BIBED-Seg: Block-in-Block Edge Detection Network for Guiding Semantic Segmentation Task of High-Resolution Remote Sensing Images

Baikai Sui, Yungang Cao, *Member, IEEE*, Xueqin Bai, Shuang Zhang, and Renzhe Wu

***Abstract*—Edge optimization of semantic segmentation results is a challenging issue in remote sensing image processing. This article proposes a semantic segmentation model guided by a block-in-block edge detection network named BIBED-Seg. This is a two-stage semantic segmentation model, where edges are extracted first and then segmented. We do two key works: The first work is edge detection, and we present BIBED, a block-in-block edge detection network, to extract the accurate boundary features. Here, the edge detection of multiscale feature fusion is first realized by creating the block-in-block residual network structure and devising the multi-level loss function. Second, we add the channel and spatial attention module into the residual structure to improve high-resolution remote sensing images' boundary positioning and detection accuracy by focusing on their channel and spatial dimensions. Finally, we evaluate our method on International Society for Photogrammetry and Remote Sensing (ISPRS) Potsdam and Vaihingen data sets and obtain ODS F-measure of 0.6671 and 0.7432, higher than other excellent edge detection methods. The second work is two-stage segmentation. First, the proposed BIBED is individually pretrained, and subsequently, the pretrained model is introduced into the entire segmentation network to extract boundary features. In the second segmentation stage, the edge detection network is used to constrain semantic segmentation results by loss cycles and feature bootstrapping. Our best model obtains the OA of 90.2%, 87.7%, and 81.5%, the IOU of 76.0%, 69.6%, and 61.3% on the ISPRS and WHDLD datasets, respectively.**

***Index Terms*—Channel attention mechanism, edge detection, high-resolution remote sensing, multiple-residual convolution blocks, semantic segmentation, spatial attention mechanism.**

I. Introduction

EDGE optimization in semantic segmentation has always been the focus of research, especially for high-resolution remote sensing images with complex and diverse targets, which is a great challenge. Most of the existing methods are refined from postprocessing by adding postprocessing steps such as morphological filtering and CRF [1], [2]. For example, after the semantic segmentation of marine aquaculture based on FCN, Pan et al. [3] used CRF to refine the edges, which improves the edge definition of marine aquaculture and further improves the extraction accuracy. Also, many studies put edge features into the network as a branch to optimize the loss function and further enhance the edge information of the segmentation results. Guo et al. [4] designed an edge prediction branch to predict the boundary of the salt body, which guides feature learning through the supervision of boundary loss so that the network can distinguish the features on both sides of the semantic boundary. Although these methods can improve the edge blur of classification results to a certain extent, they cannot retard the misclassification phenomenon within the boundary. They can easily enhance the "false" edge in the classification results. Therefore, this article focuses on the remote sensing edge information extraction of the remote sensing image itself and its guiding value for the training of semantic segmentation (see Fig. 1). We aim to eliminate the influence of boundary ambiguity and intraclass dissimilarity on semantic segmentation results through accurate boundary information. This has strict requirements for edge feature detection methods, and also brings challenges to the existing algorithms.

Fig. 1. Semantic segmentation. Left: Segmentation based on DCNN, object boundary is blurred. Right: We propose to mitigate this effect with a clear object boundary map.

Edge detection is an image processing technique for finding objects' boundaries (points with noticeable brightness changes) within digital images. It can significantly reduce the amount of data, eliminate irrelevant information, and retain the essential structural attributes of the images. With the development of remote sensing technology, the resolution of remote sensing images is improved gradually. The features of ground object information shown by high-resolution remote sensing images are becoming more and more complex. The characteristics of the edge of intraclass and interclass are becoming more and more abundant, bringing significant challenges to remote sensing interpretation. Therefore, high precision edge detection is of great significance to the interpretation (especially

Manuscript received 24 August 2022; revised 2 November 2022 and 26 November 2022; accepted 8 January 2023. Date of publication 17 January 2023; date of current version 6 February 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 41771451 and in part by the Sichuan Youth Science and Technology Innovation Team under Grant 2020JDTD0003. *(Corresponding author: Yungang Cao.)*

The authors are with the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu 611756, China (e-mail: 13012482890@163.com; yungang@swjtu.cn; baixueqin1998@163.com; zhang_shuang1999@163.com; mrwurenzhe@my.swjtu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3237584

for segmentation [5], [6], target detection, and recognition[7], [8]) of high-resolution remote sensing images. From low-level visual cues using hand-crafted features [9], [10], [11], [12] to deep-learning models [13], [14], [15], [16], the accuracy of edge detection has been significantly improved.

At first, in the field of computer vision, differential operators were primarily used mainly for image edge detection, divided into two main categories: The first-order differential operators commonly used are Robert [17], Prewitt [18], and Sobel [19], and the second-order differential operators commonly used are LOG [20], Laplace [21], and so on. These operators have many advantages, such as simple principles, easy implementation, and convenient calculation. However, the defects are also obvious: the poor anti-interference ability and the unideal detection effect. Subsequently, many new traditional algorithms have emerged, such as spherical fitting [22], wavelet transform [23], self-adaptive smoothing filter [24], particle swarm optimization [25], and the Canny algorithm [26]. The Canny algorithm has a higher signal-to-noise ratio and shows the best detection effect. Especially for remote sensing images, it has become the most commonly used and practical edge detection algorithm. Based on the traditional edge detection algorithms, many scholars have carried out a lot of research on the edge detection of remote sensing images [27], [28], [29]. However, the antinoise ability of traditional algorithms is still weak, the detection results are prone to "weak" edges, and there are great defects in the accurate extraction of interclass boundaries in the image.

With the rapid development of deep learning in image processing, CNN is widely used because of its excellent semantic information extraction ability [30], [31]. Edge detection based on deep CNN has become a new trend. Famous image edge detection methods based on CNN include N4-Fields [32], DeepContour [45], DeepEdge [33], and HED [34]. These methods show excellent results in natural color images, and their edge detection accuracy is much higher than traditional detection algorithms. However, many of the abovementioned methods may have some problems in the edge detection of remote sensing images: similar to VGG, ordinary convolutional neural networks cannot tap deeper and complex spatial semantic information; single tandem feature extraction is easy to lose key information in shallow features of images; and lack of training and research for multitarget and multispectral images. It easily leads to unclear extraction results and low detection accuracy. This inaccurate edge information does not significantly promote the semantic segmentation results of remote sensing.

Therefore, we deeply excavate the semantic information and spatial features of high-resolution remote sensing and propose an end-to-end edge detection network based on multiple residual convolutional blocks, named BIBED, to obtain more precise boundary information. Enables these features to constrain and guide the semantic segmentation task effectively. Two experiments are mainly carried out in this article. The first experimental result shows that for the edge detection task of high-resolution remote sensing images, BIBED can realize high-precision detection of remote sensing images. It is significantly better than traditional edge detection methods (such as Canny) and other deep learning edge detection models. The second experimental result shows that for the semantic segmentation task of high-resolution remote sensing images, the effective edge features obtained by BIBED can significantly improve the segmentation accuracy of high-resolution remote sensing.

The key contributions of the article are as follows.

1) *BIBED Network:* We propose an end-to-end edge detection network that effectively detects ground object boundaries in high-resolution remote sensing images. First, we establish a block-in-block network structure to fuse edge features from low-level to high-level. Second, in the block structure, according to the characteristics of large pixel series of high-resolution remote sensing images, the residual structure is introduced to solve the problem of gradient disappearance and improve model stability and accuracy while continuously deepening the network. In addition, according to the characteristics of multiband and complex spatial information of high-resolution remote sensing images, the channel attention mechanism and spatial attention mechanism are constructed in the residual structure, focusing on the band and spatial dimension, respectively. To improve the accuracy of edge detection and positioning.
2) *BIBED Loss:* We construct a multiscale feature fusion loss function in BIBED, which further integrates low-dimensional and high-dimensional features by training and assigning weights to improve the accuracy of edge detection; Aiming at the problem of the imbalance of the proportion of positive and negative samples in the boundary images, we introduce the balanced coefficient into focal and cross-entropy loss function, to reduce the weight of a large number of simple negative samples in training to mine complex samples, to achieve the balance between positive and negative samples, and prevent overfitting of network training.
3) *BIBED-Seg Model:* We construct a two-stage semantic segmentation model based on BIBED, i.e., edge detection first and then classification. BIBED is used to guide the semantic segmentation task of high-resolution remote sensing images, i.e., to improve the final semantic segmentation results through effective edge features. In addition to this, we design a dual-loss joint constraint training, i.e., boundary loss and segmentation loss.
4) *Proof Experiments:* We conduct extensive experiments on the Potsdam, Vaihingen, and WHDLD datasets to demonstrate the effectiveness and advancement of BIBED-based semantic segmentation. Among them, in the edge detection experiments, it is verified that BIBED outperforms other state-of-the-art models for edge accuracy detection of high-resolution remote sensing images by comparison experiments. In the semantic segmentation experiments, we compare several excellent semantic segmentation networks to demonstrate further the BIBED-Seg model's ability to improve semantic segmentation accuracy.

II. Related Work

This article is related to edge detection methods based on deep learning and edge-aware segmentation.

A. Deep Learning-Based Edge Detection

In recent years, deep learning-based methods generally use convolutional neural networks to extract multilevel hierarchical features. In 2014, Bertasius et al. [33] proposed DeepEdge to achieve image contour feature extraction at different scales by building a multiscale depth network. Yaroslav et al. [32] proposed a new architecture based on combining convolutional neural networks with the nearest neighbor search for difficult image processing operations, named N4-Fields. In 2015, Shen et al. [45] proposed a new training strategy and loss function (called positive-sharing loss) to extract contour edges based on deep networks effectively. Xie et al. [34] proposed an end-to-end edge detection model that leverages the outputs from different intermediate layers with skip connections. In 2019, Liu et al. [46] introduced a richer convolutional feature, which makes good use of feature hierarchies in CNNs, for edge detection. He et al. [39] proposed a new multiscale feature output strategy, where an individual layer is supervised by labeled edges at its specific scale rather than directly applying the same supervision to all CNN outputs.

Our method aims to achieve accurate extraction of remote sensing image edges (boundaries) under complex features by multispatial-channel attention blocks and residual networks and by constructing multiscale feature extraction and training strategies.

B. Deep Learning-Based Edge-Aware Segmentation

Edge optimization and enhancement in image classification and segmentation have been a hot research direction. At first, people focused on postprocessing of classification to solve this problem, such as edge optimization of classification results by CRF [1]. Later, with the rapid development of deep learning, attention was focused on combining edge optimization with deep learning models to generate more accurate classification results, i.e., edge-aware-based classification, and semantic segmentation methods. Michieli et al. [47] proposed a novel approach (GMENet) for segmentation tasks combining object-level context conditioning, part-level spatial relationships, and shape contour information. Chen et al. [48] proposed an edge-aware convolution kernel to extract RGB-D image feature maps more efficiently using the geometric information contained in the depth channels to improve the semantic segmentation accuracy. Kuang et al. [49] proposed a new body and edge-aware network for 2-D medical image segmentation, called BEA-SegNet, which fused the body segmentation result and the edge features to get the final result.

This edge-aware approach is also commonly used in the direction of remote sensing semantic segmentation. Yuan et al. [50] combined the two tasks of cloud segmentation and cloud edge detection together to encourage better detection near cloud boundaries, resulting in an end-to-end approach for accurate cloud detection. Cheng et al. [51] proposed an edge-aware convolutional network for the segmentation of remote sensing harbor images, which was achieved by loss of edge-aware regularization.

Edge-aware joint training has been widely used in remote sensing segmentation work. Our approach is based on joint training, where the clear boundary results are jointly input to the semantic segmentation network along with the original image. It enables the overall model to allow the network to learn useful edge information in addition to loss optimization and, thus, output better and more accurate segmentation results.

III. Methodology

A. BIBED-Seg Model

In the classification of remote sensing images, the semantic segmentation based on deep learning is limited by the number and quality of training samples. The network classifies pixels by learning the spectral information of the image and the spatial semantic information between pixels. However, the classification results still have the problems of fuzzy boundaries and inaccurate positioning. And the same ground object is prone to misclassification of pixel values, resulting in the decline of classification accuracy. It is our desire to let neural networks learn useful boundary information to assist semantic segmentation tasks.

Our overall segmentation process, BIBED-Seg Net, is divided into two steps (see Fig. 2). Step 1: Edge (boundary) detection. Here, we propose and design a BIBED network for boundary extraction of high-resolution remote sensing images. Step 2: Semantic segmentation. The boundary feature map obtained by BIBED participates in the training and prediction of the semantic segmentation network. That is, the pretraining model of BIBED is used to constrain the classification network in advance, enhance the classification boundary, reduce the classification error rate between and within classes, and improve the segmentation accuracy.

