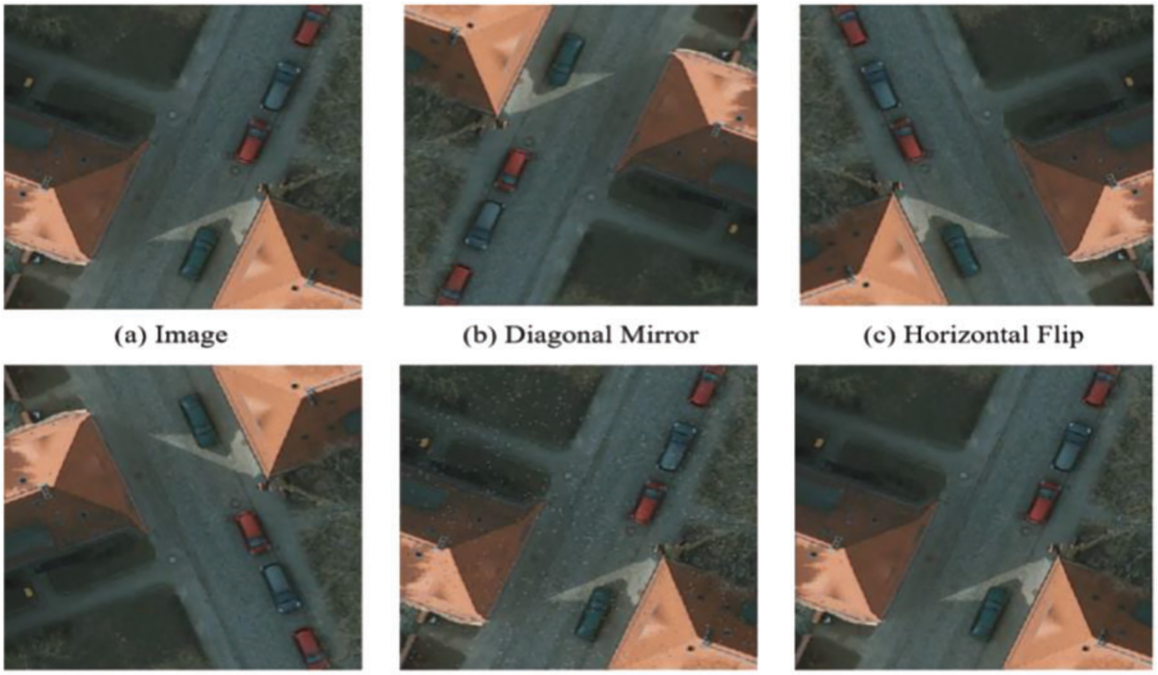
To objectively evaluate the segmentation effect of each model, this paper uses the Confusion Matrix (Table 4) to calculate the Intersection over Union (IoU), Mean Intersection over Union (MIoU), and F1-score of the segmentation results given by each model.



(a) Image

(b) Label

Figure 11: (a) and (b) represent the cropped RSI and their labels, respectively



(a) Image

(b) Diagonal Mirror

(c) Horizontal Flip

(d) Vertical Flip

(e) Pepper noise

(f) Gaussian noise

Figure 12: Data enhancement effect

Table 4: Confusion matrix

Confusion matrix		Predicted value	
		Positive	Negative
True Value	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

The evaluation indicators are defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$\text{F1\_score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

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

$$\text{MIoU} = \frac{1}{\text{k} + 1} \sum_{\text{i}=0}^{\text{k}} \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}} \tag{11}$$

where TP, FP, FN, and TN represent True Positive, False Positive, False Negative, and True Negative, respectively. Precision is the percentage of positive samples that are correctly predicted in the prediction results. Recall is the percentage of predicted true samples out of the total number of true samples. Accuracy is the ratio of correct predictions to the total number of predicted samples, and the F1-score is an overall index combining the Precision and Recall. IoU is the ratio of the intersection of predicted and true values of a category to the concurrent set, and MIoU is the average of the ratio of the intersection and concurrent set of all categories.

