Edge-Enhanced GCIFFNet: A Multiclass Semantic Segmentation Network Based on Edge Enhancement and Multiscale Attention Mechanism

Long Chen, Zhiyuan Qu, Yao Zhang, Jingyang Liu, Ruwen Wang, and Dezheng Zhang

***Abstract*—In recent years, remote sensing images (RSIs) have witnessed significant improvements in both quality and quantity. With the application of deep-learning techniques, these RSIs can be more effectively utilized to harnessed to aid in environment monitoring and urban planning. Semantic segmentation, as a common task in RSIs processing, confronts numerous challenges, including inaccurate classification, fuzzy boundaries, and other problems. This article proposes a novel semantic segmentation network known as the edge-enhanced global contextual information guided feature fusion network to address these challenges. This network consists of an edge-enhanced part and a backbone network part. First, in the encoding stage, the recurrent criss-cross attention block is employed, which incorporates spatial attention, mechanisms to capture global information. Second, in the decoding stage, a channel attention residual block module is proposed to facilitate the fusion of high-level and low-level features. Moreover, we enhance the network's ability to extract edge information during training by sharing parameters between the backbone and employing a specialized loss function. The network proposed in this article utilizes both channel attention and spatial attention at different stages, effectively utilizing edge information. Finally, we conduct experiments using the Yinchuan dataset and the LoveDA dataset. The experimental results show that the proposed network demonstrates excellent performance on both datasets.**

***Index Terms*—Attention, convolutional neural network (CNN), edge segmentation, remote sensing images (RSIs), semantic segmentation.**

I. Introduction

In RECENT years, remote sensing satellite technology has developed rapidly, and the quality and quantity of remote sensing images (RSIs) have been greatly improved, which makes the RSIs used in engineering practice have a reasonable data basis. RSIs contain a large amount of ground information, which is an important data source for guiding land resources and urban planning. With the help of RSIs, the utilization of water resources, land resources, forestry and animal husbandry resources, and mineral resources can be monitored and planned; they can also be used to monitor natural disasters, such as earthquakes and fires, and they can also be used for navigation, urban road traffic planning, and urban and rural planning [1], [2].

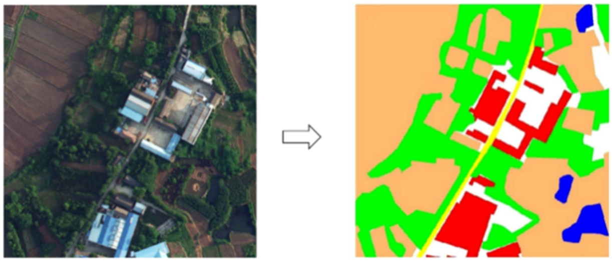


Fig. 1. Schematic representation of remote sensing images semantic segmentation task.

RSIs contain wealthy information, and their effective utilization demands precise and thorough analysis. Early researchers relied on manual annotation to acquire accurate land category information, but this approach incurred significant costs. However, with the progressive adoption of convolutional neural networks (CNNs) in image processing tasks, the exploration of deep learning in RSIs has gained considerable momentum. This includes research areas, such as change detection [3], land cover classification [4], and object detection [5].

The land cover classification tasks involve dividing an entire image into different regions, such as buildings, roads, water bodies, vegetation, etc. This task can be achieved through semantic segmentation methods [6] that classify individual pixels. Fig. 1 illustrates a typical RSI semantic segmentation task [7].

Semantic segmentation tasks in RSIs come with several challenges, such as discontinuity prediction for larger objects during the segmentation. RSIs contain rich texture information, which can lead to confusion at the edges where multiple objects converge. In addition, issues like imbalanced samples in RSIs datasets can result in suboptimal training outcomes. National territory spatial planning is a guide for national spatial development, a spatial blueprint for sustainable development, and the basic basis for all types of development activities. In national territory spatial planning, workers need to accurately classify land use types, and now the time-consuming and laborious manual

Manuscript received 2 November 2023; revised 11 January 2024; accepted 18 January 2024. Date of publication 23 January 2024; date of current version 12 February 2024. This work was supported by the Key Research and Development Program of Ningxia Hui Autonomous Region under Grant 2019BFG02009. *(Corresponding author: Dezheng Zhang.)*

Long Chen, Zhiyuan Qu, Jingyang Liu, Ruwen Wang, and Dezheng Zhang are with the Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing 100083, China (e-mail: workcarecl@163.com; tripensqzy@gmail.com; s20200660@xs.ustb.edu.cn; g202008834@xs.ustb.edu.cn; zdzchina@ustb.edu.cn).

Yao Zhang is with the Department of Surgery, University of Alberta, Edmonton, AB T6G 2R3, Canada (e-mail: yao.zhang@ualberta.ca).

Digital Object Identifier 10.1109/JSTARS.2024.3357540

labeling method is usually used. Without solving the problems mentioned above, the achievements of semantic segmentation of RSIs cannot be applied to practical work and the efficiency of the staff would not be improved. Aiming at these problems and the actual characteristics of the research data in this article, we carry out the research on deep-learning-based semantic segmentation of RSIs with multiple land classes, and the main contributions are as follows.

1) A network named global contextual information guided feature fusion (GCIFFNet) is proposed based on an encoder–decoder structure including the recurrent criss-cross attention (RCCA) module in the encoding stage, utilizing the spatial attention mechanism to obtain global information, and completes the fusion of high- and low-level features in the decoding stage. It also comes with improved loss functions during the training process.
2) To solve the problem of discontinuous prediction, the Channel Attention Residual Block (CARB) module is proposed to use the channel attention (CA) mechanism and to fuse high-level and low-level features.
3) To improve the edge accuracy of the segmentation model, an edge-enhanced segmentation model is designed to utilize the semantic edge information of RSIs. The improved model with this subnetwork is named edge-enhanced GCIFFNet and realizes the accurate semantic segmentation of RSIs.

II. Related Works

A. Semantic Segmentation

Hinton proposed AlexNet [8] to lay the foundation for the application of deep learning in image processing. In recent years, networks, such as FCN [9], VGGNet [10], ResNet [11], and MobileNet [12], have appeared. However, these networks, characterized by their simple structures and limited parameters, struggle to capture sufficient global information. The U-Net based on an encoder–decoder structure [13] was developed to address this issue. The DeepLab series, from DeepLabV1 [14] to DeepLabV3+ [15], tackles the challenge of easily losing spatial information in semantic segmentation. Although DeepLabV3+ enhances model accuracy, its method of segmenting images into multiple chunks for processing during semantic segmentation, making the segmentation of irregularly shaped objects ineffective.

With the development of deep learning, attention mechanism is also introduced into image processing [16]. Initially proposed by Fukushima, the attention mechanism enhances the ability of the model to segment different parts of an image [17]. This approach is exemplified by networks, such as SE-Net [18], PSANet [19], CBAM [20], and DANet [21]. While the attention mechanism significantly improves semantic segmentation, it faces challenges in handling small targets in RSIs, which leads to imprecise segmentation of small objects. Subsequently, many excellent networks based on self-attention mechanisms have been proposed [22], such as GCNet [23] and ANN [24].

Vaswani et al. [17] at Google introduced the transformer model to address the drawback of slow training of RNN. The transformer model incorporates a self-attention mechanism to achieve efficient parallel processing. Dosovitskiy et al. [25] extended the transformer model into the image domain and proposed vision transformer, which achieved state-of-the-art (SOTA) results on large-scale datasets. Subsequently, many transformer-based semantic segmentation networks have been proposed, such as SETR [26], Segformer [27], and Segmenter [28], breaking the longstanding dominance of CNNs in the field of semantic segmentation.

B. Edge Information Enhanced Methods

In image processing, edges represent the boundaries where distinct object classes intersect, leading to pronounced variations. These edges can be detected through differentiation techniques [29]. Recently, there are more and more models using neural networks to extract edges, such as CASENet [30], DeepEdge [31], and CEDN [32]. However, integrating edge segmentation with the semantic segmentation tasks necessitates designing and independently training two distinct networks, which is a complicated and tedious process.

Alternatively, semantic segmentation can also be executed through multitask training. This method involves simultaneous processing of semantic segmentation and edge detection, allowing for mutual enhancement of the results [33], [34]. A potential drawback of this approach is that it may have a negative impact during training.

In the task of object detection and semantic segmentation within RSIs, edges are always confusing. He et al. [35] proposed uncertainty-aware network, which achieves SOTA performance in the three public datasets. Based on the encoder–decoder framework, Hang et al. [36] proposed a CNN-based model to identify oceanic eddies. This network comprises an eddy identification branch and an edge extraction branch. The latter is used to learn the edge information of eddies to enhance the recognition effect of the network.

C. Remote Sensing Semantic Segmentation

Contemporary research predominantly employs semantic segmentation algorithms for feature extraction from RSIs. Badrinarayanan proposed SegNet [37], a notable example in this domain. Yang et al. [38] performed SegNet for extracting construction land in rural areas and observed its exceptional overall performance. However, some detailed information present in RSIs, such as building edges and fine roads, which may be ignored or misclassified by the SegNet model. To address these limitations, Liu et al. [39] proposed improving DeepLabV3+ for RSIs by introducing a dual-attention mechanism module and designing a network model with two different connections. Hang et al. [40] developed a multiscale progressive segmentation network to solve the issue of accurate semantic segmentation for both large and small objects in high-resolution remote sensed imagery, and validated the effectiveness of the method on the Vaihingen and Potsdam datasets. Zhou et al. [41] designed a feature decoupling module and proposed a class-guided feature decoupling network by exploiting the co-occurrence relationships between different classes of objects in the scene, and