B. Overview of BIBED

For high spatial resolution remote sensing images, this article proposes a semiautomatic edge detection method based on convolution neural networks with residual structure, which is used to assist the semantic segmentation of high-resolution remote sensing images. The network and edge detection process are shown in Fig. 3. First, the input high spatial resolution remote sensing image passes through the network structure of different blocks to extract the feature information at different scales; Second, the feature maps of different scales are fused; Finally, the fused feature map is binary classified by sigmoid activation function, and the final detection result is obtained.

C. Network Structure of Block-in-Block

Given the excellence of the vgg16 model [34] in edge detection, the edge detection of remote sensing images is also based on DCNN in this article. But different from the pure convolution block network structure in vgg16, we construct a block-in-block network architecture. The first block does not use the residual structure but a simple convolution structure composed of a 7×7 convolution layer, a 3×3 convolution layer, and a 3×3 max-pool layer. Others are mainly composed of

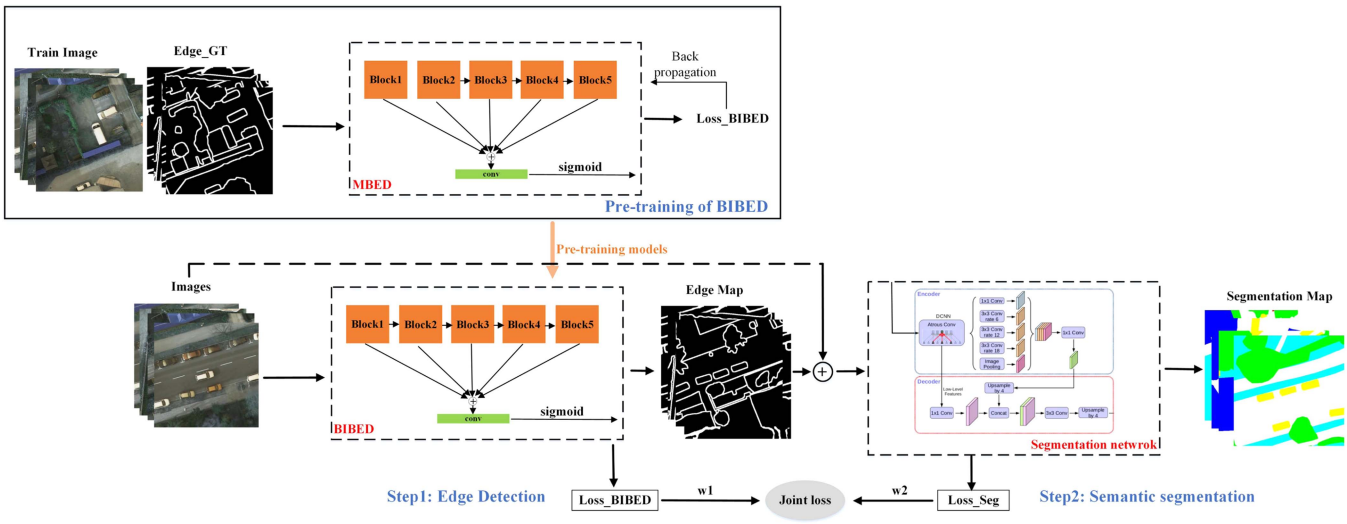

Fig. 2. Framework of BIBED-Seg model. (The semantic segmentation network structure is deeplabv3+ [42]).

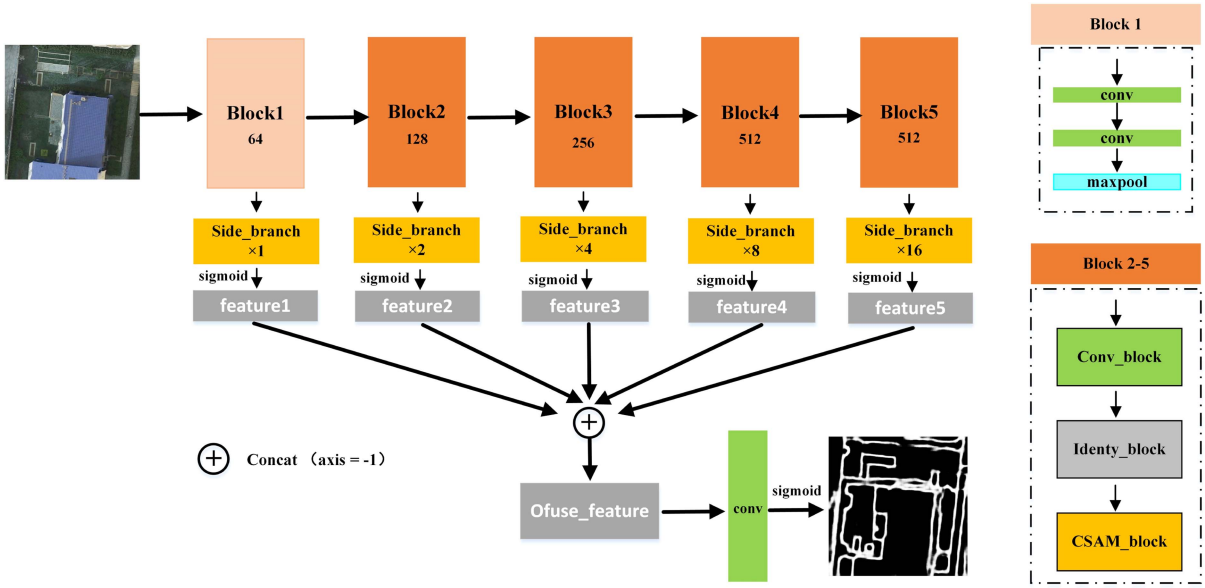

Fig. 3. Overall framework of BIBED.

three types of small blocks with residual structure: Conv_block, Identity_block, and CSAM_block (see Fig. 4).

Identity_block: A standard residual network structure, which can solve the problem of gradient disappearance while deepening the network depth, enables the model to be continuously optimized and further improves the accuracy of the model. It is suitable for recognizing and detecting high-resolution remote sensing image data with extremely complex spectral and spatial information. As shown in Fig. 4(b), the structure comprises two convolution modules and a direct skip connection. Each convolution module comprises a 3×3 convolution layer, batch normalization layer (BN), and Relu activation function. The steps of all convolution layers are 1.

Conv_block: Residual network structure with convolution skip connection. To better retain the feature information

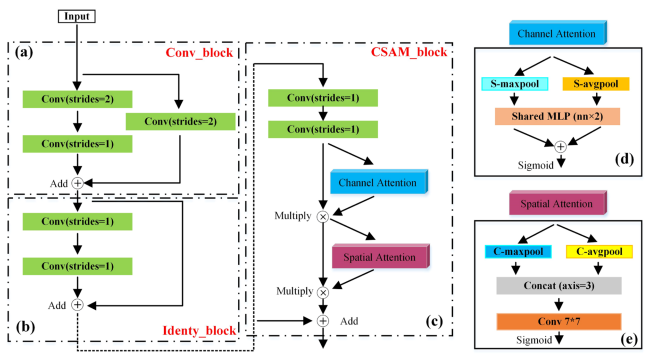

Fig. 4. Network structure of each large block (block2 – block5 in BIBED).

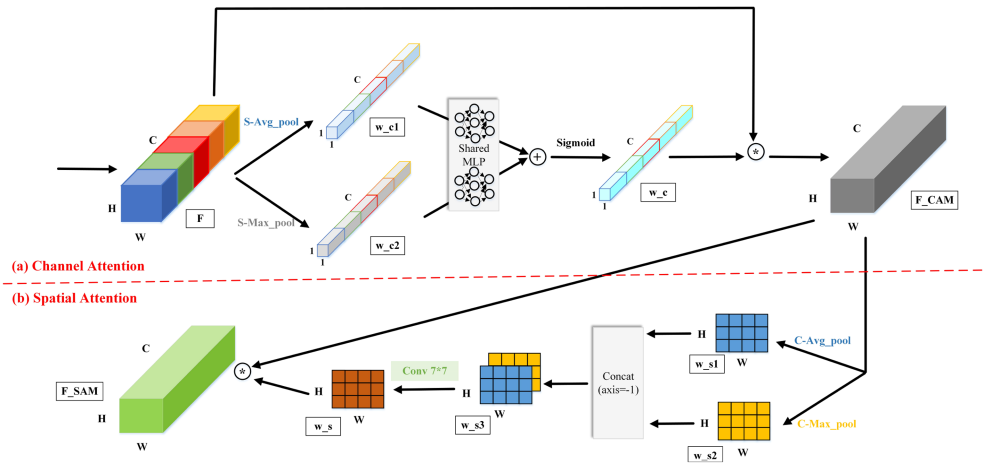

Fig. 5. Feature extraction framework of CSAM in BIBED.

extracted from the upper layer network, this structure cancels the traditional maximum pooling layer but uses a step size of 2 in the convolution layer instead of the pooling layer to achieve the purpose of downsampling. Therefore, to connect skip features effectively, adding a convolution module with the same step size of 2 to the skip connection is necessary. As shown in Fig. 4(a), the structure comprises two convolution modules and a skip connection containing one convolution block. Similarly, each convolution module comprises a 3×3 convolution layer, a BN layer, and a Relu activation function. The difference is that the step size of the convolution layer in the first convolution module on the left and the convolution layer in the skip connection is set to 2, and the rest is set to 1.

CSAM_block: Residual structure with channel and spatial attention mechanisms. Different from RGB natural images, remote sensing images usually have multiple channels. The channel attention structure [35] can focus on the channel features under the condition of compressing spatial features and give higher weight to the channel, significantly influencing the detection results. High-resolution remote sensing images have higher spatial resolution and more complex spatial correlation features. The spatial attention structure proposed by Woo [36] can focus on the spatial characteristics of images under the condition of compressing channel features, which is suitable for target detection and positioning of high-resolution images. Therefore, aiming at the difficulties of detection and positioning of edge and low detection accuracy of high-resolution remote sensing images, this article uses the method of combining channel attention mechanism (CAM) and spatial attention mechanism (SAM) with the residual network in the last small block (CSAM_bolck), in order to improve the accuracy of boundary detection of high-resolution multispectral remote sensing images. As shown in Fig. 4(c), the skip connection in the CSAM_block is composed of two 3×3 convolution modules, CAM and SAM. First, the input features after convolution pass through the CAM, then channel weight coefficient is output, and multiply it with the feature map through the convolution layer to obtain new channel features; Second, the feature is input into the SAM, and the corresponding spatial weight coefficient is output, which is multiplied with the channel feature map to obtain a new spatial feature; Finally, the new spatial feature map is fused with the output feature map from the previous piece of structure. The compression of channel and spatial features by the two attention structures is realized by global maximum pooling and global average pooling operations, as shown in Fig. 4(d) and (e).

Fig. 5 shows our method's feature extraction process of the channel and spatial attention. CAM: In the channel attention mechanism, as shown in Fig. 5(a), the global average and maximum pooling layers use different information to compress space features. The input feature (F), which is H×W×C size, can be compressed into two features, w_c1, and w_c2, with the size of 1×1×C, and then through a shared MLP structure composed of two sharable parameter neural network layers and finally added the two features output by MLP based on element-wise to obtain the channel weight coefficient feature (w_c with the size of 1×1×C) through a sigmoid activation function, where $w\_c = \{w^1, w^2, \ldots, w^c\}$, c is the number of channels in images; SAM: In spatial attention mechanism, as shown in Fig. 5(b), by using different information, the global average and maximum pooling layers can compress the input feature (F_CAM) of H×W×C into two features, w_s1 and w_s2 with the size of H×W×1, then connect the two features in the channel dimension and pass through a 7×7 convolution layer. Finally, the spatial weight coefficient (w_s with the size of H×W×1) is obtained through a sigmoid activation function,

$$\text{where } w\_s = \begin{Bmatrix} w_{0,0} & w_{0,1} & \cdots & w_{0,W} \\ w_{1,0} & w_{1,1} & \cdots & w_{1,W} \\ \vdots & \vdots & \vdots & \vdots \\ w_{H,0} & w_{H,1} & \cdots & w_{H,W} \end{Bmatrix}.$$

The specific calculation formulas of CAM and SAM in this article are as follows:

$$F\_\text{CAM} = w\_c * F = \sum_{i=1}^{c} w^i f^i \tag{1}$$

$$F\_\text{SAM} = w\_s * F\_\text{CAM} = \sum_{i=1}^{C} \sum_{h=0}^{H} \sum_{w=0}^{W} w_{h,w} f^i_{h,w}. \tag{2}$$

