Hindawi
Complexity
Volume 2022, Article ID 7750281, 13 pages
https://doi.org/10.1155/2022/7750281

Research Article

River Segmentation of Remote Sensing Images Based on Composite Attention Network

Zhiyong Fan,[1] **Jianmin Hou**,[1] **Qiang Zang**,[1] **Yunjie Chen**,[2] **and Fei Yan**[1]

[1]*Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China*
[2]*School of Math and Statistics, Nanjing University of Information Science and Technology, Nanjing 210044, China*

Correspondence should be addressed to Zhiyong Fan; zhiyongfan1981@163.com

Received 5 July 2021; Accepted 30 November 2021; Published 5 January 2022

Academic Editor: Zhijie Wang

River segmentation of remote sensing images is of important research significance and application value for environmental monitoring, disaster warning, and agricultural planning in an area. In this study, we propose a river segmentation model in remote sensing images based on composite attention network to solve the problems of abundant river details in images and the interference of non-river information including bridges, shadows, and roads. To improve the segmentation efficiency, a composite attention mechanism is firstly introduced in the central region of the network to obtain the global feature dependence of river information. Next, in this study, we dynamically combine binary cross-entropy loss that is designed for pixel-wise segmentation and the Dice coefficient loss that measures the similarity of two segmentation objects into a weighted one to optimize the training process of the proposed segmentation network. The experimental results show that compared with other semantic segmentation networks, the evaluation indexes of the proposed method are higher than those of others, and the river segmentation effect of CoANet model is significantly improved. This method can segment rivers in remote sensing images more accurately and coherently, which can meet the needs of subsequent research.

1. Introduction

Semantic segmentation of remote sensing images is widely used. The river is an important feature target, which has important influences on the ecological environment, climate change, and human activities. Therefore, the accurate extraction of river information from remote sensing images has great application values and is of great significance for planning and construction of watercourse, monitoring of water and soil resources, and comprehensive management of watershed [1–3]. River segmentation plays an important role in the process of extracting river information.

The traditional river segmentation methods of remote sensing images mainly include morphology, wavelet transform, clustering, threshold, and partial differential equation. Sghaier et al. [4] proposed a river extraction algorithm combining partial literal science measurement and shape-related knowledge to separate rivers and lakes from images; Youssefi et al. [5] put forward a river segmentation method based on the Bayesian classifier, which should first establish the training map through morphological evaluation and then improve the segmentation results using the Bayesian method; Kang et al. [6] adopt a Gabor filter and morphological operator to enhance river information and achieve noise suppression and then give the river segmentation through an automatically determined universal threshold; and Tian et al. [7] use the corner feature, texture feature, and entropy feature of remote sensing river images to input SVM for training and divide each pixel into river or background by decision function, while Han et al. [8] proposed an improved active contour model, which takes the median absolute deviation as the external energy constraint term to design a new energy weight to accelerate the model evolution and complete the task of river segmentation.

With the continuous innovation of remote sensing technology, the details of the background information of surface features in remote sensing images are more abundant, and the non-river interference noise is quite complex, which leads to the increasing difficulty of target information extraction and river information recognition. In addition, the above algorithms are mostly semi-automatic, and subtle changes in the recognition images will lead to a lot of work in adjusting manual parameters. Besides, the algorithm itself has problems such as poor robustness, low recognition accuracy, and tedious process, which bring great challenges to river segmentation. In recent years, deep learning has become a heated topic in the study of artificial intelligence. Deep learning theory represented by convolutional neural network typically has made some achievements in the field of image classification [9–11], semantic segmentation [12, 13], feature extraction [14], smart grid [15, 16], etc. The method based on a convolutional neural network is able to complete the modeling process through automatic learning of features, avoiding the incomplete modeling process caused by human intervention in the early stage. The fully convolutional network (FCN) [17] converts the last three layers of the network into 1×1 convolutional kernel, classifies images at the pixel level, and solves the problem of image segmentation at the semantic level. Ronneberger et al. [18] build the U-Net network based on FCN, take inverse convolution as the up-sampling structure, and achieve feature information fusion using the splicing technique to obtain more spatial information; Badrinarayanan et al. [19] proposed the SegNet network, removing the fully connected layer in the network and directly connecting the one-to-one corresponding encoding and decoding network to retain a large amount of useful feature information in the images and improve the accuracy of network segmentation; Yu et al. [20] raised BiSeNet network, which is mainly composed of spatial path and context path, extracting high-dimensional nonlinear features and low-dimensional spatial features, respectively, so that the network has a broad receptive field and rich spatial feature information to realize explicit image segmentation; and using LinkNet [21] as the backbone network, Zhou et al. [22] add dilated convolution layer [23] in the central region to maximize the range of receptive field and promote multi-scale feature fusion without causing resolution loss of feature map, trying to keep the details of the object space.