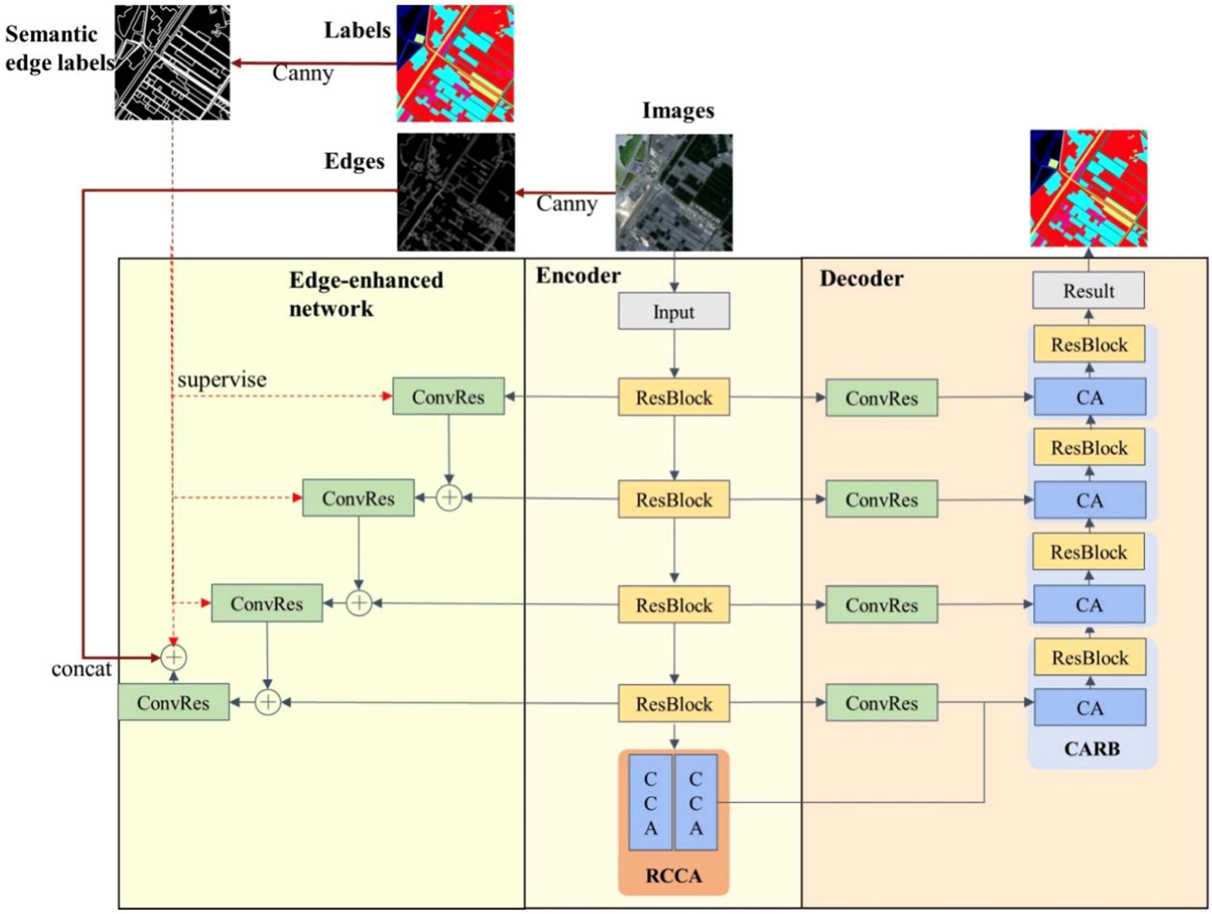


Fig. 2. Overall structure of our proposed edge-enhanced GCIFFNet. It is composed of three parts: encoder, decoder, and an edge-enhanced network. The edge-enhanced network helps learning features of the edges by sharing parameters of backbone network. RCCA and CARB modules are introduced in the encoding and decoding stages, respectively.

the overall accuracy is greater than 90% in the extraction of high-resolution RSIs.

III. Methods

In this article, we propose a semantic segmentation network based on an encoder–decoder structure named GCIFFNet. In addition, to augment the utilization of edge information, we proposed an auxiliary edge enhanced. Combined with this subnetwork, we name the final model edge-enhanced GCIFFNet. The network mainly consists of three parts: the encoder, the decoder, and the edge-enhanced network. The comprehensive network structure is delineated in Fig. 2.

A. Network Architecture

In this article, we employ the ResNet 50 architecture as the backbone network to obtain a feature map with high-level semantic information. At the end of the encoding stages, we incorporate the cross-correlation attention module (CCA), proposed by CCNet [42], which extracts dense contextual information through two sequentially connected CCA modules, known as RCCA.

In addition, we also adopted the UNet skip-connection structure. In the upsampling process, low-level features are refined through an improved residual module named ConvRes. Then, these features are fed to the CARB, enabling their fusion with the high-level features. The fused features contain not only discriminative semantic information but also detailed spatial information. The CARB module consists of a CA module and a residual module. By using the RCCA module and CARB module, our proposed network can utilize both spatial attention and CA with low computational effort.

We also introduce an edge-enhanced network to alleviate blurry boundaries and utilize edge information. During the training process, this network serves as another output head supervised by semantic edge labels.

In the following sections, we will describe each module in detail.

B. RCCA Module

The attention mechanism derives from human habit and imbues neural networks by concentrating on pertinent information. This process is shown in (1). The elements Q, K, and V are calculated by multiplying the inputs with different matrices.

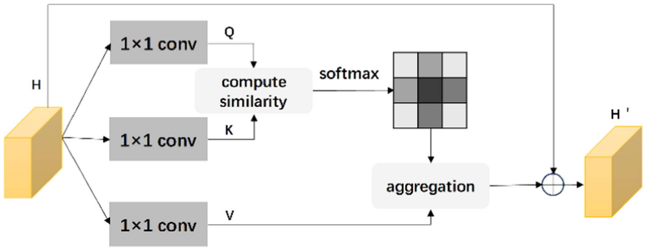


Fig. 3. Schematic diagram of CCA module.

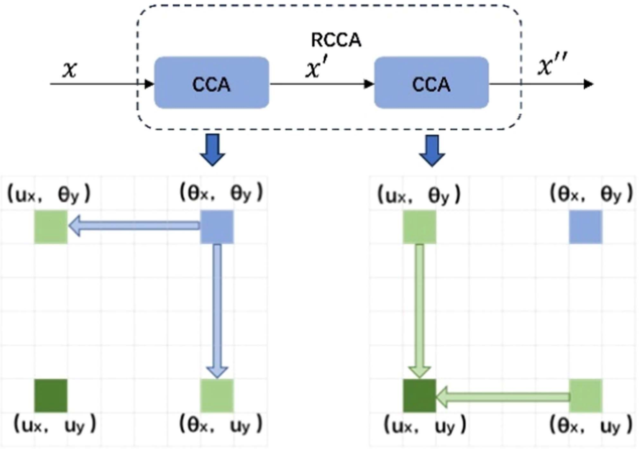


Fig. 4. Structure of RCCA module.

The similarity between Q and K is computed through the dot product, and then the result is normalized by softmax to get a matrix with values between 0 and 1, which establishes a dynamic weighting relationship. V represents the features after the linear transformation of the inputs, and a filtered feature F is obtained by multiplying a normalized matrix with a value between 0 and 1 with V

$$\text{Attention }(Q, K, V) = \text{ softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{1}$$

There is a certain level of correlation between various elements within the input, and the application of attention mechanisms among the internal elements of the input is known as the self-attention mechanism. In the field of computer vision, the input feature map of size $H \times W \times C$ is regarded as $H \times W$ vectors of length C. Self-attention lies in calculating the interrelationships between these $H \times W$ vectors, i.e., the similarity of the vectors at each position. Therefore, in the semantic segmentation task, the use of the self-attention mechanism can improve the phenomenon that only local features can be extracted in the fully CNN. The self-attention mechanism can be used to effectively capture contextual information across extended spatial ranges and is especially beneficial in addressing the problem of discontinuous prediction in the semantic segmentation of large size targets in RSIs.

Our network incorporates the CCA module, as shown in Fig. 3. It is essentially a form of self-attention mechanism but with a greatly reduced number of parameters compared to Non-Local [43] and DANet. In Fig. 3, H represents the local features with size $C \times H \times W$. Initially, two feature maps are obtained for the local features using two 1×1 convolution modules, and their sizes will be changed to $C' \times W \times H$. Subsequently, the similarity computation is used to compute the similarities between these feature maps.

The definition of similarity computation (affinity) in Fig. 3 is shown as

$$d_{i,u} = Q_u\ \Omega^{\mathrm{T}}_{i,u}. \tag{2}$$

In this equation, Q_u represents a vector obtained by extracting the features from C' channels at a specific position u in the feature maps. Similarly, we can extract features from the second feature map corresponding to positions in the same row and column as u, forming a $(H + W - 1) \times C'$ vector. $\Omega_{i,u}$represents the ith element of this vector. $\Omega(i, u)$ and Qu are dot-multiplied to compute the similarity to obtain $d_{i,u}$. By performing the above operations on each position of the feature map, the final feature map D is obtained, which has the size of $(H + W - 1) \times W \times H$. Finally, the softmax operation is applied across the channels to generate the attention feature map A.

The CCA module initiates its process by performing a 1×1 convolution of the feature map H to obtain the feature map V. Then, it extracts cross-feature maps from all the channels at a certain position on the feature map V to form a cross-feature vector with a spatial size of $(H + W - 1) \times C$. This feature vector is aggregated with the previously obtained feature map A according to (3), so that each position of the feature map V contains information about other positions in the same column with it. In this way, the CCA module can fuse features from different directions to better capture the contextual relationships of distinct locations in the image. Finally, feature map A and feature map V are weighted and aggregated with feature map H after weighted aggregation to obtain the final output feature vector of the CCA module

$$H'_u = \sum_{i\in|\Phi_u|} A_{i,u}\Phi_{i,u} + H_u. \tag{3}$$

To acquire dense contextual information, we utilize two CCA modules connected in series, as illustrated in Fig. 4. The first CCA module establishes a connection between a dark blue pixel and two light green pixels. Subsequently, the second CCA module connects these two light green pixels and a dark green pixel. In this way, the dark blue pixel point is brought into contact with the dark green pixel point, which is far away from it, thereby facilitating the capture of more extensive global contextual information.

Our proposed network contains RCCA module and CARB module, which allows the network to utilize both spatial attention and CA. We choose the RCCA module to capture long-range correlations instead of modules, such as pyramid squeeze attention (PSA) or coordinate attention. The reason is that both PSA and coordinate attention utilize the CA, and our CARB module also utilizes CA. Employing PSA or coordinate attention to capture long-range correlations leads to the redundant application of the squeeze-and-excitation mechanism in the decoding process. It

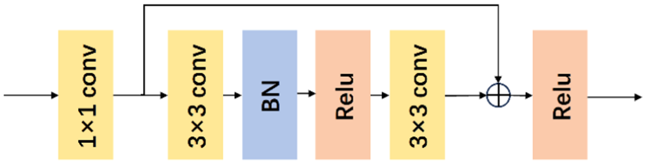


Fig. 5. Structure of ConvRes module.