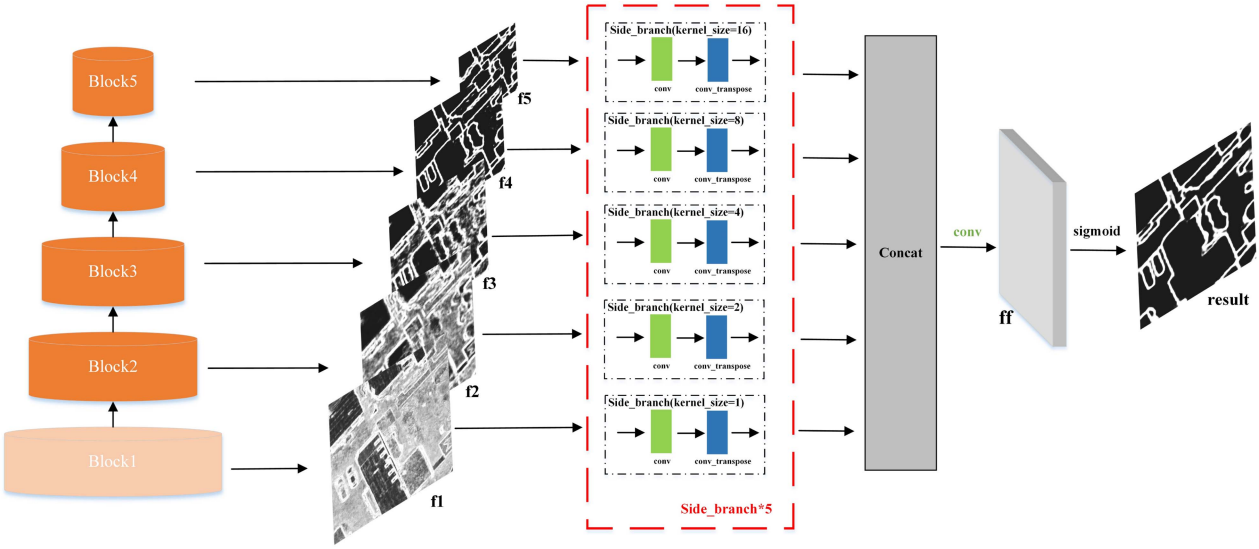

Fig. 6. Structure of the multiscale fusion extraction.

Referring to different hierarchical network structures of ResNet [37], we design different BIBED network structures of different levels (different numbers of blocks) based on the above blocks, named BIBED-N, N indicates the number of CSAM blocks in the network.

D. Edge Extraction of Multiscale Feature Fusion

Compared with the HED network, which only extracts the last layer features of vgg16. This article constructs a multifeature fusion edge extraction structure based on a multilevel and multiscale feature map. We believe that the information on features output by every Block is useful. Therefore, as shown in Fig. 6, all feature maps (f1 – f5) obtained by different blocks are fused, which can fuse information from shallow layer to deep layer and from low dimension to high dimension, retain details, and remove redundancy. Finally, the final boundary results through a 3×3 convolution layer and a sigmoid activation function. Where side_branch acts as the effect of upsampling, the reduced dimension features are restored to the original image size to facilitate subsequent feature fusion. It is mainly composed of convolution and anticonvolution layers (Up-sampling layers).

E. Loss Function

BIBED-Seg has two loss common constraints (Loss of BIBED and loss of segmentation), and its loss function can be expressed as

$$L_S = w_1 L_{\text{BIBED}} + w_2 L_{Seg}. \tag{3}$$

Since L_S is oriented to the classification task and BIBED will be pretrained separately, we will focus on L_{Seg} in the segmentation task so that $w_1 = 0.3$ and $w_2 = 0.7$ in this article. Where the loss of segmentation is generally a multiclass cross-entropy loss or multiclass Dice loss function. And the edge detection loss function will be the focus of the research in this article.

Based on the abovementioned multiscale fusion ideal, this article designs a multilevel feature fusion loss function. A total of six losses are produced in the network training process, including the loss of features at five different scales (Block1-Block5) L_{side}, and the loss of fusion feature L_{fuse}. The six losses are trained at the same time. The L_{side} and L_{fuse} in the training process are weighted and summed to obtain the final overall loss, as shown in

$$L_{\text{BIBED}} = w_{\text{side}} L_{\text{side}} + w_{\text{fuse}} L_{\text{fuse}} \tag{4}$$

$$L_{\text{side}} = \sum_{i=1}^{n} \lambda_i l_i. \tag{5}$$

L_{MBED} is the overall loss function of the network, w_{side} and w_{fuse} are weights for the loss of side features and fused features ($w_{\text{side}} = 0.4$, $w_{\text{fuse}} = 0.6$), l_i is the side loss function of the output of the *Block_i*, λ_i is the weight coefficient of the *feature_i* in different blocks, n is the number of Blocks ($n = 5$). Since our experiment focuses on accurate image boundary detection instead of complex edge information, we give greater weight to the deeply extracted features considering feature fusion. So, the λ_i in this article are set as 0.1, 0.2, 0.3, 0.3, and 0.1, respectively.

For the overfitting problem of network training due to the imbalance of positive and negative samples in the edge map, the loss functions with balance factors in favor of binary classification are used.

The loss functions of L_{side} in this article use the focal loss function [38] with balanced factors, as shown in (6). The loss function can excavate complex samples by reducing the weight of many simple negative samples in training to solve the problem of the unbalanced proportion of positive and negative samples in the remote sensing edge detection task in this article

$$\text{loss} = \begin{cases} -\beta(1-y')^{\gamma} \log y', & y = 1 \\ -(1-\beta)\, y'^{\gamma} \log(1-y'), & y = 0 \end{cases} \tag{6}$$

where y' is the probability value output through the sigmoid activation function, generally between 0-1; y represents the actual value (0 or 1) of ground truth; β is the balance factor that can balance the positive and negative samples, β is set to 0.7 in this article, the larger the β, the higher the attention to positive samples; γ factor can solve the problem that simple and easy samples cannot be distinguished, which is generally greater than 1, γ is set to 2. For positive samples, the higher the prediction probability y' , the smaller the loss value. For negative samples, the smaller the prediction probability y', the greater the loss value.

The loss function of L_{fuse} in this article designs a cross-entropy loss with class-balanced. Due to inconsistent annotations by different annotators, we introduce a threshold for loss computation. For a ground truth $Y = \{y_j, j = 1, 2, \ldots |num\_\text{pixels}|\}$, $y_j \in [0, 1]$, we define a constant ε, $Y^+ = \{y_j, y_j > \varepsilon\}$ and $Y^- = \{y_j, y_j = 0\}$. Only pixels corresponding to Y^+ and Y^- are computed in loss function. So, we define L_{fuse} as

$$L_{\text{fuse}}(Y', Y) = -\alpha \sum_{j \in Y^-} \log(1 - y'_j) - \theta \sum_{j \in Y^+} \log(y'_j) \quad (7)$$

where $Y' = \{y'_j, j = 1, 2, \ldots, |num\_\text{pixels}|\}$, $y_j \in (0, 1)$ denotes a predicted edge map, α and θ are balanced factors that balance the edge and nonedge pixels $\alpha = \mu \cdot |Y^+|/(|Y^+| + |Y^-|)$, $\theta = |Y^-| /(|Y^+| + |Y^-|)$, μ controls the weight of positive samples relative to negative samples.

F. Evaluation Index

To quantitatively evaluate the extracted boundary results, we first need to binarize the results, taking values of 0 and 1, which requires us to set a threshold for the detection results. There are two threshold methods. The first is the optimal dataset scale (ODS), which adopts a fixed threshold for all images in the dataset to maximize the F-score on the whole dataset. The second is called the optimal image scale (OIS), which selects an optimal threshold for each image to maximize the F-score of the image. We report the F-measure of both ODS and OIS in our experiments. F-measure is the harmonic average of accuracy (P) and recall (R), which is expressed by the following formula:

$$F = \frac{2 \cdot \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (8)$$

In addition, we quantitatively evaluate the edge detection results by average precision (AP) and mean intersection over Union (MIOU), which is a typical measure of semantic segmentation. It is evaluated by calculating the ratio of the intersection and Union of the actual value and the predicted value

$$\text{MIOU} = \frac{1}{n} \sum_{i=1}^{n} \frac{x_{ii}}{\sum_{j=1}^{n} x_{ij} + \sum_{j=1}^{n} x_{ji} - x_{ii}} \quad (9)$$

where n denotes the total number of categories (for segmentation, n = 6), x_{ii} denotes the number of correctly classified pixels in each class, and x_{ij} denotes the number of predictions of *i* to *j*, i.e., prediction errors FN. x_{ji} denotes the number of predictions of *j* to *i*, i.e., prediction errors FP. ($i = 01, \ldots, n$; $j = 01, \ldots, n$).

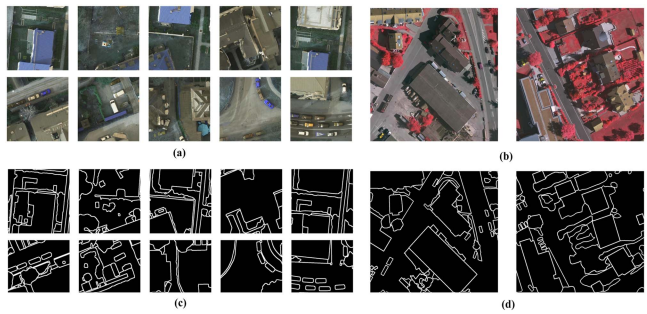

Fig. 7. Boundary data. (a) and (c) are the images and boundary ground truth maps of the Potsdam dataset with the size of 512×512, (b) and (d) are the images and boundary ground truth maps of the Vaihingen dataset with the size of 1024×1024.

IV. Experiment and Results

A. Experimental Data

Experiments are performed on the Potsdam and Vaihingen 2-D dataset of International Society for Photogrammetry and Remote Sensing (ISPRS) to assess the performance of the proposed method in edge detection and segmentation.

Potsdam dataset: Potsdam data set provides 38 images with four bands (R, G, B, NIR) from a typical historic city with giant building blocks, narrow streets, and dense settlement structures in Germany with a pixel resolution of 6000×6000 and spatial resolution of 0.05 m. The semantic segmentation's ground truth with and without boundaries contains six most common land cover categories, which are labeled in different colors: impervious surfaces (white); low vegetation (cyan); trees (green); buildings (blue); cars (yellow); backgrounds (red). We crop each image into 100 patches with the size of 512×512 for training and testing. The number of images in the training, validation, and test data sets is 2400, 600, and 800.

Vaihingen dataset: Vaihingen data set provides 33 images with three bands (R, G, B) from a relatively small village with many detached buildings and small multistory buildings in Germany with a spatial resolution of 0.09 m. The semantic segmentation's ground truth with and without boundaries contains six most common land cover categories, which are labeled in different colors: impervious surfaces (white); low vegetation (cyan); trees (green); buildings (blue); cars (yellow); and backgrounds (red). We crop each image into 20 patches with the size of 1024×1024 for training and testing. The number of images in the training, validation, and test data sets is 450, 50, and 160.

Edge ground truth acquisition: Based on the classified ground truth data with boundary and without boundary, the difference processing is carried out to obtain the boundary edge map we need, see Fig. 7.

B. Performance on BIBED

This part mainly discusses the effectiveness of the BIBED-N network with different channel-spatial attention layers in remote sensing image boundary extraction, and we believe that too much

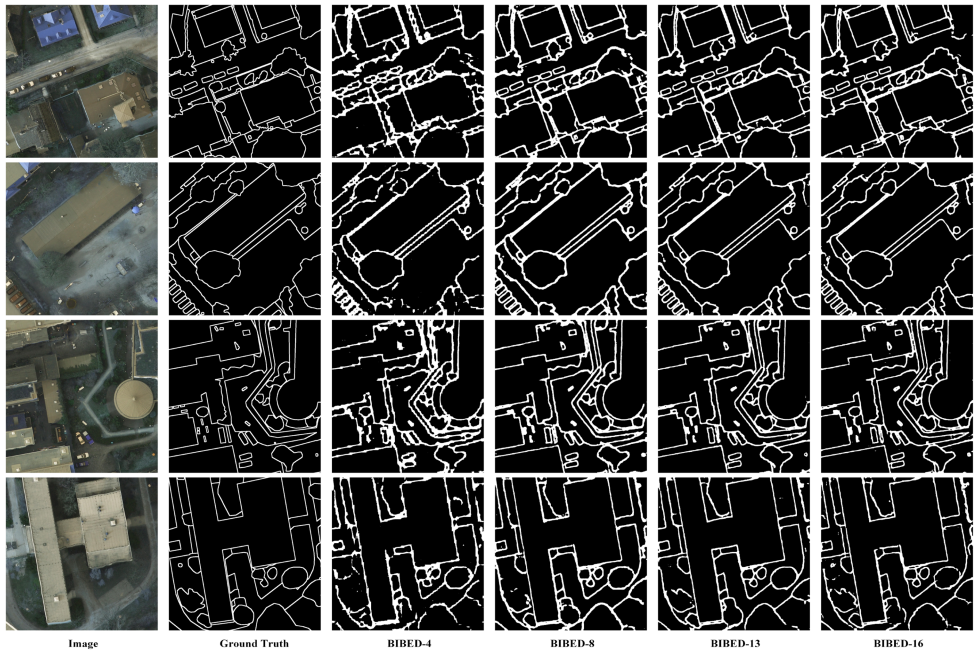

Fig. 8. Boundary detected by BIBED-N on the Potsdam dataset.