In addition, to evaluate the time and space complexity of the network, the number of parameters in each model and the number of floating point operations (FLOPs) are compared.

4.4 Experimental Environment and Parameters

The software and hardware configurations used for the experiments are listed in Table 5.

Table 5: Experimental hardware and software configuration table

Configuration	Versions
GPU	NVIDIA GeForce RTX 3090 Ti (24 GB)
CPU	Intel Core i7-11700 @ 2.50 GHz
Memory	32 G
Operating system	Windows10

(Continued)

Table 5 (continued)

Configuration	Versions
Programming	Python3.7
Framework	PyTorch1.10.1

The optimizer uses the stochastic gradient descent (SGD) algorithm with a "Poly" learning rate strategy, and sets the initial learning rate to 0.01, momentum to 0.9, and weight decay to 0.0005. The training Batch Size is set to 8, with a total of 80,000 iterations. The loss value of the network model and the accuracy of each category are output every 40 iterations. We used cross-entropy loss functions in our experiments.

4.5 Analysis of Experimental Results

To evaluate the performance of the network models designed in this paper, we cropped the large images from the Potsdam and Vaihingen datasets into training images of 512 × 512 pixels for the experiments, and quantitatively evaluated and compared U-Net, DeepLabv3+, SwinUnet, SegFormer, TCUnet [47] and DBENet [48] against the models proposed in this paper.

From Table 6, it can be found that the MIoU of the proposed model reached 76.41% on the Potsdam dataset, and the mean F1-score (mFscore) reached 86.73%. Compared with U-Net, DeepLabv3+, SwinUnet, SegFormer, TCUnet and DBENet, this represents an improvement of 8.74%, 6.95%, 9.40%, 0.98%, 5.65% and 3.63% in MIoU and 6.96%, 5.30%, 8.35%, 1.36%, 6.04% and 5.41% in mFscore. Compared with the IoU of DeepLabv3+, the proposed network model gives an improvement of 7.29%, 5.45%, 9.45%, 11.18%, 6.74%, and 1.59% in the six landcover categories of impervious surfaces, buildings, low vegetation, trees, cars, and background, respectively. The specific quantitative evaluation indicators are listed in Tables 6 and 7.

Table 6: Comparison of overall evaluation metrics of different network models on the Potsdam dataset (%)

Evaluation indicators	U-Net [13]	DeepLabv3+ [18]	SwinUnet [32]	SegFormer [25]	TCUnet [47]	DBENet [48]	Ours
MIoU	67.67	69.46	67.01	75.43	70.76	72.78	76.41
mFscore	79.77	81.43	78.38	85.37	80.69	81.32	86.73
Parameters	29.06	62.57	27.14	3.72	1.72	18.15	107.16
FLOPs	202.56	253.93	31.02	7.90	3.24	99.88	278.49

Table 7: IoU (%) for different network models on the Potsdam datasets

Evaluation indicators	U-Net [13]	DeepLabv3+ [18]	SwinUnet [32]	SegFormer [25]	TCUnet [47]	DBENet [48]	Ours
Impervious surfaces	75.65	76.35	76.31	83.56	80.24	80.78	83.64
Buildings	83.12	86.08	84.87	90.76	86.07	89.96	91.53
Low vegetation	65.44	65.91	62.47	74.56	63.39	70.23	75.36
Trees	67.57	65.75	65.54	74.30	70.42	71.33	76.93
Cars	75.18	73.29	68.36	78.85	77.58	75.56	80.03
Background	39.04	49.38	44.27	50.55	46.91	48.87	50.97