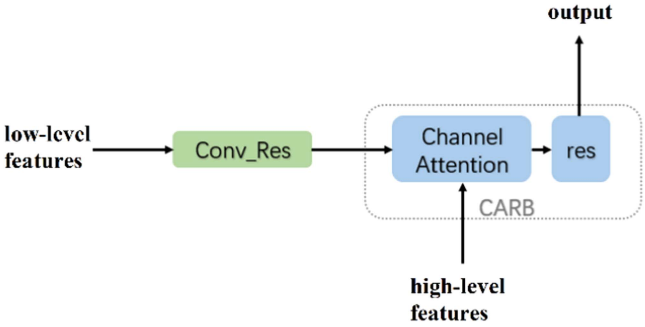


Fig. 6. Structure of CARB module used during decoding stage.

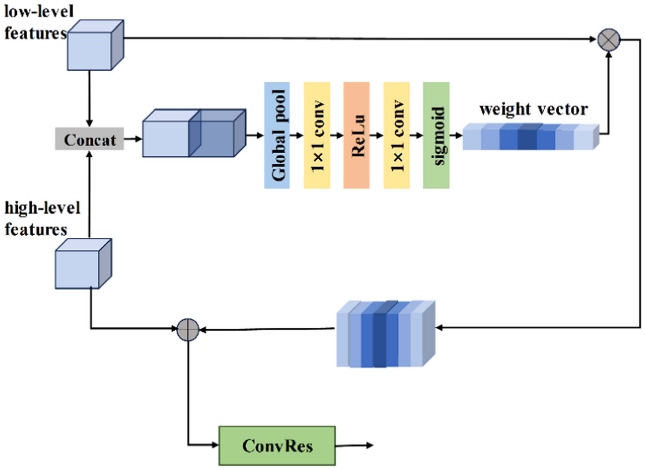


Fig. 7. Structure of CARB module.

makes the network pay more attention to the channel information and ignores the importance of spatial location information.

C. ConvRes Module

The role of the ConvRes module is to relearn the low-level features to refine the features. As shown in Fig. 5, ConvRes unifies the number of channels of the low-level features through a 1 × 1 convolution module. Subsequently, it refines the low-level feature maps again through the residual module. This module is similar to the basic residual module in ResNet, and primarily consists of two 3 × 3 convolutional modules. The ConvRes module not only standardizes the number of channels but also enables the neural network to learn important information in low-level features. In this article, ConvRes standardizes the number of channels in two cases. In the semantic segmentation network, the number of channels is unified to 512. In the edge-enhanced segmentation network, the number of channels is reduced to 21. This approach not only minimizes computational demands but also prevents gradient vanishing or gradient explosion.

D. CARB Module

RSIs contain many complex land classes, and there are often cases of cross-existence of land classes. In order to ensure intraclass consistency, neural networks need to extract global contextual information with discriminative properties.

Fig. 6 illustrates a single decoder module in this article. The decoder module applied in this article refers to UNet. However, instead of directly feeding the low-level features into the decoder via skip connections, they are initially routed through the ConvRes module, which refines the low-level features again before feeding them into the decoder. At the time of high- and low-level feature fusion, the high- and low-level features are inputted into the CARB for amalgamation. As a result, this module's final output features contain global semantic information and detailed features at different scales.

The CARB module is shown in Fig. 7, which can alleviate the intraclass inconsistency problem by utilizing high-level features to filter low-level features and retaining more discriminative features in the decoding process. This module refers to the squeeze-and-excitation module in SENet. Initially, the channel numbers of both high-level and low-level features are unified to 512. Following this, the high-level and low-level features are spliced to obtain a feature vector of size $H \times W \times 1024$. This vector is then subjected to global pooling and subsequently processed through a convolutional module, effectively reducing the channel count back to 512. The vector with a channel number of 512 is passed through the ReLu layer, the 1 × 1 convolutional layer, and a Sigmoid layer to obtain the final 1 × 1 × 512 vector. The final 1 × 1 × 512 vector, comprising weight coefficients, is instrumental in recalibration of the low-level features. It functions to suppress the channels with small discriminative properties in the low-level features, and strengthens the channels with more significant discriminative attributes, thereby facilitating effective feature channel selection. Subsequently, the recalibrated low-level features are combined with high-level features to generate the output of the CA module. The obtained output is then fed to the ConvRes module to get the refined fused features.

In extracting features from the backbone network, the high-level features help to acquire accurate categorical information, while low-level features contain clearer details but have a smaller receptive field, so the discriminative ability of the low-level features is poor. The two features are spliced together to adaptively learn the interchannel correlations in this new feature map to obtain a weight vector. This weight vector is used to filter the low-level features to highlight the feature channels that contain more category information. The decoders in this article use such a structure several times in the upsampling process to fuse the low-level features to obtain discriminative features stage by stage without losing the underlying detailed information. It is proved by subsequent experiments that the CARB module we proposed in this article can improve the performance of semantic segmentation.

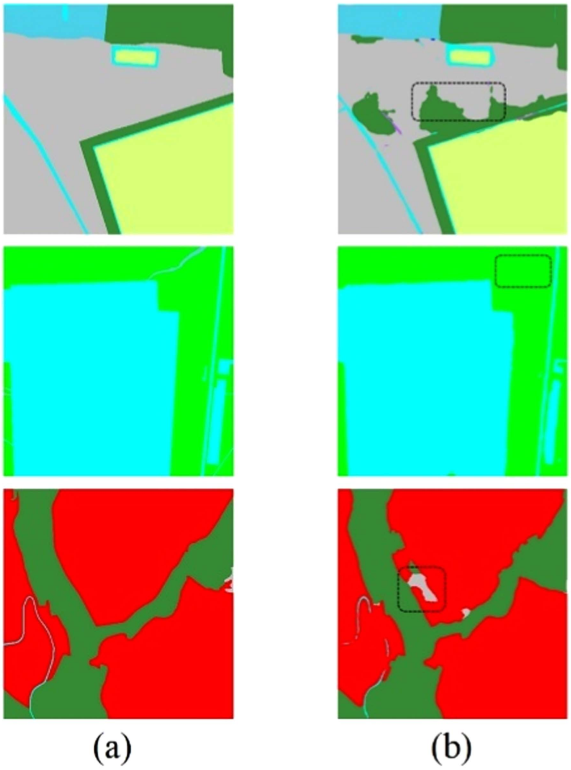


Fig. 8. Segmentation results of GCIFFNet, which show that the edge extraction effect still needs to be improved. (a) Ground truth. (b) GCIFFNet results.

E. Edge-Enhanced Network

Fig. 8 illustrates that GCIFFNet still needs to be improved from the perspective of edge extraction effect. To address the issue of "interclass feature similarity" in RSIs, it is essential for the network to pay enough attention to both the semantic edges and their adjacent pixels. In addition, the network must effectively learn the features of the semantic edges and the images on both sides of the edges. This approach is critical for significantly enhancing the precision of semantic segmentation to a large extent [44].

1) Selection of Edge Extraction Operators: The core point of semantic segmentation is to determine the boundaries between different classes of objects. It is important to note that semantic edges are different from image edges. Specifically, image edges will focus on the edges between categories as well as the edges within the same categories, while semantic edges will not focus on the edges between the same categories. For example, image edges can be extracted from two houses that are very close to each other, while semantic edges cannot be extracted from these two houses. To improve the accuracy of semantic segmentation, we need to utilize these two types of edge information. The image edges are extracted from the original images, and the semantic edges are obtained directly from the labeled images.

Our objective is that the edge information extracted from the original image can be more similar to the semantic boundary, which can reduce the network training difficulty. Fig. 9 displays the edge extraction results of different algorithms. This figure reveals that Sobel operator [45], Laplacian operator [46], and Prewitt [47] operators are more sensitive to noise. Consequently, the edges they extract are not clear, and they often identify nonsemantic boundaries as edges in the original image. The results extracted by the Roberts [48] operator omit a lot of edge information. In contrast, Canny [49] operator extracts better edge information due to the use of Gaussian filtering to remove the apparent noise and nonmaximum suppression to refine the edges. Therefore, in this article, the Canny operator is selected as the way of edge information extraction.

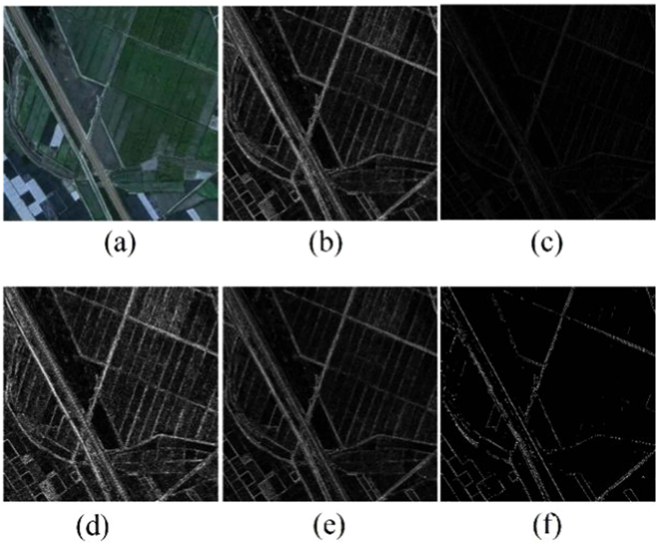


Fig. 9. Edge labels extracted by different operators. (a) Image. (b) Sobel. (c) Roberts. (d) Laplacian. (e) Prewitt. (f) Canny.

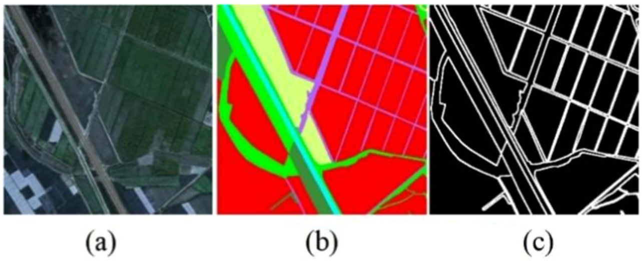


Fig. 10. Extraction of semantic edge labels. (a) Image. (b) Class label. (c) Semantic label.

We also extract the semantic edges of labeled images as supervisory signals to the edge-enhanced module. The accurate semantic boundaries need not be relabeled and can be extracted directly from the labels. The process of semantic boundary extraction is consistent with the above. Fig. 10 shows an image in the training set with its corresponding semantic segmentation label and semantic boundary label.

2) Edge-Enhanced Network Architecture: As is shown in Fig. 2, the edge-enhanced network is supervised by extracted edge labels and semantic edge labels during training. The edge-enhanced network helps learning features of the edges by sharing parameters of backbone network.