TABLE I
NETWORK STRUCTURES OF BIBED WITH DIFFERENT CSAM BLOCKS

	Fiter	BIBED-4	BIBED-8	BIBED-13	BIBED-16
Block_1	64	conv 7×7, 64, stride 1 conv 3×3, 64, stride 1 max pool 3×3, stride 2			
	64	side_branch, factor 2			
Block_2	128	Conv_block×1 Identity_block×1 CSAM_block×1	Conv_block × 1 Identity_block × 1 CSAM_block × 2	Conv_block×1 Identity_block×1 CSAM_block×2	Conv_block×1 Identity_block×1 CSAM_block×2
	128	side_branch, factor 4			
Block_3	256	Conv_block×1 Identity_block×1 CSAM_block×1	Conv_block×1 Identity_block×1 CSAM_block×2	Conv_block×1 Identity_block×1 CSAM_block×3	Conv_block×1 Identity_block×1 CSAM_block×3
	256	side_branch, factor 8			
Block_4	512	Conv_block×1 Identity_block×1 CSAM_block×1	Conv_block×1 Identity_block×1 CSAM_block×2	Conv_block×1 Identity_block×1 CSAM_block×6	Conv_block×1 Identity_block×1 CSAM_block×9
	512	side_branch, factor 16			
Block_5	512	Conv_block×1 Identity_block×1 CSAM_block×1	Conv_block×1 Identity_block×1 CSAM_block×2	Conv_block×1 Identity_block×1 CSAM_block×2	Conv_block×1 Identity_block×1 CSAM_block×2
	512	side_branch, factor 32			
	1	Concatenate (axis=-1), Conv 1×1, 1, sigmoid			

or too little CSAM will affect the edge detection results. Here, we mainly design four BIBED-N networks, and several experiments are carried out on Potsdam and Vaihingen data sets.

1) Potsdam Dataset: In this section, we conduct experiments on the Potsdam data set with the BIBED-N network. The detection results are shown in Fig. 8. When the number of CSAM blocks is few, such as BIBED-4, due to too few convolution layers, the edge detection result is poor, and the edge is rough and easy to break. With the increase in the number of CSAM blocks, the detection effect is significantly improved, in which BIBED-13 and BIBED-16 perform well, and the detection of small targets and edge fractures is improved considerably. Table II records the performance ability of BIBED-N under the Potsdam data set, including edge evaluation, parameters, and network computation. The best performance is denoted in bold, and the second-best is marked with underlines in the table. As can be seen from Table II, the increase in the number of CSAM_block will effectively improve the accuracy of edge detection on

TABLE II
VALIDITY OF EDGE DETECTION FROM BIBED-N ON POTSDAM DATASET

Methods	ODS-F	OIS-F	AP	Edge-IOU	Param.	Flops	Times (s)
BIBED-4	0.5646	0.6036	0.6489	0.429	33.64M	70.51M	5400
BIBED-8	0.6136	0.6573	0.7033	0.453	44.14M	92.54M	6370
BIBED-13	**0.6671**	0.7275	**0.7668**	**0.522**	63.42M	132.97M	9504
BIBED-16	0.6644	**0.7298**	0.7627	0.517	77.03M	161.51M	11850

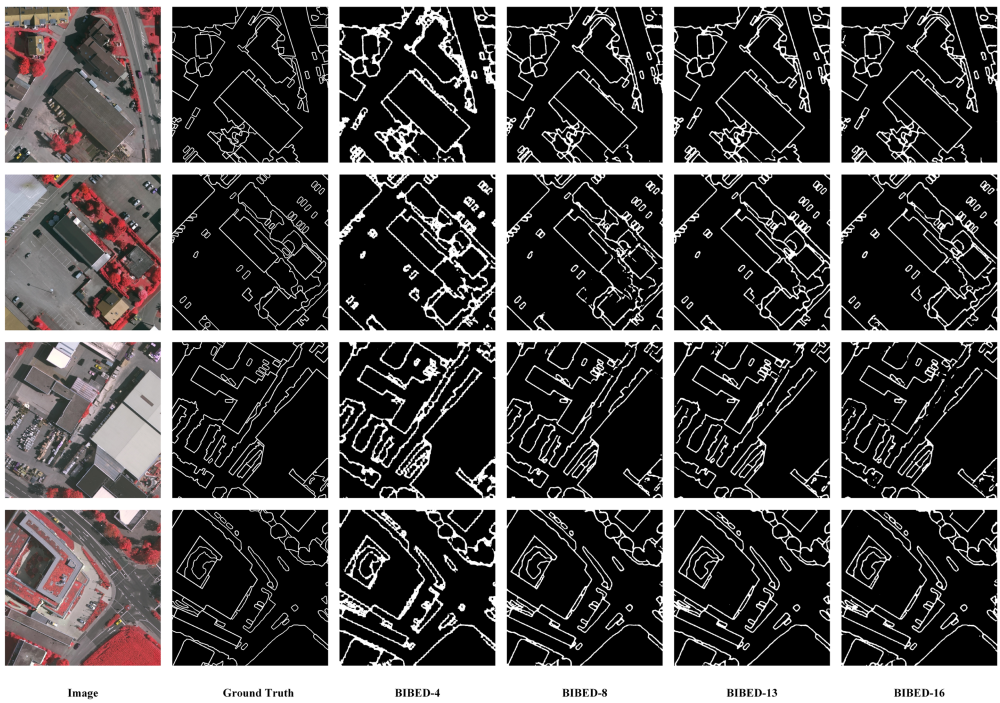

Fig. 9. Boundary detected by BIBED-N on Vaihingen dataset.

TABLE III
VALIDITYO OF EDGE DETECTION FROM BIBED-N ON VAIHINGEN DATASET

Methods	ODS-F	OIS-F	AP	Edge-IOU	Param.	Flops	Times (s)
BIBED-4	0.6832	0.7016	0.7427	0.523	33.64M	70.51M	7000
BIBED-8	0.7199	0.7327	0.7673	0.562	44.14M	92.54M	8400
BIBED-13	**0.7432**	0.7868	0.8034	**0.617**	63.42M	132.97M	10920
BIBED-16	0.7375	**0.7879**	**0.8047**	0.595	77.03M	161.51M	13200

BIBED, although the amount of calculation will also increase. However, compared with 4 and 8 blocks, 13 blocks significantly improve, while 16 blocks have no significant improvement in detection accuracy compared with 13 blocks. On the contrary, ODI-F and Edge-IOU have decreased, and calculation and time have also increased significantly. Therefore, BIBED-13 performs best for the Potsdam dataset, with the highest ODS-F and Edge-IOU of 0.6671 and 0.522, respectively, while OIS-F is 0.7275, only 0.23% lower compared to BIBED-16.

2) Vaihingen Dataset: In this section, we conduct experiments on the Vaihingen data set with the BIBED-N network. The detection results are shown in Fig. 9. The statistics of edge evaluation, parameters, and network computation are shown in Table III. Like the Potsdam data set, BIBED-13 also showed the best effect on the Vaihingen data set, with ODS-F of 0.7432 and Edge-IOU of 0.617.

To sum up, this article finally selects BIBED-13 as the final leading network for edge detection of remote sensing images, including subsequent ablation experiments, comparison experiments, and BIBED-Seg semantic segmentation, completed based on the BIBED-13 network.

Then, we discuss the improvement effect of multilevel feature fusion results on edge detection. Fig. 10 shows the output features of the different blocks (Block1-Block5) and the results of the fusion features from BIBED-13. It can be seen that the output features gradually become coarse as the number of convolution layers deepens. The convolutional features of the output (f) of block5 become blurred, and the intermediate layers

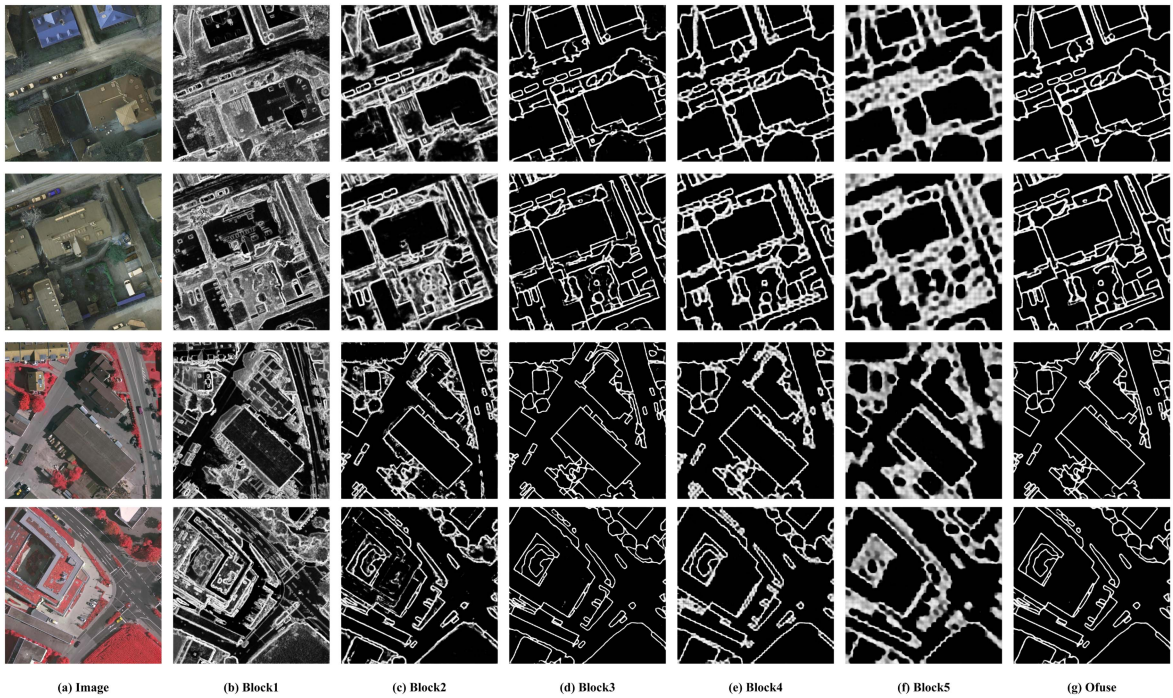

Fig. 10. Side output is produced by BIBED-13 (b)–(f). (g) Represents the edge feature after fusion of (b)–(f). The upper parts are selected from the Potsdam data set. The lower half is selected from the Vaihingen data set.

TABLE IV
VALIDITY OF SIDE BRANCH FROM BIBED-13

Block	Dataset	ODS-F	Edge-IOU	Dataset	ODS-F	Edge-IOU
Block1(b)	Potsdam	0.356	0.217	Vaihingen	0.402	0.289
Block2(c)		0.439	0.281		0.487	0.361
Block3(d)		**0.629**	**0.458**		**0.724**	**0.567**
Block4(e)		0.592	0.420		0.698	0.537
Block5(f)		0.441	0.282		0.492	0.368

(b)–(e) contain essential details that do not appear and are very important. Therefore, we believe it is necessary to fuse such crucial information, and (g) is the edge feature after fusing the intermediate output features. It can be seen that the output features (g) become clear without missing the critical boundary information and the rich texture information at the beginning, which is the boundary feature we want to get.