From Table 8, it can be found that the MIoU of the proposed network model reached 88.95% on the Vaihingen dataset, and the mFscore reached 94.08%. Compared with U-Net, DeepLabv3+, SwinUnet, SegFormer, TCUnet and DBENet, this represents an improvement of 7.86%, 3.79%, 0.72%, 0.83%, 2.36% and 1.93% in MIoU and 4.72%, 2.29%, 0.52%, 0.50%, 1.87% and 1.92% in mFscore. Compared with the IoU of DeepLabv3+ the proposed network model produces an improvement of 1.98%, 1.66%, 1.44%, 0.93%, 7.45%, and 6.26% in the six landcover categories of impervious surfaces, buildings, low vegetation, trees, cars, and background, respectively. The specific quantitative evaluation indicators are presented in Tables 8 and 9.

Table 8: Comparison of overall evaluation metrics for different network models on the Vaihingen datasets (%)

Evaluation indicators	U-Net [13]	DeepLabv3+ [18]	SwinUnet [32]	SegFormer [25]	TCUnet [47]	DBENet [48]	Ours
MIoU	81.09	85.16	88.23	88.12	86.59	87.02	88.95
mFscore	89.36	91.79	93.56	93.58	92.21	92.16	94.08
Parameters	29.06	62.57	27.14	3.72	1.72	18.15	107.16
FLOPs	202.56	253.93	31.02	7.90	3.24	99.88	278.49

Table 9: IoU (%) for different network models on the Vaihingen datasets

Evaluation indicators	U-Net [13]	DeepLabv3+ [18]	SwinUnet [32]	SegFormer [25]	TCUnet [47]	DBENet [48]	Ours
Impervious surfaces	86.31	90.64	92.39	92.02	91.76	91.88	92.62
Buildings	92.45	94.40	95.62	95.58	94.32	94.56	96.06
Low vegetation	74.31	84.97	88.24	85.53	83.53	84.79	86.41

(Continued)

Table 9 (continued)

Evaluation indicators	U-Net [13]	DeepLabv3+ [18]	SwinUnet [32]	SegFormer [25]	TCUnet [47]	DBENet [48]	Ours
Trees	80.67	86.99	89.57	87.43	85.63	85.23	87.92
Cars	69.55	70.09	72.23	77.13	75.32	76.42	77.54
Background	83.25	83.87	91.28	91.06	89.01	89.24	90.13

From Tables 6 and 8, it can be found that the experimental results obtained for the Vaihingen dataset are similar to those obtained for the Potsdam dataset. This proves that the proposed network model is effective in introducing a cross-regional self-attention mechanism and an improved depthwise-separable ASPP to extract cross-scale features and enhance the performance of the network model.

The segmentation results of the various models on the Potsdam and Vaihingen datasets are shown in Figs. 13 and 14. Looking at Fig. 13, we see that the models (such as Unet, DeepLabv3+, and SwinUnet) have relatively vague segmentation boundaries for buildings (labeled with blue in Fig. 13). For the segmentation of low vegetation (cyan) and trees (green), the models (such as Unet, DeepLabv3+, TCUnet, and DBENet) can achieve the segmentation of the main parts, but the boundary between the two categories is still unclear. The proposed model exhibits greater sensitivity in identifying low vegetation and trees. The segmentation boundaries are more refined than in the other six methods, and the position localization is more accurate. We see that the proposed model provides better segmentation than the other six models, especially in the red rectangular boxes. From Fig. 14, it can be observed that the other six models produce a certain degree of fuzziness at the edges of buildings and smaller-scale low vegetation. The proposed model effectively alleviates this issue and accurately locates small-scale information. The experimental results show that the model described in this paper is more detailed in the segmentation of high-resolution RSI datasets, captures small-scale information more accurately, and reproduces the details better.

4.6 Ablation Experiment

We denote the CrossFormer backbone network, the ASPP module, and the CBAM as modules 1, 2, and 3, respectively. To verify the performance of each component of the whole network, we conducted ablation experiments on the Potsdam dataset.