Specifically, the input image is extracted into four stages of features by ResNet50, and the feature maps of these four stages are input into the ConvRes module for refinement learning. Since the size of the feature maps at different stages is also different, the refined features need to be upsampled before they are added with the refined deep features at the next stage. In this article, bilinear interpolation is employed for upsampling, following which the resulting fused feature map is fed into the ConvRes module again. The network can learn more intricate edge information from the low-level features through this multilevel feature learning approach. At the same time, we use the semantic information of the higher level features to filter the edges of the nonsemantic boundaries and get accurate semantic boundaries. To mitigate the potential issue of insufficiently detailed semantic boundaries derived from deeper layers, we incorporate the Canny operator. This operator extracts the edge image from the original image, adding it as a new channel to the deepest fused features. Subsequently, the number of channels is unified by the Convres module. In this part, the ConvRes module is utilized to reduce the number of channels of the output features to 21. This reduces the computational demands and avoids gradient vanishing or gradient explosion.

It is worth noting that the auxiliary network is only involved in the training process and does not affect the speed of edge-enhanced GCIFFNet during testing.

F. Loss Functions

To alleviate the issue of class imbalance in the Yinchuan high-resolution remote sensing dataset, our approach involves a specifically designed loss function comprising two parts: weighted cross-entropy loss (WCE Loss) [50], [51] for GCIFFNet and focal loss [52] for edge-enhanced network. We will describe the two losses in detail as follows.

First, a WCE loss is used as shown as follows, where c represents the number of categories in the current semantic segmentation task, y_c represents the labeled values, and p_c represents the predicted value:

$$L = -\sum_{c=1}^{\text{class}} w_c y_c \log\left(p_c\right) \tag{4}$$

where the weight of each class w_c is calculated as described as follows:

Then, we traversed the entire dataset and calculate the total number of pixel points of the category in the dataset, n_{c_pixel}, furthermore, record the number of images in the dataset which the category occurs, recorded as n_c, and the frequency of occurrence of the category as n_{c_freq} as follows:

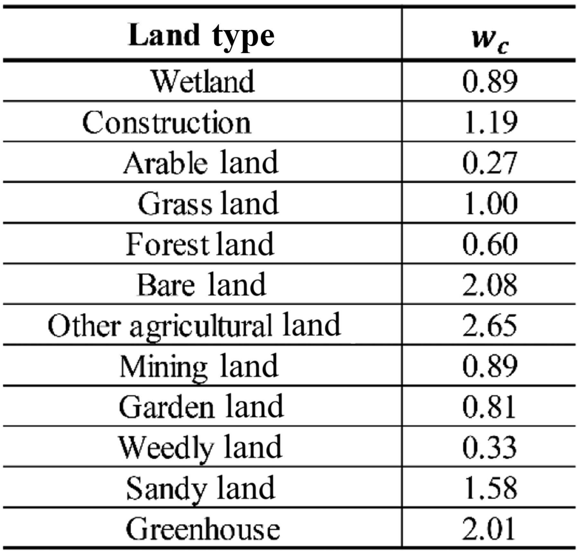
$$n_{c\_\text{freq}} = \frac{n_{c\_\text{pixel}}}{n_c \times H \times W}. \tag{5}$$

After obtaining n_{c_freq} for all categories, we denote these categories' median frequency of occurrence as $n_{\text{freq\_mid}}$. The weight w_c is then shown as follows:

$$w_c = \frac{n_{\text{freq\_mid}}}{n_{c\_\text{freq}}}. \tag{6}$$

The Yinchuan high-resolution remote sensing dataset uses the median frequency balance to obtain the weight values of 12 land categories, as shown in Table I.

TABLE I
WEIGHTS IN YINCHUAN DATASET



Land type	w_c
Wetland	0.89
Construction	1.19
Arable land	0.27
Grass land	1.00
Forest land	0.60
Bare land	2.08
Other agricultural land	2.65
Mining land	0.89
Garden land	0.81
Weedly land	0.33
Sandy land	1.58
Greenhouse	2.01

For the edge-enhanced network, we apply focal loss instead. WCE loss only considers the issue of the unbalanced number of samples and does not consider the difficulty of sample categorization. In contrast, focal loss can control both the weights of positive and negative samples and the weights of easy-to-classified and hard-to-classified samples.

The specific form of focal loss is shown in (7), where the weighting factor α is used to balance positive and negative samples, $\alpha \in [0, 1]$ for class 1 and $1 - \alpha$ for class -1. At the same time, $(1 - p_t)^\gamma$ is introduced to distinguish between easy-to-classified samples and hard-to-classified samples, $p \in [0, 1]$ for class 1 and $1 - p$ for class -1 same as α. In (7), the exponential part is used to control the sensitivity of the loss to the easy-to-classified and hard-to-classified samples. When this value is larger, the loss contributed by easy-to-classified samples will be minor, and the model will pay more attention to the hard-to-classified samples. Focal loss is equivalent to CE loss when this value is set to 0. The experiments in this article set α to 0.25 and set γ to 2.0

$$\text{FL} = \begin{cases} -\alpha(1-p)^\gamma \log\left(p\right), y = 1 \\ -\left(1-\alpha\right)p^\gamma \log\left(1-p\right), y \neq 1. \end{cases} . \tag{7}$$

In this article, we integrate the two losses with assigned weights. Since edge extraction is only involved in edge-enhanced network during training, and the main task is still semantic segmentation. It is crucial to establish hyperparameters that balance the contribution of these two tasks to the neural network loss. Therefore, the loss of the whole network is presented in (8), where Loss_{seg} represents the loss function of semantic segmentation, $\text{Loss}_{\text{edge}}$ represents the loss function of the edge-enhanced network, and β represents the loss weight parameter of the auxiliary task, and the experimental value of β in this article is set to 0.4

$$\text{Loss} = \text{Loss}_{\text{seg}} + \beta\text{Loss}_{\text{edge}}. \tag{8}$$

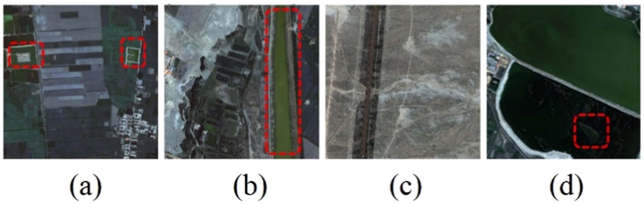


Fig. 11. Showcases of class imbalance in Yinchuan dataset. (a) Small wetland. (b) Large wetland. (c) Hard-to-segment bare land. (d) Dark green vegetation in wetland.

TABLE II
PIXEL DISTRIBUTIONS IN YINCHUAN DATASET

Land type	Pixel distribution (%)
Wetland	7.35
Construction land	5.41
Arable land	23.88
Grassland	6.41
Forest land	10.8
Bare land	3.09
Other agricultural land	2.42
Mining land	7.23
Garden land	7.9
Weedly land	19.17
Sandy land	3.13
Greenhouse	3.18

IV. DATASET

The dataset used in this article contains 6417 images of 1024 × 1024 size, all captured by the Gaofen-2 remote sensing satellite with a spatial resolution of 1 m. The dataset is segmented into training, validation, and test sets in a 7:2:1 ratio, consisting of 4533 pictures in the training set, 1295 pictures in the validation set, and 648 pictures in the test.

The following characteristics exist in the dataset.

1) *Wide variation in characteristics within the same category:* In the Yinchuan dataset, the same landform exhibits a wide range of characteristics, as shown in the red box in Fig. 11(a) and (b), where the landforms differ greatly in shape and scale and also have different color characteristics. Some streams are very dark in color and can be easily confused with the dark green cropland. The features exhibited by the bare ground in Fig. 11(c) are not the same, which can lead to intraclass inconsistency in segmentation. There are small patches of dark green vegetation in the wetland shown in Fig. 11(d), but the image features of the vegetation are very similar to the wetland, which will bring some challenges to semantic segmentation.

2) *Class imbalance:* There are 12 categories in the dataset used in this article, and the pixel occupancy of each category is shown in Table II and Fig. 12.

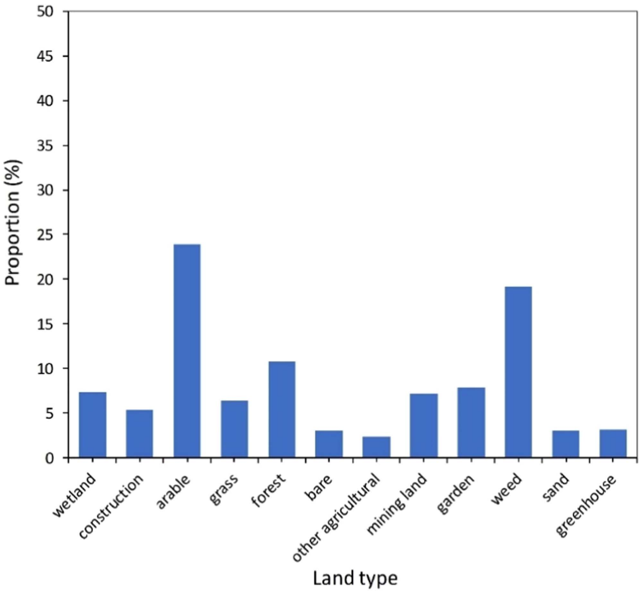


Fig. 12. Histogram of pixel proportions for different land categories.

3) *Unclear edges between different land cover categories:* The edges between different land features in the Yinchuan high-resolution RSI dataset are unclear, and the image details are not obvious. The intersecting edges of landforms may appear multiple other landforms, resulting in poor segmentation results.

V. EXPERIMENTS AND RESULTS

In order to verify the effectiveness of our proposed method, we have done many comparison and ablation experiments on the Yinchuan dataset. First, the GCIFFNet network is compared with the mainstream semantic segmentation network nowadays to verify the advancement of GCIFFNet. Then, ablation experiments are carried out on each GCIFFNet module we used, proving that each module can improve the semantic segmentation effect. Furthermore, experiments are carried out using different loss functions, proving that our loss function can achieve the optimal effect on the Yinchuan dataset.

To improve the edge extraction accuracy, we propose an edge-enhanced network and conduct comparison experiments. First, the positive effect of edge extraction using the Canny operator on semantic segmentation accuracy is verified. Then, the edge-enhanced GCIFFNet with the addition of the edge-enhanced module is subjected to ablation experiments with the original GCIFFNet network. Finally, the complete edge-enhanced GCIFFNet is compared with the mainstream semantic segmentation networks at this stage using the boundary intersection over union (BIoU) evaluation metric.

A. Evaluation Metrics

In the experiments of this article, a total of four metrics is used. They are aAcc, mAcc, mIoU, and the edge segmentation evaluation metric BIoU.