However, we find that the features of the network's first block (block1) are too rich and contain too much complex texture information, and the features of the last block (block5) are too fuzzy, resulting in excessive edge adhesion. As shown in Table IV, the ODS-F and Edge-IOU of the output characteristics of block1 and block5 are much lower than those of other blocks. These are the factors that affect the accuracy of results from fusion features. To solve this problem, we put forward two ideas: Idea 1: Remove the last block structure of the network and build a four-block network. The results show that boundary extraction has not been improved, but the accuracy of ODS-F has decreased by 4.5% on the Potsdam data set and 2.03% on the Vaihingen data set (see Table V). It shows that the effect of the five-block structure is better than that of the four-block structure. Although the output characteristics of the last block are fuzzy, it still plays an essential role in the side-loss training and fusion stage. Idea 2: In the feature fusion and side branch training stage, we focus on the features of block-2,3,4 on the network, especially the block-34. Therefore, during training, the weights of the side loss are set as 0.1, 0.2, 0.3, 0.3, and 0.1, respectively. As shown in Table V, the idea of weighted training and fusion of Idea 2 has achieved satisfactory results on both data sets. Compared with the direct summation, its edge evaluation gets the highest score. On the Potsdam dataset, ODS-F is 66.71%, and Edge-IOU is 0.522, an increase of 2.24% and 0.032, respectively, over the normal fusion. On the Vaihingen dataset, ODS-F is 74.32%, and

TABLE V
VALIDITY OF EDGE DETECTION FROM BIBED-13 BASED ON OUR IDEA

Dataset	Methods	ODS-F	OIS-F	Edge-IOU
Potsdam	Ordinary	0.6447	0.7003	0.490
	Idea 1	0.6114	0.6744	0.440
	Idea 2	**0.6671**	**0.7275**	**0.522**
Vaihingen	Ordinary	0.7239	0.7689	0.579
	Idea 1	0.7016	0.7336	0.553
	Idea 2	**0.7432**	**0.7868**	**0.617**

TABLE VI
VALIDITY OF DIFFERENT ATTENTION MECHANISMS BASED ON BIBED

Components	Dataset	ODS-F	Dataset	ODS-F
BIBED with non-local	Potsdam	0.6498	Vaihingen	0.7257
BIBED with DANet		0.6601		0.7398
BIBED with CCA		0.6589		0.7401
BIBED with CSAM		**0.6671**		**0.7432**

Edge-IOU is 0.617, an increase of 1.93% and 0.038, respectively, over the normal fusion.

C. Ablation Study of BIBED

To better understand our model and prove the effectiveness of each module, we conducted ablation experiments using our testing dataset.

Effect of different attention mechanisms: To demonstrate the effectiveness and sophistication of the CSAM mechanism, we also introduced other attention mechanisms into the BIBED model for ablation experiments. The selected other attention mechanisms include nonlocal self-attention [54], DANet dual attention [55], and criss-cross attention [56]. Table VI shows the ODS-F of the edge detection results of this group ablation experiment on the ISPRS dataset, and it can be seen that the proposed structure combining CSAM in this article is more suitable for remote sensing boundary detection than several other attention structures.

Effect of the network components: First, we discuss the performance of our proposed method with CSAM (residual structure with channel and spatial attention mechanisms). The experiments were conducted on BIBED-13, and several experiments were carried out: edge detection of removing CAM (residual structure with SAM), removing SAM (residual structure with CAM), removing CSAM (only residual structure), and pure convolution (vgg16), respectively. And we evaluate the ODS-F, OIS-F, and Edge-IOU on the Potsdam and Vaihingen data set. As shown in Table VII, compared with the pure convolution network, the residual convolution structure can improve the edge detection ability of the network, improving the ODS-F from 0.5774 to 0.6133 and Edge-IOU from 0.433 to 0.467 for Potsdam data set. For the Vaihingen data set, the ODS-F and Edge-IOU also increased by 4.68% and 3.7%, respectively. And the residual structure combining channel and spatial attention mechanisms allows the network detection capability to be improved even more. The best performance is achieved by using both attention modules together. Compared with the residual convolution network, combined with CSAM can improve the ODS-F from 0.6133 to 0.6671, OIS-F from 0.6487 to 0.7275, and Edge-IOU from 0.467 to 0.522 for the Potsdam data set. For the Vaihingen data set, the ODS-F, OIS-F, and Edge-IOU also increased by 4.68%, 6.89%, and 5.3%, respectively.

Fig. 11 shows the visualization results of the boundary detection with different network components. The top three rows show the results of the Potsdam data set, and the bottom two are for the Vaihingen data set. In comparison, the boundary of ground objects detected by the pure convolution network based on vgg16 is relatively blurred. There are problems such as false detection, missing detection, and line discontinuity in the boundaries of many ground objects (small buildings, trees, and low vegetation). The network with residual structures alleviates these problems, but the boundaries of the detected features are still blurred, and the limits of many different types of parts are connected. The proposed method, i.e., the residual convolution network combining CSAM, is best for edge detection on both data sets. The detected boundaries are relatively straightforward. The results have the sharpest edges of small targets, sparse vegetation, and trees, alleviating the problem of boundary adhesion.

Therefore, it can be concluded that the method proposed in this article, which introduces both channel attention and spatial attention in the residual convolution neural network, can effectively improve the edge detection capability of high-resolution remote sensing images.

Effect of different block nums: Second, we discuss the influence of different numbers of blocks in the proposed method on the edge detection accuracy of remote sensing images. The experiments are conducted with the BIBED-13 network as the backbone, and the number of network blocks is analyzed from 2 to 6, respectively, while keeping the number of CSAM_block consistent. The four precisions of ODS-F, OIS-F, AP, and Edge-IOU with different numbers of blocks are recorded, as shown in Fig. 12. The line graph on the left is from the Potsdam dataset, and the right is from Vaihingen. It is found that the accuracy of edge detection increases with the number of blocks at the beginning. When the number of blocks in the network is five, the edge detection accuracy of both data sets reaches the highest, while when the number of blocks is 6, the accuracy decreases. It is because the number of blocks is too large. The edge features extracted will become more blurred, and a lot of helpful information will be canceled out, leading to a decrease in accuracy after feature fusion. Therefore, the leading network of this article selects five blocks as the optimal number of blocks.

D. Comparisons With Other State-of-the-Art (Edge Detection)

In this part, we compare the proposed method with other edge detection algorithms, including the Canny algorithm [26], DeepContour [33], HED [34], CED [16], BDCN [39], PiDiNet [52], and eGAN [53]. The experimental data are completed on the Potsdam and Vaihingen data set.

Potsdam data set: The comparison of Potsdam is summarized in Table VIII. The best performance is denoted in bold, and the second-best is marked with underlines in the table. BIBED-13 obtains the ODS F-Measure of 0.6671, the OIS F-Measure of 0.7275, the AP of 0.7668, and the Edge-IOU of 0.522, outperforming all other competing methods. Compared with the VGG16-based HED detection algorithm, the ODS-F and Edge-IOU of our method improved by 4.9% and 6.7%, respectively. This is followed by the SEM module-based BDCN network, as well as PiDiNet and eGAN, with similar detection results. Take BDCN, for example, whose detection results of ODS-F and Edge-IOU can reach 0.6446 and 0.513. But both are lower than our proposed BIBED-13 network. In addition, the traditional Canny detection algorithm does not perform well in front of other deep learning models, with an ODS-F of only

TABLE VII
VALIDITY OF DIFFERENT NETWORK COMPONENTS FROM BIBED-13 (CAM IN THE TABLE REPRESENTS CHANNEL ATTENTION MECHANISM, SAM IN THE TABLE REPRESENTS SPATIAL ATTENTION MECHANISM)

Components	Dataset	ODS-F	OIS-F	Edge-IOU	Dataset	ODS-F	OIS-F	Edge-IOU
Pure conv	Potsdam	0.5774	0.6163	0.433	Vaihingen	0.6429	0.6881	0.527
Res_conv		0.6133	0.6487	0.467		0.6897	0.7179	0.564
Res+CAM		0.6509	0.7044	0.499		0.7055	0.7422	0.587
Res+SAM		0.6537	0.7028	0.509		0.7396	0.7692	0.602
Res+CSAM		**0.6671**	**0.7275**	**0.522**		**0.7432**	**0.7868**	**0.617**

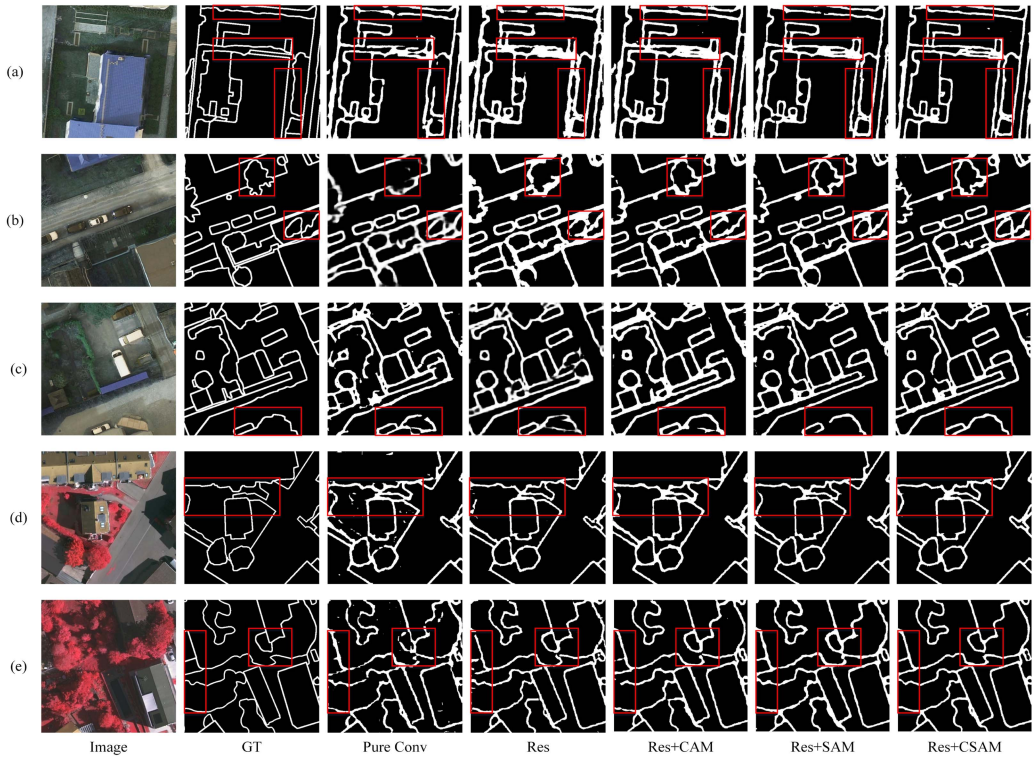

Fig. 11. Visualizations of edge detection results with or without CSAM (Ablation study). (a)–(c) are selected from the Potsdam data set. (d)–(e) are selected from the Vaihingen data set.

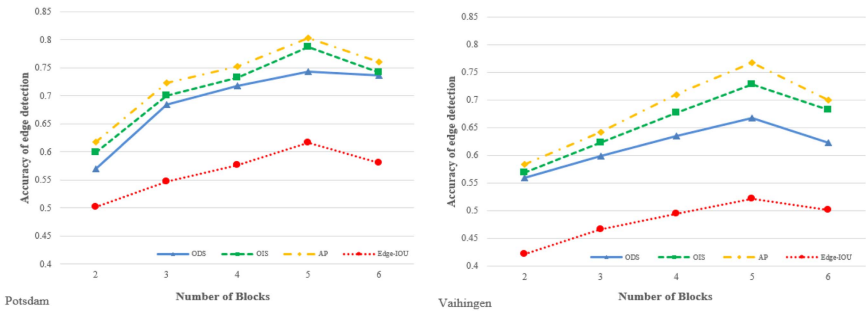

Fig. 12. Comparison of edge detection accuracy with different numbers of blocks (2 to 6).

0.4097 and an Edge-IOU of only 0.237, the worst performance among all compared methods. This also indicates that the Canny algorithm is unsuitable for high-resolution remote sensing images' interclass boundary feature extraction.

To form a more intuitive comparison, we visualize the edge detection results (binarization to 0 and 255) in Fig. 13. The first three rows are the detection results of the Potsdam dataset. It can be seen that the Canny algorithm has the worst accuracy because its detection operator is more inclined to image edge calculation than boundary extraction, and its detection results have more false edges. And we debugged a variety of minimum and maximum thresholds, which cannot solve this problem well.