"Deeplabv3+" represents the original DeepLabv3+ model. "Deeplabv3+ +1+2" denotes the original backbone network and ASPP in DeepLabv3+ have been replaced by modules 1 and 2. We conducted a longitudinal comparison to analyze the performance of the proposed modules.

Comparison between "Deeplabv3+ +all" and "Deeplabv3+ +1+3" (where the former utilizes an improved ASPP module and the latter uses the original ASPP module): After incorporating modules 1 and 3, the improved ASPP module exhibits an improvement of 1.73% in MIoU and 1.71% in mFscore compared to the original ASPP module.

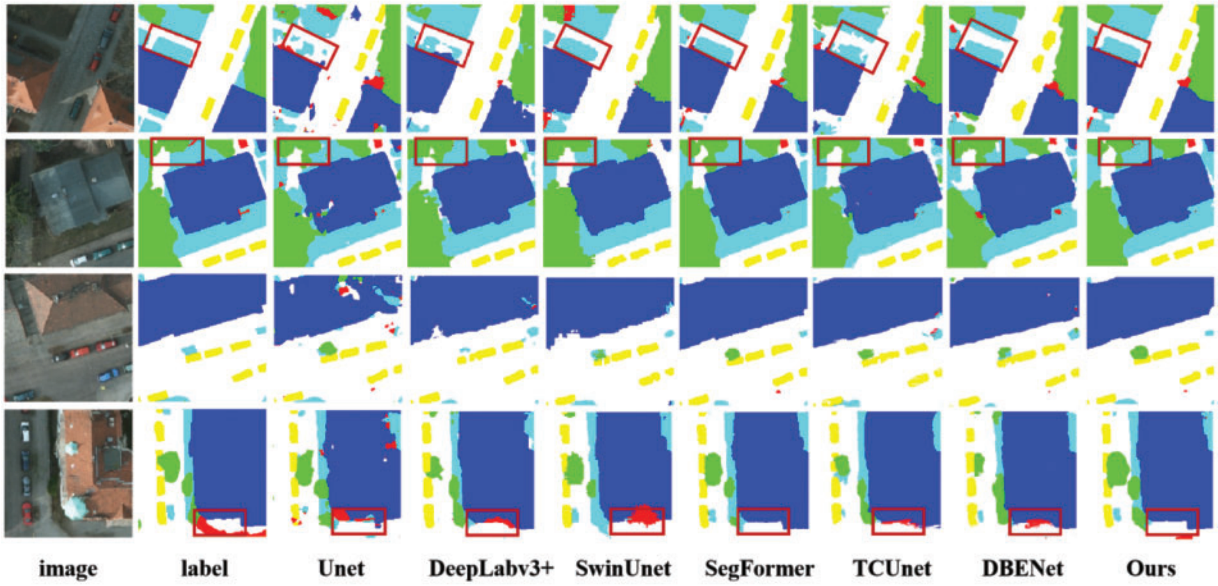


Figure 13: Visualization results of Potsdam datasets

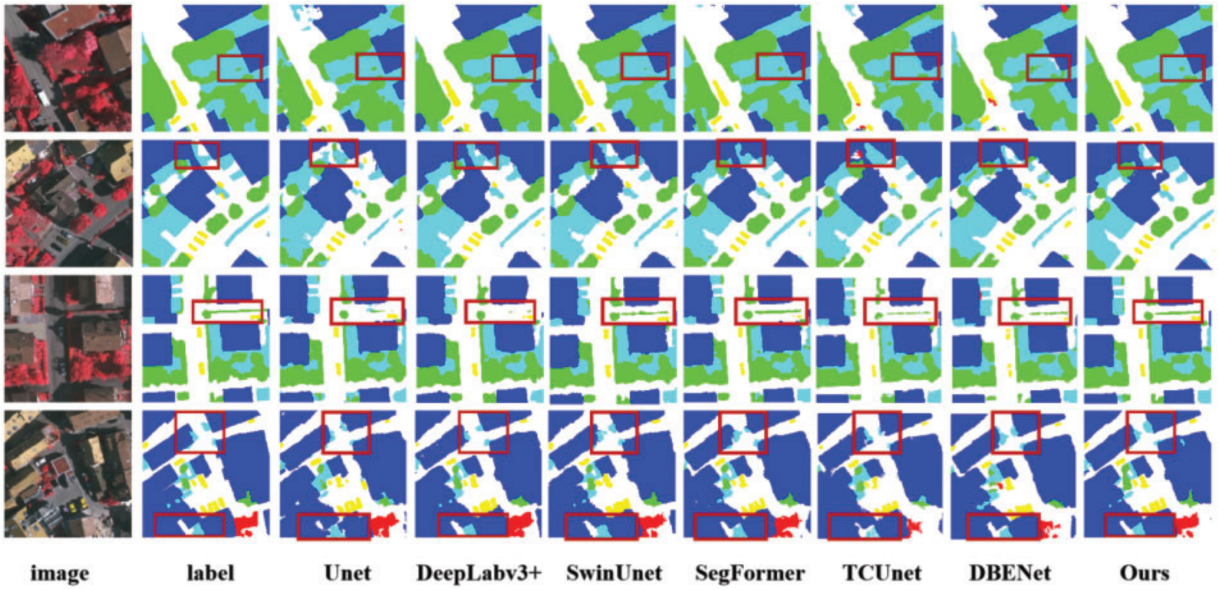


Figure 14: Visualization results of Vaihingen datasets

“Deeplabv3+ +all” compared to “Deeplabv3+ +2+3” (where the former uses the CrossFormer backbone network and the latter employs the original convolutional network): After replacing the backbone network, there was a noticeable improvement. Based on the metrics, the CrossFormer-based structure shows an increase of 3.85% in MIoU and 3.79% in mFscore compared to the CNN-based structure.

“Deeplabv3+ +all” compared to “Deeplabv3+ +1+2” (where the former incorporates CBAM, while the latter does not use CBAM): CBAM focuses on crucial feature information, enhancing the model's perception of multiscale features. In terms of MIoU and mFscore, "Deeplabv3+ +all" gives improvements of 1.6% and 1.77%, respectively, over "Deeplabv3+ +1+2".

The results of the ablation experiments are shown in Fig. 15. From the data in Table 10, the combination of modules 1 + 2 and 1 + 3 significantly improves the accuracy. In Fig. 15, the 1 + 2 combination significantly improves the extraction of cars and low vegetation, and the extracted boundaries are smooth and clear. This indicates that CrossFormer and the improved ASPP enhance the ability to recognize small-scale features. The combination of modules 1 + 3 is highly effective in recognizing buildings with fewer "voids". This indicates that our enhancement of cross-scale feature information exchange is effective.