TABLE III
EXPERIMENTAL ENVIRONMENT

Operating system	Ubuntu 16.04
GPU	2 × NVIDIA TITAN RTX
CUDA version	CUDA 10.1
Programming language	Python 3.6
Deep-learning framework	Pytorch

TABLE IV
HYPERPARAMETER SETTINGS

Hyperparameter	**Value**
Input image size	512 × 512
Initial learning rate	0.01
Batch size	8
Learning rate policy	Poly
Weight decay	0.0001
Optimizer	SGD
Momentum	0.9

The overall pixel accuracy (aAcc) is used to count the proportion of correctly predicted pixels to the sum of all pixels, as shown in (9), which can also be expressed in the form of (10)

$$\text{aAcc} = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{ij}} \tag{9}$$

$$\text{aAcc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{10}$$

The average pixel accuracy mAcc refers to averaging the pixel accuracy for all categories as follows:

$$\text{mAcc} = \frac{1}{n} \sum_{i=1}^{n} \text{Acc}_i. \tag{11}$$

The BIoU is utilized as an evaluation metric for boundary segmentation [53]. As shown in (12), where G_d represents the set of pixels in the labeled image whose edges with the labeled image are not greater than d, P_d represents the set of pixels in the predicted image whose edges with the predicted image are not greater than d, G represents the labeled image, and P represents the predicted image. In the experiments on the Yinchuan dataset in this article, d is equal to 10

$$\text{BIoU} = \frac{|(G_d \cap G) \cap (P_d \cap P)|}{|(G_d \cap G) \cup (P_d \cap P)|}. \tag{12}$$

B. Experimental Environment

The experimental environment is shown in Table III.

The experimental parameters are set as shown in Table IV.

C. Comparative Experiments of GCIFFNet

In this article, we selected FCN, UNet, PSPNet [54], and DeepLabV3 for comparative experiments. We also chose DANet and CCNet based on attention mechanisms, as well as SETR, Segformer, and Segmenter grounded in transformers. This comparison with networks employing attention mechanisms illustrates that this article achieves superior experimental outcomes through the integration of multiple attention mechanisms.

TABLE V
COMPARATIVE EXPERIMENTS OF GCIFFNET

Model	**aAcc (%)**	**mIoU (%)**	**mAcc (%)**
FCN [9]	79.58	52.87	63.88
Unet [13]	79.95	51.05	60.28
DeepLabV3 [15]	84.39	64.6	79.63
PSPNet[54]	78.32	56.4	74.55
CCNet [34]	82.01	62.18	75.15
DANet [21]	86.29	66.74	83.04
SETR(MLA)[26]	77.59	53.12	67.73
Segformer (MIT-B0)[27]	84.83	64.18	76.15
Segmenter (Vit-B)[28]	86.23	66.67	76.70
GCIFFNet	89.27	74.15	89.43

The experimental results are shown in Table V. aAcc of GCIFFNet is 89.27%, mIoU is 74.15%, and mAcc is 89.43%. All three evaluation indexes are higher than other networks, and the experimental results show the superiority of GCIFFNet network model. The performance of the transformer-based model on the Yinchuan dataset is unsatisfactory, which may be caused by the three characteristics of the Yinchuan dataset (Section IV).

The segmentation results are shown in Fig. 13.

In Fig. 13, the original image, the ground truth, and the segmentation results of different networks are shown from left to right. The first row demonstrates that GCIFFNet can maintain commendable intraclass consistency. The second row reveals that GCIFFNet can pay attention to more detailed information when segmenting the construction land, and the effect is better when segmenting similar land classes. The third row proves that GCIFFNet is more effective in segmenting the small pieces of cultivated land that appear in a large piece of garden land, i.e., it is more sensitive to small-sized objects.

D. Ablation Experiments of GCIFFNet

To tackle the challenges of uneven classes and intraclass inconsistency in the dataset, we propose GCIFFNet. We capture the global information by using the RCCA module. At the same time, we added the CARB module to ensure that the model has better intraclass consistency, while preserving detailed information. The results of the ablation experiments based on the proposed network design are shown in Table VI.

The first to the fourth group of experiments show that each module of our proposed network enhances semantic segmentation to different degrees, thereby validating the efficacy of the modules we employed. Notably, the fifth group of experiments utilizes both the RCCA and CARB modules, achieving the

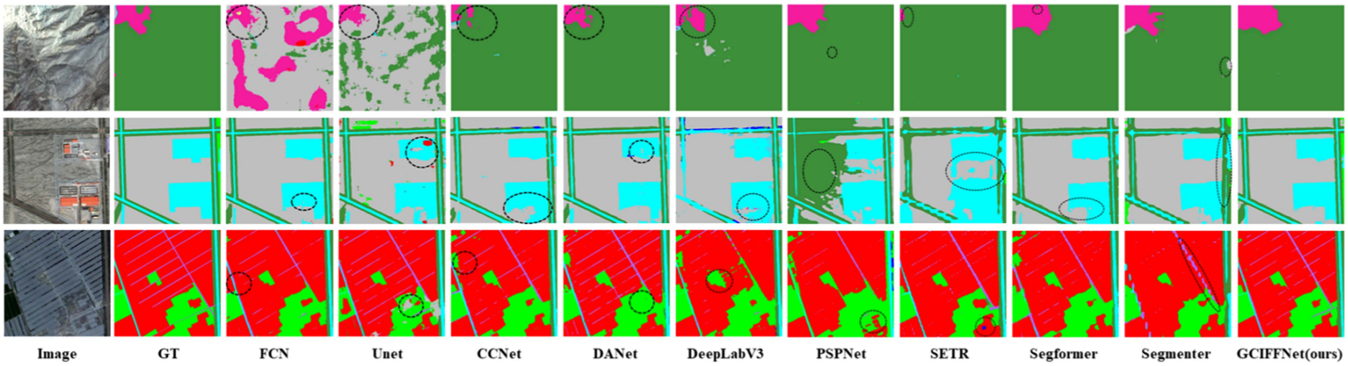


Fig. 13. Comparison of segmentation results between GCIFFNet and commonly used semantic segmentation algorithms.

TABLE VI
ABLATION EXPERIMENTS OF GCIFFNET

Model	RCCA	CARB	ConvRes	aAcc	mIoU	mAcc
Baseline	×	×	×	79.58%	52.87%	63.88%
+RCCA	√	×	×	82.01%	62.18%	75.15%
+CARB	×	√	×	85.59%	66.75%	77.93%
+ConvRes	×	×	√	79.65%	55.14%	72.09%
+RCCA &CARB	√	√	×	89.67%	73.17%	83.03%
GCIFFNet	√	√	√	89.36%	73.58%	84.42%

TABLE VII
ABLATION EXPERIMENTS OF GCIFFNET

Land types	Base (%)	+RCCA (%)	+RCCA +CARB (%)	GCIFFNet (%)
Wetland	84.31	86.54	88.06	89.7
Construction land	85.39	86.57	87.69	85.07
Arable land	88.33	92.19	93.16	93.25
Grassland	45.64	72.96	84.04	82.7
Forest land	61.49	76.59	89.61	91.6
Bare land	63.4	60.2	72.35	67.9
Other agricultural land	51.4	44.15	62.19	65.21
Mining land	15.8	63.23	90.42	88.71
Garden land	82.08	81.56	84.66	85.56
Weedly land	94.0	81.24	92.68	91.21
Sandy land	15.68	71.85	86.85	88.56
Greenhouse	79.0	74.77	64.29	83.52

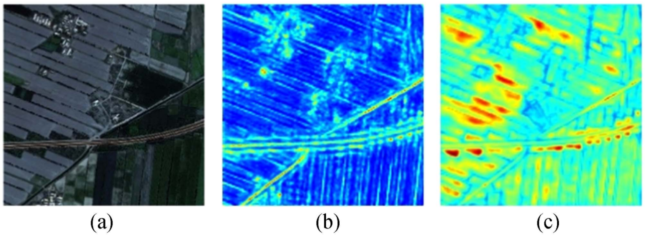


Fig. 14. Visualized feature maps obtained by RCCA module. (a) Original image. (b) Feature output of backbone network. (c) Feature after RCCA module.

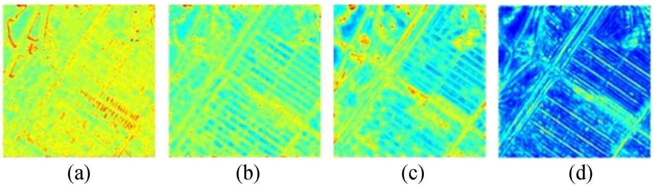


Fig. 15. Visualization of feature output obtained by different layers in backbone network. (a) Feature output by res_1. (b) Feature output by res_2. (c) Feature output by res_3. (d) Feature output by res_4.

experimental results markedly superior to those of the second and fourth groups. The sixth group of experiments adds the ConvRes module based on the fifth group of experiments, resulting in the highest overall segmentation accuracy, with mIoU reaching 73.58%, mAcc reaching 84.42%, and aAcc only 0.31% lower than the fifth group. In summary, our module combination reaches the optimum in several evaluation indexes, and aAcc is slightly inferior but comparable, proving the effectiveness of each module of GCIFFNet.

Table VII presents the pixel accuracies of the network model with different modules added to the baseline network on each land category. GCIFFNet exhibits superior performance on the extraction of wetland, cropland, forest land, other agricultural land, garden land, sandy land, and greenhouses. The performance of the other categories remains comparably consistent with the control group, demonstrating no significant decline in the recognition accuracy of a particular category.

1) Visualization of Intermediate Features: Fig. 14 illustrates the visualization of the new features obtained after the RCCA module. The feature map extracted by the RCCA module can pay more attention to some hard-to-classify areas on the sides of the road and among the cultivated land, this can suppress irrelevant features. The global information can be obtained by the RCCA module, which increases the image's intraclass consistency. The use of RCCA improves the segmentation accuracy of large-sized objects and objects with varied shapes, and reduces the problem of discontinuous prediction.

Fig. 15 illustrates the visualized feature maps extracted at four different stages of the backbone network. The shallow extracted feature maps of the backbone network contain more detailed information, while the deep features give more semantic information. Specifically, the shallow feature map shown in Fig. 15(a)

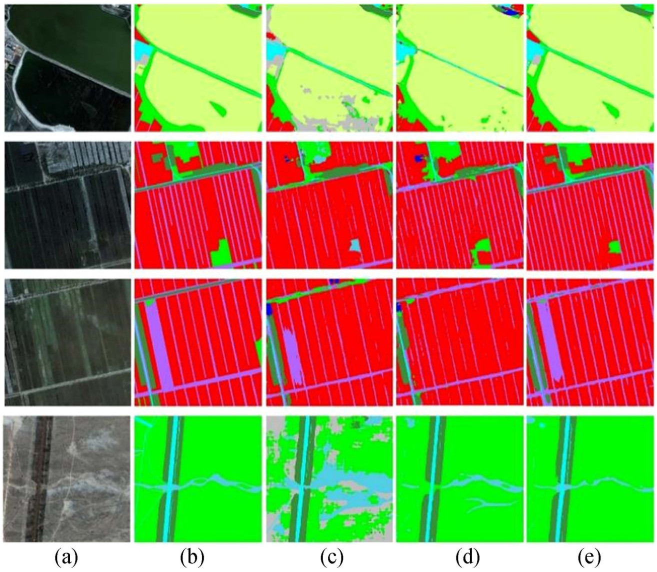


Fig. 16. Visualization of segmentation results on Yinchuan dataset by ablation models. (a) Image. (b) GT. (c) Base. (d) Base_rcca. (e) GCIFFNet.