TABLE VIII
VALIDITY OF BOUNDARY DETECTION RESULTS IN DIFFERENT METHODS ON THE POTSDAM DATASET

Methods	ODS-F	OIS-F	AP	Edge-IOU
Canny	0.4097	0.4147	-	0.237
DeepContour	0.5932	0.6075	0.6388	0.438
HED	0.6181	0.6388	0.6634	0.455
CED	0.6397	0.7179	0.7403	0.507
BDCN	0.6446	0.7154	0.7432	0.513
eGAN	0.6408	0.7163	0.7501	0.509
PiDiNet	0.6478	0.7167	0.7455	0.511
BIBED-13	**0.6671**	**0.7275**	**0.7668**	**0.522**

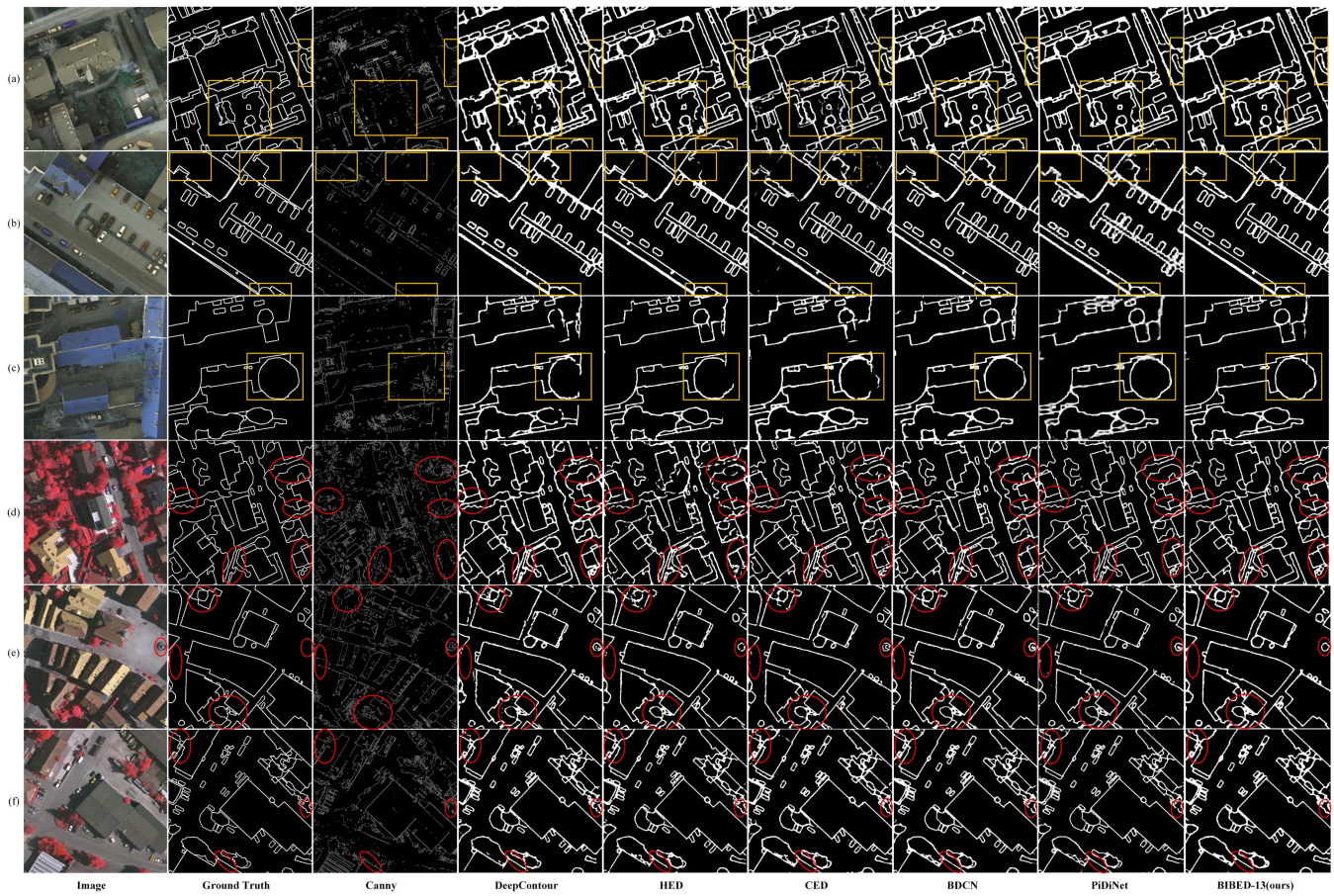

Fig. 13. Visualizations of edge detection results output by our proposed BIBED and other baseline methods for comparison. (a)–(c) are selected from the Potsdam data set. (d)–(f) are selected from the Vaihingen data set.

If the threshold is too large, then too few edges are detected, resulting in the boundary not being detected either. Our proposed BIBED-13 network detects the boundary better than other deep learning models, and the results are closer to the ground truth boundary map. It shows better learning and expression ability in boundary detection of special features (small targets and nonregular features), and the continuity of the boundary is improved.

Vaihingen data set: The comparison of Vaihingen is summarized in Table IX. The best performance is denoted in bold, and the second-best is marked with underlines in the table. BIBED-13 obtains the ODS F-Measure of 0.7432, the OIS F-Measure of 0.7868, the AP of 0.8034, and the Edge-IOU of 0.522, which also outperforms all other competing methods. Compared to HED, the four accuracies are 4.51%, 7.57%, 7.62%, and 3.8% higher, respectively. Compared to the second-best performer, PiDiNet, the four accuracies improved by 0.96%, 2.31%, 2.11%, and 0.6%, respectively. And the Canny algorithm still performs the worst. As shown in Fig. 10, the bottom three rows are the detection results of the Vaihingen dataset. Compared to

TABLE IX
VALIDITY OF BOUNDARY DETECTION WITH DIFFERENT METHODS ON THE VAIHINGEN DATASET

Methods	ODS-F	OIS-F	AP	Edge-IOU
Canny	0.4886	0.5122	-	0.398
DeepContour	0.6823	0.7057	0.6977	0.568
HED	0.6981	0.7111	0.7272	0.579
CED	0.7288	0.7629	0.7677	0.591
BDCN	0.7304	0.7645	0.7719	0.604
eGAN	0.7293	0.7733	0.7814	0.599
PiDiNet	0.7336	0.7637	0.7823	0.611
BIBED-13	**0.7432**	**0.7868**	**0.8034**	**0.617**

TABLE X
VALIDITY OF SEMANTIC SEGMENTATION RESULTS BY DIFFERENT BIBED-SEG MODEL WITH DIFFERENT SEGMENTATION NETWORKS

Methods	Dataset	OA	Kappa	MIOU	Dataset	OA	Kappa	MIOU	Dataset	OA	Kappa	MIOU
FCN	Potsdam	79.2	73.6	61.9	Vaihingen	70.2	61.6	53.6	WHDLD	73.7	62.7	47.9
BIBED-FCN		84.9	80.3	69.1		77.8	70.8	61.7		78.6	68.8	55.8
UNet		81.7	76.5	65.1		79.9	74.3	58.8		74.9	61.6	48.4
BIBED-UNet		87.5	84.9	72.4		84.6	79.6	66.7		80.0	70.2	57.5
Deeplabv3+		85.5	81.1	70.3		83.7	78.3	64.0		76.3	64.8	51.8
BIBED-v3+		**90.2**	**89.8**	**76.0**		**87.7**	**83.7**	**69.6**		**81.5**	**73.4**	**61.3**

the Potsdam data set, this data set has more buildings and is heavily wooded with more prominent boundaries, so the overall detection is better.

E. Performance of BIBED-Seg Net

This part mainly discusses the influence of boundary features on pixel-level semantic segmentation results for high-resolution remote sensing images by our proposed method BIBED-Seg Net. To verify the validity and reliability of our proposed method, we conducted several experiments on the Potsdam, Vaihingen, and WHDLD datasets. WHDLD contains 4940 RGB images in 256×256 captured by Gaofen 1 Satellite and ZY-3 Satellite over Wuhan urban area. By image fusion and resampling, the resolution of the images is to reach 2 m/pixel. The images are labeled with six classes, i.e., bare soil, building, pavement, vegetation, road, and water. The training datasets for the semantic segmentation experiments in this article are all effectively data-augmented and image-enhanced, and the experimental results are more stable.

First, we conduct trial experiments on several advanced semantic segmentation models, namely, FCN [40], UNet [41], and Deeplab V3+ [42]. These models are combined with the BIBED network proposed in this article, thus forming the new boundary-based semantic segmentation networks BIBED-FCN, BIBED-UNet, and BIBED-V3+. We visualized and compared the classification results of the BIBED-Seg network with the original semantic segmentation network, as shown in Fig. 14. It can be seen that the semantic segmentation results of the network combined with BIBED are significantly improved compared with the semantic segmentation results without edge information. The interclass boundaries are more accurate and precise, and the number of misclassified pixels and noise points within classes is significantly reduced. In particular, the classification results are more accurate for those irregular targets, such as trees, low vegetation, and rounded buildings.

TABLE XI
VALIDITY OF SEMANTIC SEGMENTATION RESULTS IN DIFFERENT METHODS EDGE-SEG ON POTSDAM DATASET

Methods	OA	Kappa	MIOU
Canny-Seg	65.6	63.2	51.5
HED-Seg	86.4	84.8	70.7
BDCN-Seg	88.5	87.3	73.6
BIBED-Seg	**90.2**	**89.8**	**76.0**

The accuracy statistics of the splitting results are shown in Table X. The best performance is denoted in bold, and the results of BIBED-Seg Net are marked with underlines in the table. When combined with valid boundary information, 1) compared to FCN, BIBED-FCN improved the overall accuracy of segmentation results by 5.7% and MIOU by 7.2% for the Potsdam dataset. The OA improved by 7.6%, and MIOU improved by 7.7% for the Vaihingen dataset. The OA improved by 4.9%, and MIOU improved by 7.9% for the WHDLD dataset. 2) Compared to UNet, BIBED-UNet improved the OA of segmentation results by 5.8% and MIOU by 7.3% for the Potsdam dataset. The OA improved by 4.7%, and MIOU improved by 7.9% for the Vaihingen dataset. The OA improved by 5.1%, and MIOU improved by 9.1% for the WHDLD dataset. 3) Compared to Deeplabv3+, BIBED-v3+ improved the OA of segmentation results by 4.7% and MIOU by 5.7% for the Potsdam dataset. For the Vaihingen dataset, the OA improved by 4.0%, and MIOU improved by 5.6%. For the WHDLD dataset, the OA improved by 5.2%, and MIOU improved by 9.5%. It can be seen that BIBED-Seg can significantly improve the accuracy of semantic segmentation of Seg networks, especially IOU. BIBED-v3+ performs the best

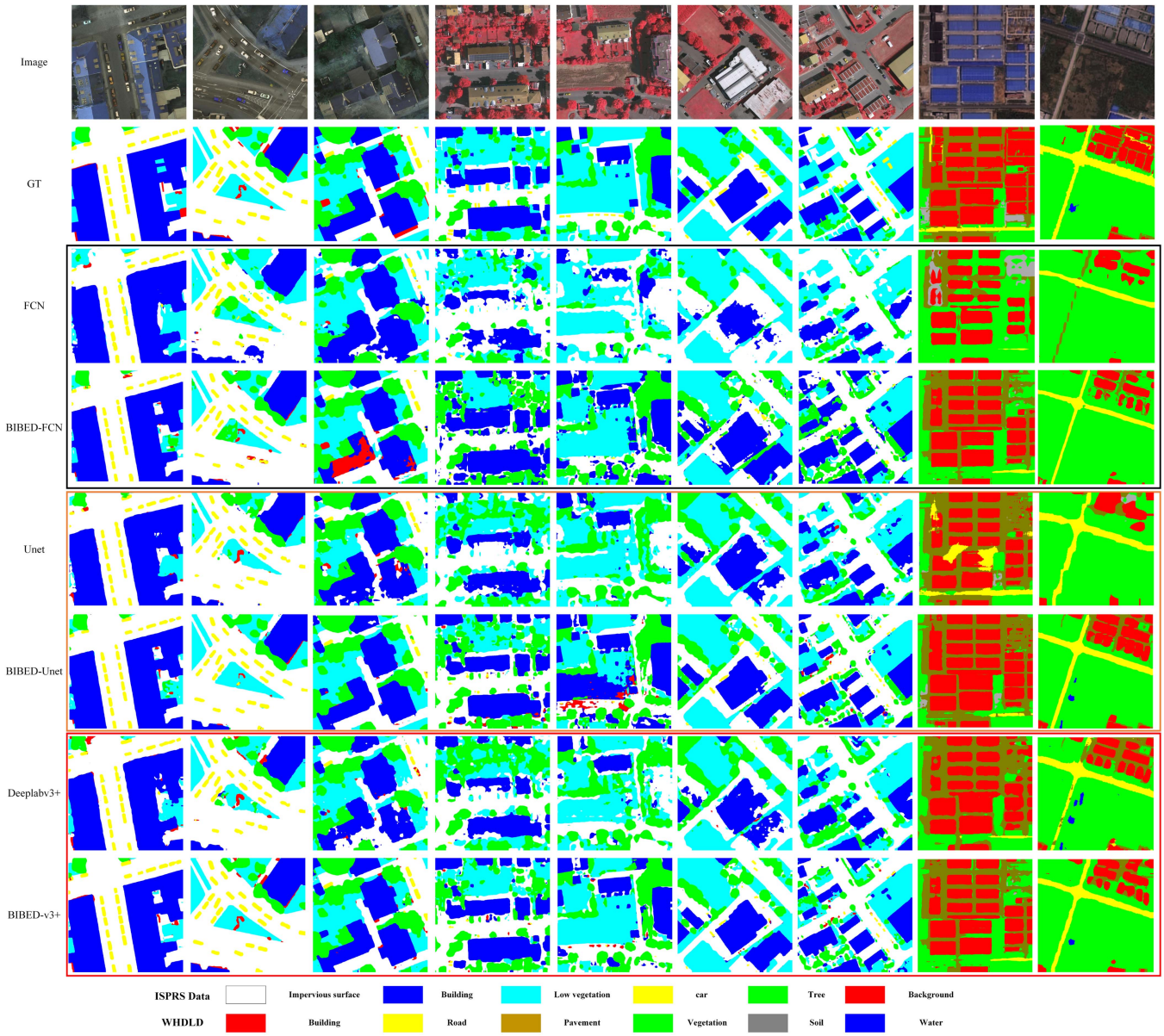

Fig. 14. Visualizations of semantic segmentation results output by different BIBED-Seg models with varying networks of segmentation. The top three columns are selected from the Potsdam data set. Columns 3 to 5 are selected from the Vaihingen data set. The last two columns are selected from the WHDLD dataset.

among all methods, with OA accuracy of 90.2%, 87.7%, and 81.5% for the two datasets, respectively.