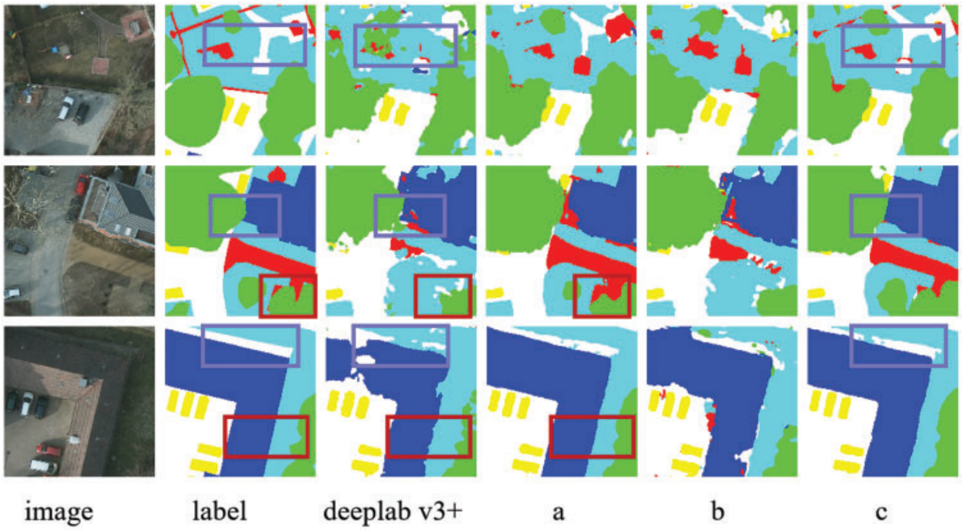


Figure 15: Examples of semantic segmentation results on the Potsdam dataset. a. Deeplabv3+ +1+2. b. Deeplabv3+ +2+3. c. Deeplabv3+ +1+3

Table 10: Ablation experimental design

Model name	Modules			Evaluation index (%)	
	1	2	3	MIoU	mFscore
Deeplabv3+				69.46	81.43
Deeplabv3+ +1+2	√	√		74.81	84.96
Deeplabv3+ +2+3		√	√	72.56	82.94
Deeplabv3+ +1+3	√		√	74.68	85.02
Deeplabv3+ + all	√	√	√	76.41	86.73

5 Conclusions

This paper has described a cross-regional multiscale DeepLabv3+ improvement model for the difficult problem of small object segmentation in RSIs. The cross-scale embedding layer was established by replacing the backbone feature extraction network of DeepLabv3+ with CrossFormer, and a cross-regional self-attention mechanism was introduced to enhance the connection between local and global contextual information. A modified ASPP structure with CBAM was then used to extract multiscale features to enhance the model's ability to recognize small objects. In this study, we replaced the normal convolution in ASPP with a DSC unit to control the number of parameters. Experiments on the Potsdam and Vaihingen datasets demonstrated that our proposed model performs better than U-Net, DeepLabv3+, SwinUnet, and SegFormer. It is more sensitive to small-scale information and works better on small-scale features. In areas with darker images, all models are prone to misclassification. Providing calculation guidance for low-contrast areas will be considered in future work.

Acknowledgement: The authors are highly thankful to the National Natural Science Foundation of China, to the Hainan Provincial Natural Science Foundation of China. Furthermore, we would also like to thank ISPRS for providing the remote sensing images.

Funding Statement: This research was funded by the National Natural Science Foundation of China (Grant Number 62066013) and Hainan Provincial Natural Science Foundation of China (Grant Numbers 622RC674 and 2019RC182).

Author Contributions: The authors confirm the following contributions to this paper: study conception and design: H. Xing, Q. Tong; data collection: K. Cao, Z. Zhu; writing—review & editing, analysis and interpretation of results: Q. Tong, H. Xing, M. Zhang, K. Cao; draft manuscript preparation: Q. Tong, Z. Zhu, K. Cao. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data presented in this study are available upon request from the corresponding author.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] J. Li, Y. Pei, S. Zhao, R. Xiao, X. Sang and C. Zhang, "A review of remote sensing for environmental monitoring in China," *Remote Sens.*, vol. 12, no. 7, pp. 1130, 2020. doi: 10.3390/rs12071130.

[2] X. Wang *et al.*, "A survey of farmland boundary extraction technology based on remote sensing images," *Electron.*, vol. 12, no. 5, pp. 1156, Feb. 2023. doi: 10.3390/electronics12051156.

[3] L. Xu *et al.*, "Extraction of cropland field parcels with high resolution remote sensing using multi-task learning," *Eur J. Remote Sens.*, vol. 56, no. 1, pp. 2181874, 2023.

[4] M. Shu and S. H. Du, "Forty years of progress and challenges of remote sensing for land survey," *J. Geoinform. Sci.*, vol. 24, no. 4, pp. 597–616, 2022.