TABLE VIII
COMPARATIVE RESULTS OF LOSS FUNCTIONS

Loss	aAcc (%)	mIoU (%)	mAcc (%)
CELoss	**89.36**	73.58	84.42
FocalLoss	83.15	51.79	61.15
Ohem	83.92	64.1	74.6
WCELoss	89.27	**74.15**	**89.43**

The bold values indicate the values that perform best in the comparison.

can distinguish each building in the complex, and Fig. 15(d) shows that the deep feature map can distinguish the category of this area. Therefore, it is necessary to combine the deep and shallow features and utilize their respective characteristics for upsampling.

2) Results of Ablation Experiments: Fig. 16 shows the original image, the ground truth, and the segmentation results of different networks from left to right, respectively.

When solely the RCCA module is integrated into the baseline network, there is a partial alleviation of the discontinuous prediction, yet the segmentation in areas requiring detailed information remains suboptimal. The addition of the RCCA module and the proposed CARB module can strengthen the network's discriminative ability and maintain intraclass consistency. At the same time, the network pays attention to more detailed information.

E. Studies of Different Loss Functions

We compared the performance of various loss functions on the Yinchuan dataset in a comparative experiment. As shown in Table VIII, the cross-entropy loss function with weights significantly improves the network's mIoU and mAcc to achieve the best results. Especially in mAcc, it is 5.01% higher than the second loss. aAcc is only about 0.1% different from the optimal result in the comparison experiment. Overall, the cross-entropy loss function with weights is better.

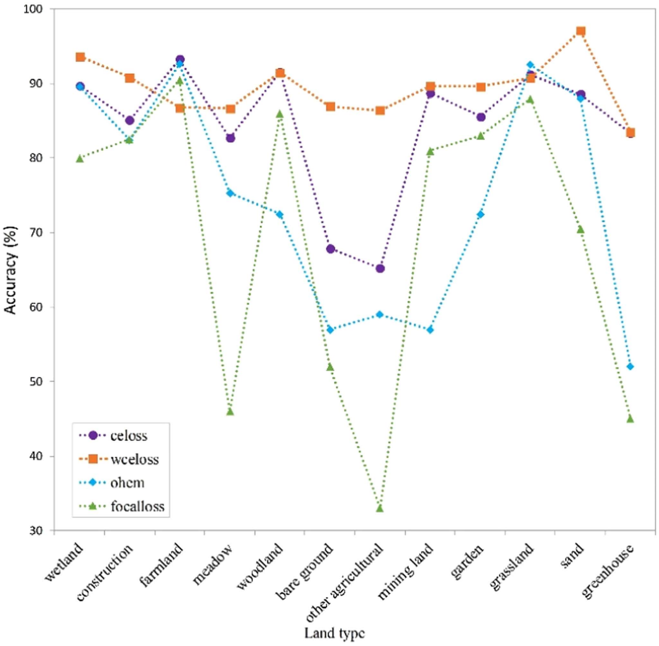


Fig. 17. Comparison of accuracy using different loss functions (CELoss, FocalLoss, Ohem, WCELoss).

TABLE IX
COMPARATIVE RESULTS OF LOSS FUNCTIONS

Land types	CELoss		WCELoss	
	IoU (%)	Acc (%)	IoU (%)	Acc (%)
Wetland	84.5	89.7	**85.53**	**93.65**
Construction land	76.4	85.07	**78.4**	**90.81**
Arable land	**86.36**	**93.25**	83.5	86.78
Grassland	69.2	82.7	**73.94**	**86.66**
Forest land	82.06	**91.6**	**84.41**	91.49
Bare land	60.05	67.9	**67.03**	**86.94**
Other agricultural land	**49.23**	65.21	46.7	**86.37**
Mining land	77.52	88.72	**82.07**	**89.67**
Garden land	74.92	85.56	74.85	**89.6**
Weedly land	84.89	**91.21**	**85.89**	90.77
Sandy land	79.81	88.56	**82.66**	**97.09**
Greenhouse	**57.97**	83.35	44.57	**83.52**

The bold values indicate the values that perform best in the comparison.

Fig. 17 shows the pixel accuracy across various land types, with the horizontal coordinates representing the land type and the vertical coordinates representing the accuracy. From the figure, it can be seen that WCELoss has optimal performance in several categories with a small sample. The accuracy on some land categories does not reach the highest value but maintains a high level.

Table IX presents the performance outcomes of the models obtained by training several loss functions on different land classes. These results clearly demonstrate that the WCELoss we use can alleviate the problems caused by the imbalance of sample categories.

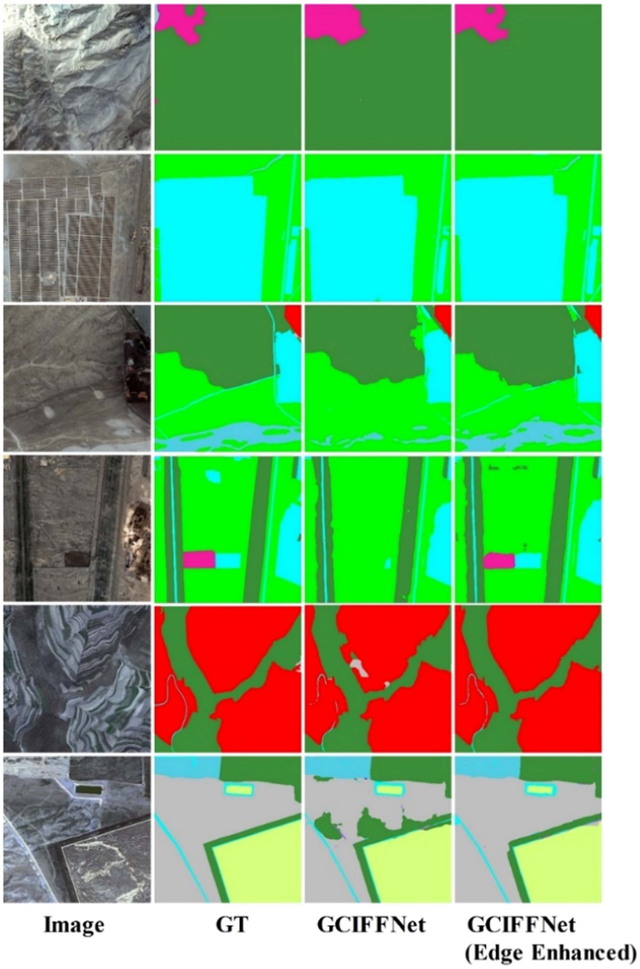


Fig. 18. Comparative results of edge-enhanced GCIFFNet and GCIFFNet on Yinchuan dataset.

TABLE X
ABLATION STUDY OF CANNY OPERATOR

Model	aAcc (%)	mIoU (%)	mAcc (%)	BIoU (%)
Without operator	**90.94**	**76.01**	86.03	87.21
With Canny	89.4	74.11	**88.4**	**88.54**
With Sobel	88.71	73.73	85.78	86.96
With Roberts	88.25	73.60	86.19	87.43
With Laplacian	87.11	71.49	86.35	88.41
With Prewitt	87.92	72.86	87.63	87.68

The bold values indicate the values that perform best in the comparison.

TABLE XI
ABLATION STUDY OF EDGE-ENHANCED NETWORK

Model	aAcc (%)	mIoU (%)	mAcc (%)	BIoU (%)
GCIFFNet	89.27	**74.15**	**89.43**	85.51
Edge-enhanced GCIFFNet	**89.4**	74.11	88.4	**88.54**

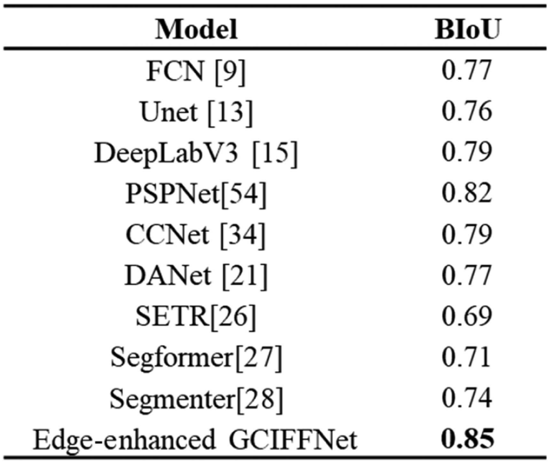
The bold values indicate the values that perform best in the comparison.

TABLE XII
RESULTS OF LAND TYPES (EDGE-ENHANCED GCIFFNET)

Land types	GCIFFNet		Edge-enhanced GCIFFNet	
	Acc (%)	IoU (%)	Acc (%)	IoU (%)
Wetland	93.71	85.79	**94.43**	**86.12**
Construction land	**92.54**	**79.62**	90.63	78.02
Arable land	**94.03**	**88.23**	93.95	87.02
Grassland	82.65	70.05	**84.22**	**70.98**
Forest land	**87.49**	**84.58**	84.1	80.5
Bare land	87.04	59.99	**89.43**	**61.72**
Other agricultural land	**86.58**	53.53	77.66	**54.25**
Mining land	**91.69**	**80.04**	89.42	77.81
Garden land	85.79	74.52	**86.7**	**76.14**
Weedly land	**90.96**	83.97	90.08	**84.2**
Sandy land	97.15	75.09	95.52	**76.18**
Greenhouse	83.56	54.35	**84.94**	**56.43**

The bold values indicate the values that perform best in the comparison.

TABLE XIII
COMPARATIVE STUDY OF EDGE-ENHANCED GCIFFNET



Model	BIoU
FCN [9]	0.77
Unet [13]	0.76
DeepLabV3 [15]	0.79
PSPNet[54]	0.82
CCNet [34]	0.79
DANet [21]	0.77
SETR[26]	0.69
Segformer[27]	0.71
Segmenter[28]	0.74
Edge-enhanced GCIFFNet	**0.85**

The bold value indicate the value that perform best in the comparison.