Second, to explore the effect of different boundary features on the semantic segmentation results, we conducted the following comparison experiments based on the Potsdam data set, which used Deeplabv3+ as the backbone: Canny-Seg, HED-Seg, BDCN-Seg, and BIBED-Seg (ours). As shown in Fig. 15, compared with the traditional Canny algorithm and the recent HED, BDCN, the BIBED model proposed by this article performs best in improving the semantic segmentation effect of high-resolution remote sensing. In short, the semantic segmentation results based on the BIBED edge detection method proposed in this article are closer to the ground truth map.

The quantitative evaluation of the above segmentation results is shown in Table XI. Semantic segmentation results by BIBED-Seg obtain the OA of 90.2%, the Kappa of 89.8%, and the MIOU of 76.0%. 3.8%, 5.0%, and 5.3% higher than results by HED-Seg; 1.7%, 2.5%, and 2.4% higher than BDCN-Seg. The semantic segmentation result combining edge information obtained by the Canny algorithm is the worst. OA is only 65.6%. It shows that the edge information detected by the Canny algorithm is mixed and disorderly, which is useless for the semantic segmentation task of high-resolution remote sensing images, which will only interfere with the model's training and reduce the segmentation accuracy.

Finally, we also explored the enhancement effects of each category. Table XII and Fig. 16 show the IOU and IOU improvement of the segmentation results of each category in the Potsdam, Vaihingen, and WHDLD data sets. Compared with the semantic segmentation results without edge features, the IOU of each class

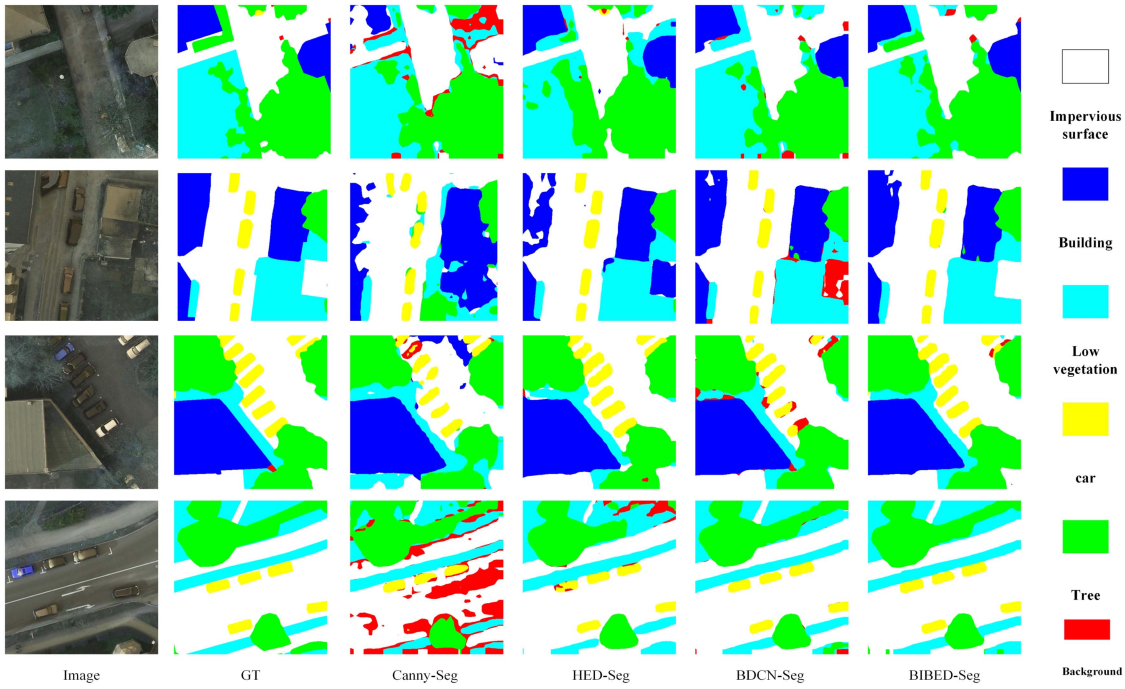

Fig. 15. Semantic segmentation results by BIBED-v3+ with different edge detection methods (including comparative experiments: The semantic segmentation results with traditional Canny, the semantic segmentation results with HED, the semantic segmentation results with BDCN, and with BIBED).

TABLE XII
IOU IMPROVEMENT OF SEMANTIC SEGMENTATION RESULTS BY BIBED-V3+ COMPARED WITH DEEPLABV3+

	Methods	Surface	Building	Low vegetation	Tree	Car
Potsdam	Deeplabv3+	81.5	83.0	66.2	70.8	64.1
	BIBED-v3+	**87.4**	**90.4**	**75.3**	**79.2**	**67.3**
Vaihingen	Deeplabv3+	65.2	70.7	55.1	59.1	57.7
	BIBED-v3+	**70.6**	**76.1**	**63.0**	**67.8**	**62.3**

	Methods	Building	Road	Pavement	Vegetation	Soil	Water
WHDLD	Deeplabv3+	45.7	44.9	27.9	73.2	33.2	86.2
	BIBED-v3+	**56.5**	**57.6**	**40.4**	**79.3**	**42.1**	**91.9**

TABLE XIII
VALIDITY OF RESULTS IN DIFFERENT EDGE-AWARE SEMANTIC SEGMENTATION METHODS

Methods	Dataset	OA	Kappa	MIOU	Dataset	OA	Kappa	MIOU
FCN-CRF [3]		84.7	83.1	70.2		77.8	75.6	59.7
BEA-SegNet [50]		88.2	87.1	73.8		85.7	79.3	66.1
EANet [51]	**Potsdam**	89.1	88.6	74.7	**Vaihigen**	86.5	80.5	67.7
FusionNet [52]		88.8	87.5	74.3		86.8	80.2	67.3
GMENet [48]		89.7	88.3	74.7		86.1	79.6	68.1
BIBED-Seg(v3+)		**90.2**	**89.8**	**76.0**		**87.7**	**83.7**	**69.6**

of the semantic segmentation results guided by BIBED has been significantly improved, among which the promotion effect of low vegetation and trees is the most significant, with an increase of 9.1% and 8.4% in Potsdam data set, and increased by 7.9% and 8.7% in Vaihingen data set. For the WHDLD dataset, the improvement in IOU was more pronounced for all categories, with the six categories improving by 10.8%, 12.7%, 12.5%, 6.1%, 8.9%, and 5.7%, respectively.

Finally, to further demonstrate the advancedness of the proposed method in this article, we conducted a comparison experiment between BIBED-Seg and several advanced edge-aware deep learning methods on two remote sensing datasets, as shown in Table XIII. All methods were done on the same training and test sets, and it was found that the BIBED-Seg model proposed in this article achieved the best accuracy evaluation results, with the highest scores in both overall pixel accuracy (OA) and intersection over Union (IOU). Recent methods that focus on remote sensing segmentation with joint training of multiple features (such as EANet, FusionNet, and GMENet) also have good results. And the postprocessing edge enhancement methods based on CRF have been slightly behind.

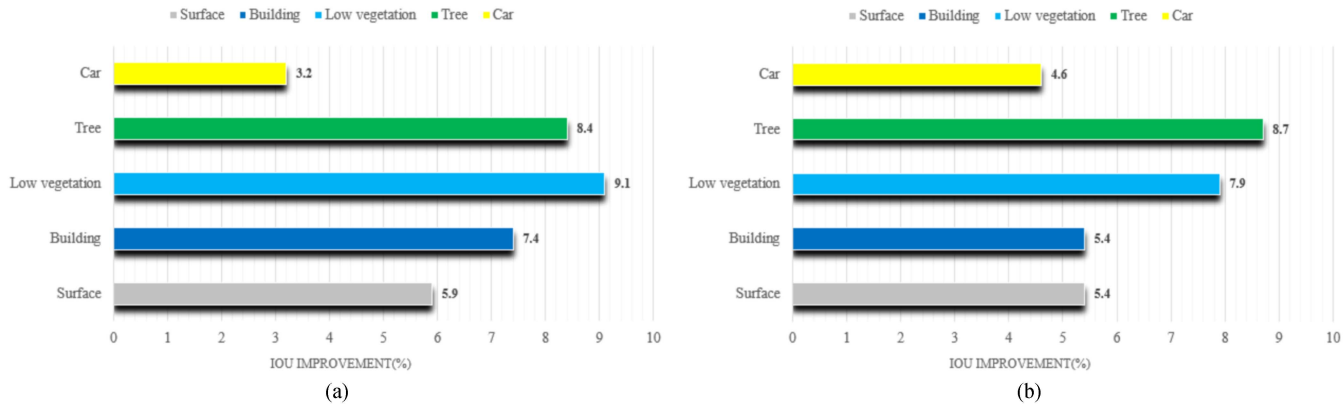

Fig. 16. IOU improvement of semantic segmentation results in different objects by BIBED-Seg (compared with results without edge). (a) From the Potsdam data set. (b) From the Vaihingen data set.

V. Discussion

The semantic segmentation based on boundary information proposed in this article can lead to more accurate and stable results obtained by advanced semantic segmentation networks, which are mainly influenced by the two leading networks in the model, the edge detection network and the semantic segmentation network. The better the learning and expression ability of the two networks, then the better the overall segmentation result. Through experiments, we can find that combining different edge detection networks and semantic segmentation networks yields very different semantic segmentation results. In the case where the latter is already mature, this is the reason why we have devoted ourselves to the study of improving the edge detection capability.

The BIBED edge detection network proposed in this article is mainly built with the residual structure due to the powerful learning ability of residual networks in recent years [43], [44]. In addition, the improvement of the BIBED network boundary detection capability is also attributed to the applicability of the CSAM on high-resolution remote sensing images and the balance of positive and negative samples in unbalanced binary classification and the mining of complex samples through the loss function.

However, edge detection, especially the detection of interclass boundaries, remains a significant challenge, and our method, although greatly improved, is still inadequate in terms of boundary continuity and adhesion. In the future, we will continue our research in boundary detection and improve these problems one by one.

VI. Conclusion

This article proposes BIBED, a block-in-block convolution residual block-based edge detection method for high-resolution remote sensing images. To reduce the amount of calculation and consider the edge features at multiple scales, we built a block-in-block residual network structure and designed a multiscale feature weight fusion loss function. To improve high-resolution remote sensing images' boundary positioning and detection accuracy, we introduce the channel and spatial attention modules into the residual structure to focus on the images' band and spatial dimension. Our method compares favorably with over five edge detection methods on ISPRS Potsdam and Vaihingen data sets and has shown the best performance.

Based on this, we propose the BIBED-Seg model, a two-stage semantic segmentation model in which edges are extracted first and then segmented. Various semantic segmentation networks are compared on three high-resolution aerial remote sensing data sets. It has been discovered that the method may optimize the bounds of segmentation results and enhance segmentation accuracy greatly, particularly in the intersection over Union.

This also supports our hypothesis that using precise and effective boundary information to drive remote sensing image segmentation will make it easier. Of course, this presents more severe issues for edge detection methods and techniques, and there is still a long way to go.

References

[1] L. Zhou, X. Kong, C. Gong, F. Zhang, and X. Zhang, "FC-RCCN: Fully convolutional residual continuous CRF network for semantic segmentation," *Pattern Recognit. Lett.*, vol. 130, pp. 54–63, 2020, doi: 10.1016/j.patrec.2018.08.030.

[2] Y. Li, Y. Ma, W. Cai, Z. Xie, and T. Zhao, "Complementary convolution residual networks for semantic segmentation in street scenes with deep Gaussian CRF," *J. Adv. Comput. Intell. Intell. Inform.*, vol. 25, no. 1, pp. 3–21, 2021, doi: 10.20965/JACIII.2021.P0003.