The key to river segmentation of remote sensing images lies in the identification of the associated river pixel information in the images. However, in the actual segmentation task, it is difficult to establish an efficient segmentation model due to the interference of some nonassociated pixels, such as bridge, shadow, riverbank, and river-like road. To solve these problems, we propose a river segmentation algorithm based on composite attention network, named CoANet, which combines attention modules with dilated convolution layer and forms a new central region between encoding and decoding to improve the accuracy of river segmentation of remote sensing images. The proposed attentive and similarity-sensitive framework makes the proposed method a good solution that is capable of segmenting detailed river region, including tributary, riverbank, and bridge.

2. River Segmentation Model in Remote Sensing Images

2.1. Network Model. This study selects D-LinkNet network architecture for river segmentation in remote sensing images. The model uses LinkNet with precoder as its backbone network, which is composed of coding area, central area, and decoding area. The main idea of the model is to complete the segmentation by encoding the river information to the feature information and then the decoding area mapping the feature information processed by the central area to the space. For the river segmentation task, the use of the dilated convolution layer in the central area can enhance the receptive field of the feature points in the central region of the network, ensuring that the information is not lost, and meanwhile, the resolution of the feature map is not reduced.

In this model, the coding area is composed of an initial convolution module with a size of 7×7 and stride = 2 and four residual modules. The residual module uses the pretrained ResNet34 [24] structure and adopts the jump connection to enhance the generalization and characterization ability. As the core of the central region, the dilated convolution layer module employs the connection mode of series and parallel connection. Besides, a composite attention module is added to the original structure of the central region to accurately obtain information on the road feature. The decoding area adopts the residual network bottleneck connection structure with 1×1 convolution kernel [25] shown in Figure 1 to improve the network computing efficiency and also uses the up-sampling of the transposed convolution to restore the original image size. The author of this study takes D-LinkNet as the basic framework and constructs a river segmentation model based on composite attention mechanism with a repeated attention module in the central region, as shown in Figure 2.

2.2. Attention Module. The essence of the attention mechanism applied to the neural network is the distribution of a series of attention coefficients, namely the weighted processing of important features, to complete the emphasis on important information and the suppression of irrelevant information. For the river segmentation of remote sensing images, the main role of the attention module is to select the river region as the focus position from the complicated background, generating a more discriminative feature representation. To achieve the above goal, this study proposes a composite attention mechanism, which combines the attention of channel domain [26] and spatial domain [27] effectively. The composite attention module proposed will be described in detail below.

2.2.1. Composite Attention Mechanism. Woo et al. proposed the convolutional block attention module (CBAM) [28], which adopts the serial structure of channel attention branch and spatial attention branch. Park et al. proposed a

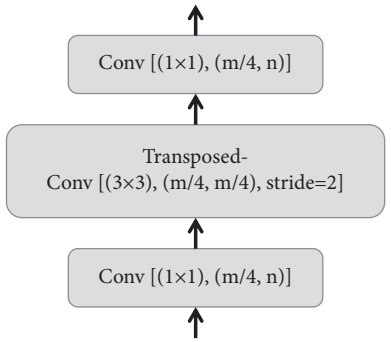

Figure 1: Residual network bottleneck structure with 1×1 convolution kernel.