F. Studies of Edge-Enhanced Network

1) Comparative Experiments of Canny Operator for Edge Extraction: In the edge-enhanced network, we use the Canny operator to extract the image edges and splice them with the deepest feature maps of the backbone network to enhance the sensitivity to the image edges. In order to prove the effectiveness of the method, we conducted a comparison experiment, and the results are shown in Table X.

Table X demonstrates that the application of the Canny operator for high-level feature enhancement results in a notable improvement of 2.37% in mAcc and 1.33% in BoundaryIoU. Overall, after adding the Canny operator to the edge-enhanced network, there is a great improvement in the accurate differentiation between categories at the edges, and there is no significant decrease in other metrics.

In contrast, alternative operators, such as Sobel, do not yield significant enhancements in mAcc and BIoU. As shown in Fig. 9, the Sobel operator, the Laplace operator, and the Prewitt operator tend to extract excessive nonedge noise, leading to suboptimal results. These three operators also extract wrong edge information in the same land category. Moreover, the Roberts operator misses a lot of edge information and cannot improve the edge segmentation effect.

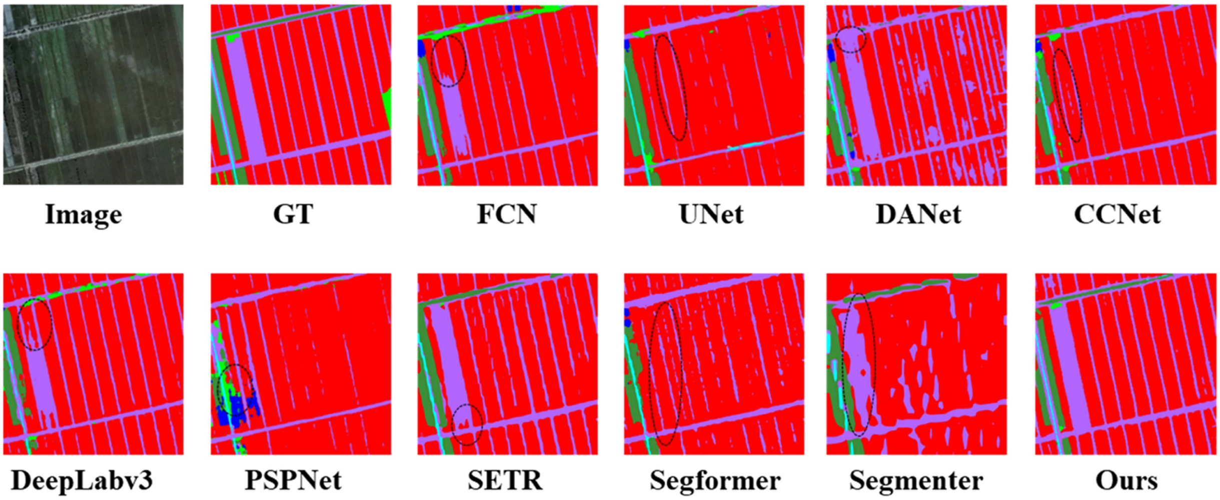


Fig. 19. Comparison of semantic segmentation results.

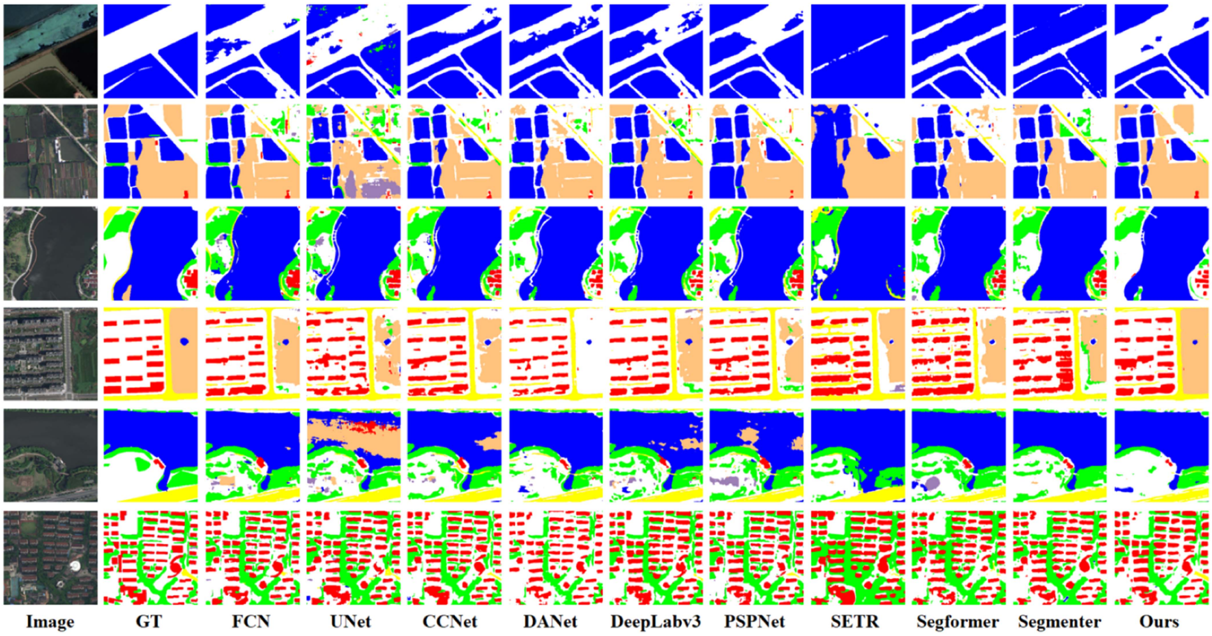


Fig. 20. Experimental results on the LoveDA dataset.

The table shows that aAcc and mIoU have decreased, mainly because the Canny operator is the extraction based on grayscale images, and some pseudo edges will be extracted in the detail-rich regions. In order to prove the above point, we use the accurate semantic edge image instead of the edge image extracted by the Canny operator on the validation set, which can be improved to 91.84%, 78.18%, 89.59%, and 89.32% in the indexes of aAcc, mIoU, mAcc, and BIoU. These results prove the rationality of utilizing edge information as auxiliary information in this article.

2) Comparative Experiments of Edge-Enhanced GCIFFNet: We compare the edge-enhanced GCIFFNet with the GCIFFNet network, including evaluation metrics, such as aAcc, mAcc, mIoU, and BoundaryIoU, with the results detailed in Table XI. BoundaryIoU is the metric used for evaluating edge segmentation, and the accuracy of the improved network in this metric is also greatly improved, reaching 88.54%. It proves that the edge-enhanced GCIFFNet we proposed can segment image edges well. The accuracy on the remaining two metrics is reduced but still higher than other networks.

Table XII shows the segmentation accuracy of edge-enhanced GCIFFNet on each land class.

As shown in Table XII, the addition of the edge-enhanced module to our network enhances the network's extraction of the edge-unclear land categories. The enhancement effect is most obvious for the land categories of wetland, grassland, bare land, garden land, and greenhouse, all of whom are prone to erroneous segmentation results due to unclear edges. On the other hand,

TABLE XIV
COMPARISON EXPERIMENT OF EDGE-ENHANCED GCIFFNET ON LOVEDA DATASET

Method	backbone	Acc/Iou per category(%)							aAcc (%)	mAcc (%)	mIoU (%)
		Background	Building	Road	Water	Barren	Forest	Agricultural			
FCN[9]	ResNet50	86.49/52.04	73.04/60.25	58.67/52.12	67.9/61.17	30.22/25.32	58.61/41.04	49.37/46.49	66.93	60.62	48.35
Unet[13]	ResNet50	78.68/47.63	73.53/53.69	56.48/48.95	55.43/50.76	36.08/25.67	64.23/35.7	44.92/41.47	61.98	58.48	43.41
DeepLabv3[15]	ResNet50	88.12/53.53	76.38/62.73	61.36/55.32	67.78/61.47	27.22/23.17	49.36/38.57	53.03/48.81	68.10	60.46	49.09
PSPNet[54]	ResNet50	88.74/52.83	**78.02**/64.51	58.14/51.94	66.12/60.54	31.79/27.69	51.05/39.78	50.12/46.84	67.59	60.57	49.16
CCNet[34]	ResNet50	87.94/53.10	76.50/62.77	59.24/53.19	69.08/62.15	29.76/24.94	53.48/41.18	50.97/47.73	67.92	61.00	49.29
DANet[21]	ResNet50	**92.35**/50.36	68.29/59.03	58.32/51.79	63.30/58.75	9.11/8.41	34.62/30.29	45.8/43.61	64.43	53.11	43.18
SETR[26]	ViT-L	60.72/45.45	77.78/55.81	**63.59**/51.97	**82.79**/61.85	**44.44/31.7**	59.63/39.04	**68.91**/53.58	66.22	65.41	48.49
Segformer[27]	MIT-B0	87.76/53.51	77.71/64.38	56.79/51.24	74.21/65.94	30.96/26.37	50.31/39.56	52.05/49.27	68.69	61.46	50.05
Segmenter[28]	ViT-B	81.92/53.64	76.97/63.36	62.12/55.01	80.60/**69.47**	28.45/24.74	58.06/43.22	61.91/55.88	70.67	64.29	52.19
Ours	ResNet50	90.17/**56.97**	75.82/**64.59**	62.08/**55.78**	73.31/67.15	35.97/28.89	**65.27/43.93**	61.08/**56.24**	**73.23**	**66.24**	**53.15**

The bold values indicate the values that perform best in the comparison.

the enhancement of construction land, arable land, and forest land, which have already clear edges, is not significant, and the accuracy is slightly reduced but still maintains a high standard. Other agricultural land has the lowest percentage of pixels in the dataset (shown in Fig. 12), and other agricultural land is not a key land category, so the reduction of pixel accuracy has no significant effect on the network effect.

Fig. 18 shows the semantic segmentation outcomes of GCIFFNet and edge-enhanced GCIFFNet. From left to right are the original image, the real label, the segmentation result of GCIFFNet, and the segmentation result of edge-enhanced GCIFFNet.

In the first line, the original image exhibits indistinct edges, where GCIFFNet fails to enhance the edges between categories. However, the accuracy is notably improved with the addition of the edge-enhanced network.

In the second and third rows, it is observed that GCIFFNet does not segment the roads. The improved network is more sensitive to the texture information of the original image, so it is more effective in this case of interclass similarity.

In the fourth line, GCIFFNet does not distinguish the light blue and pink regions in the labels, while the improved network can focus on the edges of the landforms through the auxiliary network and learn the difference between the two.