[3] X. Pan, T. Jiang, Z. Zhang, B. Sui, C. Liu, and L. Zhang, "A new method for extracting laver culture carriers based on inaccurate supervised classification with FCN-CRF," *J. Mar. Sci. Eng.*, vol. 8, no. 4, pp. 1–16, 2020, doi: 10.3390/JMSE8040274.

[4] J. Guo, L. Xu, J. Ding, B. He, S. Dai, and F. Liu, "A deep supervised edge optimization algorithm for salt body segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 10, pp. 1746–1750, Oct. 2021, doi: 10.1109/LGRS.2020.3007258.

[5] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011, doi: 10.1109/TPAMI.2010.161.

[6] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440, doi: 10.1109/CVPR.2015.7298965.

[7] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid, "Groups of adjacent contour segments for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 1, pp. 30–51, Jan. 2008, doi: 10.1109/TPAMI.2007.1144.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587, doi: 10.1109/CVPR.2014.81.

[9] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986, doi: 10.1109/TPAMI.1986.4767851.

[10] J. Kittler, "On the accuracy of the sobel edge detector," *Image Vis. Comput.*, vol. 1, no. 1, pp. 37–42, 1983, doi: 10.1016/0262-8856(83)90006-9.

[11] J. J. Lim, C. L. Zitnick, and P. Dollar, "Sketch tokens: A learned mid-level representation for contour and object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3158–3165, doi: 10.1109/CVPR.2013.406.

[12] P. Dollár and C. L. Zitnick, "Fast edge detection using structured forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1558–1570, Aug. 2015, doi: 10.1109/TPAMI.2014.2377715.

[13] G. Bertasius, J. Shi, and L. Torresani, "DeepEdge: A multi-scale bifurcated deep network for top-down contour detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4380–4389, doi: 10.1109/CVPR.2015.7299067.

[14] Y. Liu et al., "Richer convolutional features for edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 5872–5881, 2019, doi: 10.1109/TPAMI.2018.2878849.

[15] I. Kokkinos, "Pushing the boundaries of boundary detection using deep learning," in *Proc. Int. Conf. Learn. Representations*, 2016.

[16] Y. Wang, X. Zhao, Y. Li, and K. Huang, "Deep crisp boundaries: From boundaries to higher-level tasks," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1285–1298, Mar. 2019, doi: 10.1109/TIP.2018.2874279.

[17] A. Kaur, H. Singh, and D. Arora, "An efficient approach for image denoising based on edge-aware bilateral filter," in *Proc. 4th IEEE Int. Conf. Signal Process., Comput. Control*, 2017, pp. 56–61, doi: 10.1109/ISPCC.2017.8269649.

[18] D. Adlakha, D. Adlakha, and R. Tanwar, "Analytical comparison between sobel and prewitt edge detection techniques," *Int. J. Sci. Eng. Res.*, vol. 7, no. 1, pp. 1482–1485, 2016.

[19] C.-C. Zhang and J.-D. Fang, "Edge detection based on improved sobel operator," in *Proc. Int. Conf. Comput. Eng. Inf. Syst.*, 2016, pp. 129–132, doi: 10.2991/ceis-16.2016.25.

[20] J. Hu, X. Tong, Q. Xie, and L. Li, "An improved, feature-centric LoG approach for edge detection," in *Lecture Notes in Computer Science (Including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9787. Berlin, Germany: Springer, 2016, pp. 474–483, doi: 10.1007/978-3-319-42108-7_36.

[21] R. A. Purba, J. Sembiring, E. H. Sihombing, and S. Sondang, "Edge image detection combines Laplace operation with convolution technique to produce drawing materials for children," *J. Phys., Conf. Ser.*, vol. 1402, no. 6, 2019, Art. no. 066097, doi: 10.1088/1742-6596/1402/6/066097.

[22] C. L. Dembia, Y. C. Liu, and C. T. Avedisian, "Automated data analysis for consecutive images from droplet combustion experiments," *Image Anal. Stereology*, vol. 31, no. 3, pp. 137–148, 2012, doi: 10.5566/ias.v31.p137-148.

[23] P. S. P. Wang and J. Yang, "A review of wavelet-based edge detection methods," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 26, no. 7, 2012, Art. no. 1255011, doi: 10.1142/S0218001412550117.

[24] S. S. Sengar and S. Mukhopadhyay, "Moving object area detection using normalized self adaptive optical flow," *Optik (Stuttg)*, vol. 127, no. 16, pp. 6258–6267, 2016, doi: 10.1016/j.ijleo.2016.03.061.

[25] C. B. Sherin and L. Mredhula, "A novel method for edge detection in images based on particle swarm optimization," *J. Phys., Conf. Ser.*, vol. 787, no. 1, 2017, Art. no. 012012, doi: 10.1088/1742-6596/787/1/012012.

[26] R. Song, Z. Zhang, and H. Liu, "Edge connection based Canny edge detection algorithm," *Pattern Recognit. Image Anal.*, vol. 27, no. 4, pp. 740–747, 2017, doi: 10.1134/S1054661817040162.

[27] M. Han, X. Yang, and E. Jiang, "An extreme learning machine based on cellular automata of edge detection for remote sensing images," *Neurocomputing*, vol. 198, pp. 27–34, 2016, doi: 10.1016/j.neucom.2015.08.121.

[28] D. Liu, H. Wang, S. Wang, G. Pu, X. Deng, and X. Hou, "Quaternion-based improved artificial bee colony algorithm for color remote sensing image edge detection," *Math. Problems Eng.*, vol. 2015, pp. 1–10, 2015, doi: 10.1155/2015/138930.

[29] S. Guiming and S. Jidong, "Remote sensing image edge-detection based on improved Canny operator," in *Proc. 8th IEEE Int. Conf. Commun. Softw. Netw.*, 2016, pp. 652–656, doi: 10.1109/ICCSN.2016.7586604.

[30] M. Vijayan, R. Mohan, and P. Raguraman, "Contextual background modeling using deep convolutional neural network," *Multimedia Tools Appl.*, vol. 79, no. 15/16, pp. 11083–11105, 2020, doi: 10.1007/s11042-019-07800-0.

[31] Y. Tian, "Artificial intelligence image recognition method based on convolutional neural network algorithm," *IEEE Access*, vol. 8, pp. 125731–125744, 2020, doi: 10.1109/ACCESS.2020.3006097.

[32] Y. Ganin and V. Lempitsky, "N4-fields: Neural network nearest neighbor fields for image transforms," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9004. Berlin, Germany: Springer, 2015, pp. 536–551, doi: 10.1007/978-3-319-16808-1_36.

[33] W. Shen, X. Wang, Y. Wang, X. Bai, and Z. Zhang, "DeepContour: A deep convolutional feature learned by positive-sharing loss for contour detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3982–3991, doi: 10.1109/CVPR.2015.7299024.

[34] S. Xie and Z. Tu, "Holistically-nested edge detection," *Int. J. Comput. Vis.*, vol. 125, no. 1–3, pp. 3–18, 2017, doi: 10.1007/s11263-017-1004-z.

[35] L. Chen et al., "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6298–6306, doi: 10.1109/CVPR.2017.667.

[36] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11211. Heidelberg, Germany: LNCS, 2018, doi: 10.1007/978-3-030-01234-2_1.

[37] Y. Hu, H. Tang, and G. Pan, "Spiking deep residual networks," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Nov. 1, 2021, doi: 10.1109/TNNLS.2021.3119238.

[38] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020, doi: 10.1109/TPAMI.2018.2858826.

[39] J. He, S. Zhang, M. Yang, Y. Shan, and T. Huang, "Bi-directional cascade network for perceptual edge detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3823–3832, doi: 10.1109/CVPR.2019.00395.

[40] W. Zhao, H. Zhang, Y. Yan, Y. Fu, and H. Wang, "A semantic segmentation algorithm using FCN with combination of BSLIC," *Appl. Sci.*, vol. 8, no. 4, pp. 500–515, 2018, doi: 10.3390/app8040500.

[41] K. Heidler, L. Mou, C. Baumhoer, A. Dietz, and X. X. Zhu, "HED-UNet: Combined segmentation and edge detection for monitoring the antarctic coastline," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 23, 2021, Art. no. 4300514, doi: 10.1109/TGRS.2021.3064606.

[42] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11211. Heidelberg Germany: LNCS, 2018, doi: 10.1007/978-3-030-01234-2_49.

[43] F. Ren, W. Liu, and G. Wu, "Feature reuse residual networks for insect pest recognition," *IEEE Access*, vol. 7, pp. 122758–122768, 2019, doi: 10.1109/ACCESS.2019.2938194.

[44] K. Chen, K. Chen, Q. Wang, Z. He, J. Hu, and J. He, "Short-term load forecasting with deep residual networks," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3943–3952, Jul. 2019, doi: 10.1109/TSG.2018.2844307.

[45] W. Shen, X. Wang, Y. Wang, X. Bai, and Z. Zhang, "DeepContour: A deep convolutional feature learned by positive-sharing loss for contour detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3982–3991.

[46] Y. Liu, M. Cheng, X. Hu, K. Wang, and X. Bai, "Richer convolutional features for edge detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5872–5881.

[47] U. Michieli and P. Zanuttigh, "Edge-aware graph matching network for part-based semantic segmentation," *Int. J. Comput. Vis.*, vol. 130, pp. 2797–2821, 2022.

[48] R. Chen, F. Zhang, and T. Rhee, "Edge-aware convolution for RGB-D image segmentation," in *Proc. 35th Int. Conf. Image Vis. Comput. New Zealand*, 2020, pp. 1–6.

[49] H. Kuang, Y. Liang, N. Liu, J. Liu, and J. Wang, "BEA-SegNet: Body and edge aware network for medical image segmentation," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2022, pp. 939–944.

[50] K. Yuan, G. Meng, D. Cheng, J. Bai, S. Xiang, and C. Pan, "Efficientcloud detection in remote sensing images using edge-aware segmentation network and easy-to-hard training strategy," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 61–65.
[51] D. Cheng, G. Meng, S. Xiang, and C. Pan, "FusionNet: Edge aware deep convolutional networks for semantic segmentation of remote sensing harbor images," *IEEE J-Stars*, vol. 10, no. 12, pp. 5769–5783, Dec. 2017.
[52] Z. Su et al., "Pixel difference networks for efficient edge detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5097–5107.
[53] C. Y. Lee, J. G. Shon, and J. S. Park, "An edge detection–based eGAN model for connectivity in ambient intelligence environments," *J. Ambient Intell. Humanized Comput.*, vol. 13, pp. 4591–4600, 2021.
[54] X. Wang, R. B. Girshick, A. K. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.
[55] J. Fu, J. Liu, H. Tian, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3141–3149.
[56] Z. Huang et al., "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 603–612.

Baikai Sui received the B.S. degree in remote sensing science and technology and the M.S. degree in photogrammetry and remote sensing from the Department of Remote Sensing Science and Technology, Shandong University of Science and Technology, Qingdao, China, in 2018 and 2021, respectively. He is currently working toward the Ph.D. degree in surveying and mapping science and technology with Southwest Jiaotong University, Chengdu, China.

His research interests include the area of digital image processing and deep learning, including convolutional neural network classification, semantic segmentation, and image generation based on adversarial neural network.

Yungang Cao (Member, IEEE) received the M.S. degree in geodesy and survey engineering from Southwest Jiaotong University, Chengdu, China, in 2003, and the Ph.D. degree in cartography and geographic information systems from the Chinese Academy of Sciences, Beijing, China, in 2006.

He is currently an Associate Professor with the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University. His research interests include computer vision, deep learning, and remote sensing image processing.

Xueqin Bai received the B.S. degree in surveying and mapping science and technology from the College of Mining Engineering, Taiyuan University of Technology, Taiyuan, China, in 2020. She is currently working toward the Postgraduate degree in geographic information science with Southwest Jiaotong University, Chengdu, China.

Her research interests include the area of remote sensing information extraction and image processing, including convolutional neural network classification and image generation based on active and passive remote sensing images.

Shuang Zhang received the B.S. degree in surveying and mapping engineering from the Department of Surveying and Mapping Engineering, Xinjiang University, Urumqi, China, in 2021. He is currently working toward the M.S. degree in civil and hydraulic engineering with the Southwest Jiaotong University, Chengdu, China.

His research interests include the area of digital image processing and deep learning.

Renzhe Wu received the B.S. degree in surveying and mapping engineering from the Department of Surveying and Mapping Engineering, Henan University of Urban Construction, Pingdingshan, China, in 2018, and the M.S. degree in surveying and mapping engineering from the Department of Surveying and Mapping Engineering, Southwest Jiaotong University, Chengdu, China, in 2021, where he is currently working toward the Ph.D. degree in surveying and mapping science and technology.

His research interests include area of synthetic aperture radar images and optical remote sensing image processing