[5] A. M. Coutts *et al.*, "Thermal infrared remote sensing of urban heat: Hotspots, vegetation, and an assessment of techniques for use in urban planning," *Remote Sens. Environ.*, vol. 186, pp. 637–651, 2016.

[6] B. Yin, Y. Yu, and L. Su, "Research on the method of extracting urban construction land based on remote sensing data of Gaofen-1 satellite," *Front. Earth Sci.*, vol. 9, no. 5, pp. 334–340, 2019.

[7] B. Zhang *et al.*, "Geographic cognitive models and methods for intelligent interpretation of remote sensing big data," *J. Survey. Map.*, vol. 51, no. 7, pp. 1398–1415, 2022.

[8] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006. doi: 10.1126/science.1127647.

[9] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006. doi: 10.1162/neco.2006.18.7.1527.

[10] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. CVPR*, Boston, MA, USA, 2015, pp. 3431–3440.

[11] 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. doi: 10.1109/TPAMI.2016.2644615.

[12] H. Noh, S. Hong, and B. Han, “Learning deconvolution network for semantic segmentation,” in *Proc. ICCV*, Santiago, CA, USA, 2015, pp. 1520–1528.

[13] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. MICCAI*, Munich, Germany, 2015, pp. 234–241.

[14] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” presented at the 2015 ICLR Conf., Vancouver, BC, Canada, May 7–9, 2015, pp. 1511.

[15] L. C. Chen *et al.*, “Rethinking atrous convolution for semantic image segmentation,” arXiv preprint arXiv:1706.05587, 2017.

[16] L. C. Chen *et al.*, “Semantic image segmentation with deep convolutional nets and fully connected CRFs,” presented at the 2014 CVPR Conf., Portland, OR, USA, Jun. 23–26, 2014, pp. 1412.

[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, 2017. doi: 10.1109/TPAMI.2017.2699184.

[18] L. C. Chen *et al.*, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proc. ECCV, 2018*, Munich, Germany, 2018, pp. 801–818.

[19] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *Proc. IEEE Conf. CVPR*, Hawaii, USA, 2017, pp. 2881–2890.

[20] Y. Zhang, J. Chu, L. Leng, and J. Miao, “Mask-refined R-CNN: A network for refining object details in instance segmentation,” *Sens.*, vol. 20, no. 4, pp. 1010, 2020. doi: 10.3390/s20041010.

[21] J. Chu, Z. Guo, and L. Leng, “Object detection based on multi-layer convolution feature fusion and online hard example mining,” *IEEE Access*, vol. 6, pp. 19959–19967, 2018. doi: 10.1109/ACCESS.2018.2815149.

[22] A. Vaswani *et al.*, “Attention is all you need,” presented at the 2017 NIPS Conf., Long Beach, CA, USA, Dec. 4–9, 2017, pp. 1–55.

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

[24] K. Han *et al.*, “Transformer in transformer,” *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 15908–15919, 2021.

[25] E. Xie *et al.*, “SegFormer: Simple and efficient design for semantic segmentation with transformers,” *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 12077–12090, 2021.

[26] H. Touvron *et al.*, “Training data-efficient image transformers & distillation through attention,” in *Proc. 38th Int. Conf. Mach. Learn.,* 2021, pp. 10347–10357.

[27] W. Wang *et al.*, “CrossFormer: A versatile vision transformer hinging on cross-scale attention,” arXiv preprint arXiv:2108.00154, 2021.

[28] B. Zhao, X. Li, and X. Lu, “CAM-RNN: Co-attention model based RNN for video captioning,” *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5552–5565, Nov. 2019. doi: 10.1109/TIP.2019.2916757.

[29] X. Li, B. Zhao, and X. Lu, “MAM-RNN: Multi-level attention model based RNN for video captioning,” in *Proc. IJCAI*, Melbourne, VIC, Australia, 2017, pp. 2208–2214.

[30] H. Zhao, J. Jia, and V. Koltun, “Exploring self-attention for image recognition,” in *Proc. CVPR*, Seattle, WA, USA, 2020, pp. 10073–10082.