In the sixth line, the segmentation result of GCIFFNet is sticky, and the improved network alleviates this problem.

Comparing edge-enhanced GCIFFNet with other networks: In this article, the superiority of GCIFFNet is proved by comparing GCIFFNet with current commonly used semantic segmentation networks, and the comprehensive performance of edge-enhanced GCIFFNet is superior to that of GCIFFNet. Therefore, our proposed edge-enhanced GCIFFNet performs optimally in the three evaluation metrics of aAcc, mAcc, and mIoU. We compare edge-enhanced GCIFFNet with other semantic segmentation networks using the BIoU evaluation metric, and the results are shown in Table XIII.

According to the comparison of the results in the table, our proposed network achieves the highest accuracy of 0.88 in the metric of BIoU, which proves the superiority of edge-enhanced GCIFFNet in edge segmentation.

Fig. 19 shows the segmentation results of edge-enhanced GCIFFNet with other semantic segmentation networks on the test set. The features of cultivated land and forested land in the original image are similar. The other semantic segmentation networks have different degrees of mis-segmentation or omission for the regions circled by the black dashed box in Fig. 19, and edge-enhanced GCIFFNet has the best results compared to them.

G. Experiments on the LoveDA Dataset

To demonstrate the validity of our proposed model, we validate our model on the LoveDA dataset, which contains 5987 images with 0.3-m resolution. There are seven land classes in LoveDA, including building, road, water, barren, forest, agriculture, and ground. During the experiment, the training set, validation set, and test set contain 2522, 1669, and 1796 images. The parameter settings of the experiments remain the same as above. In contrast of the Yinchuan dataset, the LoveDA dataset uses background to represent all the land categories that are not the other six categories in the dataset. The results of the experiment are shown in Table XIV and Fig. 20.

As shown in Table XIV, our proposed model achieves the best performance on aAcc, mAcc, and mIoU. The dataset has a total of seven load categories, and our model scores the highest IoU metrics on five categories with the edge-enhanced module and multiple attentional mechanisms. Background category contains a tremendous amount of complex information, and our network achieves the second highest pixel accuracy for this category, with an IoU of 56.97% and at least 3% better than the other networks. Barren category is the most difficult to classify and our proposed model is at the top of the list for its segmentation.

Some of the experimental results are presented in Fig. 20.

The first row of the results shown that our network is able to discriminate the interfering ground classes in the BACKGROUND better.

The second and fourth rows of the results show that our network can solve the problem of prediction discontinuity and the experimental results are better than other networks.

The third and fifth rows of the images prove that the results of our network can be better applied in real production to assist practitioners to quickly perform landmarking.

The last row proves that our proposed network can accurately extract the shape of the land categories.

The effectiveness of our proposed network is demonstrated through experiments on the LoveDA dataset, and there is an advantage in accuracy compared to other networks. Moreover, compared to other networks, our proposed network can better help practitioners and significantly reduce their workload.

VI. Conclusion

Due to the unique characteristics of RSIs, semantic segmentation networks are prone to intraclass inconsistency when dealing with large-scale and complex regions. Simultaneously, the remote sensing dataset suffers from an imbalance in the number of samples across different land classes, resulting in reduced accuracy for certain land categories. In addition, RSIs frequently exhibit similarities between different land classes, leading to edge mis-segmentation issues.

To address the above challenges, this article proposes the GCIFFNet network. The RCCA module obtains dense contextual information, and the CARB module is proposed to fuse the low-level and high-level features at the decoder stage to obtain enhanced category features. By incorporating the RCCA and CARB modules, our proposed network can utilize both spatial attention and CA with minimal computational effort. The loss function is improved to solve the sample imbalance problem in the dataset. To solve the problem of unclear edge segmentation, an edge-enhanced network is added to obtain edge-enhanced GCIFFNet, which strengthens the network model to focus on the edge. The experiment proves that the segmentation effect of the edge is improved, whether compared with the original network or with other semantic segmentation networks. The edge-enhanced GCIFFNet achieves a 3.03% increase in the BoundaryIoU evaluation index, reaching 88.54%. Meanwhile, we validate the effectiveness of the proposed network on the LoveDA dataset, achieving the highest scores in the three evaluation metrics of aAcc, mAcc, and mIoU. Specifically, aAcc surpasses other networks by at least 2.5%, and five out of seven land categories in the dataset exhibit the highest IoU scores, affirming the superior performance of our proposed network.

However, there are still some limitations in this article. There is room for further improvement in our edge-enhanced network to enhance the effect of semantic segmentation of RSIs. Although the proposed model demonstrates strong segmentation effect, it comes with a relatively large number of parameters, resulting in lengthy training times. Subsequently, the model can be lightweight to improve the computing speed. In the future, we will focus on algorithm improvement and carry out in-depth research to deal with the numerous challenges brought by the semantic segmentation of RSIs.

References

[1] A. Mei, W. Peng, and Q. Qin, *Introduction to Remote Sensing*. Beijing, China: Higher Education Press, 2001, Art. no. 7.

[2] O. A. B. Penatti, K. Nogueira, and J. A. Dos Santos, “Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 44–51.

[3] W. Shi et al., “Change detection based on artificial intelligence: State-of-the-art and challenges,” *Remote Sens.*, vol. 12, no. 10, pp. 1688–1723, 2020.

[4] Y. Li et al., “Deep learning for remote sensing image classification: A survey,” *Wiley Interdiscipl. Rev.: Data Mining Knowl. Discov.*, vol. 8, no. 6, pp. e1264–e1271, 2018.

[5] K. Li et al., “Object detection in optical remote sensing images: A survey and a new benchmark,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 296–307, 2020.

[6] Y. Mo et al., “Review the state-of-the-art technologies of semantic segmentation based on deep learning,” *Neurocomputing*, vol. 493, pp. 626–646, 2022.

[7] J. Wang et al., “LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation,” in *Proc. Neural Inf. Process. Syst. Track Datasets Benchmarks*, 2021, pp. 1–12.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[9] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[10] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–14.

[11] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[12] A. G. Howard et al., “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, Art. no. 04861.

[13] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2015, pp. 234–241.

[14] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[15] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[16] K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra, “Draw: A recurrent neural network for image generation,” in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1462–1471.

[17] A. Vaswani et al., “Attention is all you need,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 30–45.

[18] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[19] H. Zhao et al., “Psanet: Point-wise spatial attention network for scene parsing,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 267–283.

[20] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

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

[22] L. Ding, H. Tang, and L. Bruzzone, “LANet: Local attention embedding to improve the semantic segmentation of remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 426–435, Jan. 2020.

[23] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, “GCNet: Non-local networks meet squeeze-excitation networks and beyond,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 1971–1980.

[24] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, “Asymmetric non-local neural networks for semantic segmentation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 593–602.

[25] A. Dosovitskiy et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” in *Proc. Inf. Conf. Adv. Neural Inf. Process. Syst.*, vol. 12, no. 10, Jan. 2020, pp. 1–8.

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

[27] 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.

[28] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, “Segmenter: Transformer for semantic segmentation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 7262–7272.

[29] D. Ruiling, L. Qingxiang, and L. Yuhe, “A review of research on image edge detection methods,” *Opt. Techn.*, vol. 3, pp. 415–419, 2005.

[30] Z. Yu, C. Feng, M. Y. Liu, and S. Ramalingam, “Casenet: Deep category-aware semantic edge detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5964–5973.

[31] G. Bertasius, J. Shi, and L. Torresani, “Deepedge: A multi-scale bifurcated deep network for top-down contour detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4380–4389.

[32] J. Yang, B. Price, S. Cohen, H. Lee, and M.-H. Yang, “Object contour detection with a fully convolutional encoder-decoder network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 193–202.

[33] S. Borse, Y. Wang, Y. Zhang, and F. Porikli, “Inverseform: A loss function for structured boundary-aware segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5901–5911.

[34] A. T. M. Nakamura, V. Grassi, and D. F. Wolf, “An effective combination of loss gradients for multi-task learning applied on instance segmentation and depth estimation,” *Eng. Appl. Artif. Intell.*, vol. 100, pp. 104205–104215, 2021.

[35] W. He, J. Li, W. Cao, L. Zhang, and H. Zhang, “Building extraction from remote sensing images via an uncertainty-aware network,” 2023, *arXiv:2307.12309*.

[36] R. Hang, G. Li, M. Xue, C. Dong, and J. Wei, “Identifying oceanic eddy with an edge-enhanced multiscale convolutional network,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9198–9207, Oct. 2022.

[37] 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.

[38] J. Yang et al., “Rural construction land extraction from high spatial resolution remote sensing image based on SegNet semantic segmentation model,” *Trans. Chin. Soc. Agricultural Eng.*, vol. 35, no. 5, pp. 251–258, 2019.

[39] W. Liu et al., “Remote sensing image segmentation using dual attention mechanism Deeplabv3+ algorithm,” *Trop. Geogr.*, vol. 40, no. 2, pp. 303–313, 2020.

[40] R. Hang, P. Yang, F. Zhou, and Q. Liu, “Multiscale progressive segmentation network for high-resolution remote sensing imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 5412012.

[41] F. Zhou, R. Hang, and Q. Liu, “Class-guided feature decoupling network for airborne image segmentation,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2245–2255, Mar. 2020.

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

[43] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[44] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, “Learning a discriminative feature network for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1857–1866.

[45] A. Koschan, “A comparative study on color edge detection,” in *Proc. 2nd Asian Conf. Comput. Vis.*, 1995, pp. 574–578.

[46] T. Shi et al., “Improved Roberts operator for detecting surface defects of heavy rails with superior precision and efficiency,” *High Technol. Lett.*, vol. 22, no. 2, pp. 207–214, 2016.

[47] M. Liu, J. Zhao, and N. Sun, “Edge thinning based on Prewitt operator,” *Optoelectron. Technol.*, vol. 26, no. 4, pp. 259–263, 2006.

[48] T. Shi et al., “Improved Roberts operator for detecting surface defects of heavy rails with superior precision and efficiency,” *High Technol. Lett.*, vol. 22, no. 2, pp. 207–214, 2016.

[49] J. Canny, “A computational approach to edge detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.

[50] P. T. De Boer et al., “A tutorial on the cross-entropy method,” *Ann. Oper. Res.*, vol. 134, pp. 19–67, 2005.

[51] V. Pihur, S. Datta, and S. Datta, “Weighted rank aggregation of cluster validation measures: A Monte Carlo cross-entropy approach,” *Bioinformatics*, vol. 23, no. 13, pp. 1607–1615, 2007.

[52] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[53] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, and A. Kirillov, “Boundary IoU: Improving object-centric image segmentation evaluation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15334–15342.

[54] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.