[31] D. Cheng, G. Meng, G. Cheng, and C. Pan, “SeNet: Structured edge network for sea-land segmentation,” *IEEE Geosci. Remote Sensing Lett.*, vol. 14, no. 2, pp. 247–251, 2016. doi: 10.1109/LGRS.2016.2637439.

[32] H. Cao *et al.*, “Swin-Unet: Unet-like pure transformer for medical image segmentation,” in *Proc. ECCV*, Tel Aviv, Israel, 2022, pp. 205–218.

[33] L. Ding, H. Tang, and L. Bruzzone, "LANet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE Geosci. Remote Sensing Lett.*, vol. 59, no. 1, pp. 426–435, Jul. 2020.

[34] X. He *et al.*, "Swin transformer embedding UNet for remote sensing image semantic segmentation," *IEEE Geosci. Remote Sensing Lett.*, vol. 60, no. 1, pp. 1–15, Jan. 2022.

[35] Z. Wen, H. Huang, and S. Liu, "Multi-scale attention fusion network for semantic segmentation of remote sensing images," *Int. J. Remote Sens.*, vol. 44, no. 24, pp. 7909–7926, Dec. 2023.

[36] D. Xiao *et al.*, "Csswin-unet: A Swin-unet network for semantic segmentation of remote sensing images by aggregating contextual information and extracting spatial information," *Int. J. Remote Sens.*, vol. 44, no. 23, pp. 7598–7625, Dec. 2023.

[37] D. X. Wang *et al.*, "SDSNet: Building extraction in high-resolution remote sensing images using a deep convolutional network with cross-layer feature information interaction filtering," *Remote Sens.*, vol. 16, no. 1, pp. 1–23, Dec. 2023.

[38] X. Fan, W. Zhou, X. Qian, and W. Yan, "Progressive adjacent-layer coordination symmetric cascade network for semantic segmentation of multimodal remote sensing images," *Expert. Syst. Appl.*, vol. 238, no. 12, pp. 121999, 2024. doi: 10.1016/j.eswa.2023.121999.

[39] Y. Wang *et al.*, "Geometric boundary guided feature fusion and spatial-semantic context aggregation for semantic segmentation of remote sensing images," *IEEE Trans. Image Process.*, vol. 32, no. 1, pp. 6373–6385, Oct. 2023.

[40] J. Wang, Z. Feng, Y. Jiang, S. Yang, and H. Meng, "Orientation attention network for semantic segmentation of remote sensing images," *Knowl.-Based Syst.*, vol. 267, no. 12, pp. 110415, 2023. doi: 10.1016/j.knosys.2023.110415.

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

[42] Z. Liu *et al.*, "A ConvNet for the 2020s," in *Proc. CVPR*, Tampa, FL, USA, 2022, pp. 11976–11986.

[43] Q. Wang *et al.*, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. CVPR*, Seattle, WA, USA, 2020, pp. 11534–11542.

[44] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, Seattle, Salt Lake City, UT, USA, 2018, pp. 7132–7141.

[45] S. Woo *et al.*, "CBAM: Convolutional block attention module," in *Proc. ECCV*, Munich, Germany, 2018, pp. 3–19.

[46] "ISPRS 2D semantic labeling dataset," Accessed: Jun. 10, 2021. [Online]. Available: https://www.isprs.org/education/benchmarks/UrbanSemLab/default.aspx

[47] X. Xiong *et al.*, "TCUNet: A lightweight dual-branch parallel network for sea-land segmentation in remote sensing images," *Remote Sens.*, vol. 15, no. 18, pp. 1–26, Sep. 2023.

[48] X. Ji, L. Tang, T. Lu, and C. Cai, "DBENet: Dual-branch ensemble network for sea–land segmentation of remote-sensing images," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–11, 2023. doi: 10.1109/TIM.2023.3302376.