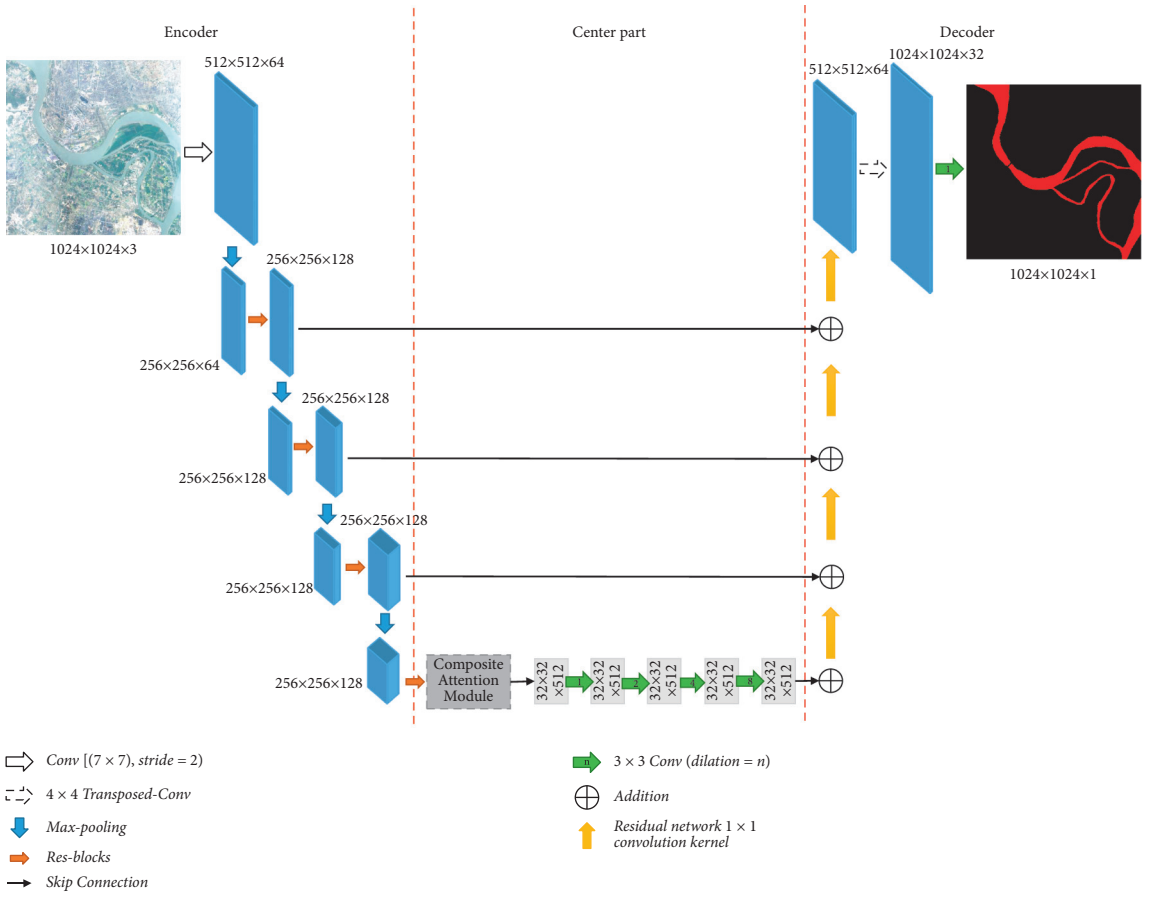

Figure 2: Composite attention network for river segmentation.

bottleneck attention module (BAM) [29], which simply adds the attention results of channel dimension and spatial dimension, making it difficult to identify small tributaries under complicated background in the river segmentation task. Therefore, to better extract features and integrate feature information, we propose a composite attention mechanism, as shown in Figure 3.

The feature matrix $Z \in \mathbb{R}^{h \times w \times c}$ of any layer passes through the channel attention branch and the spatial attention branch in parallel to obtain the channel weight matrix W_u and the spatial weight matrix W_v, respectively. By multiplying the channel weight matrix W_u and the feature matrix Z, the network can conduct weight assignment according to the importance of different characteristics of the input image. The results are multiplied by the spatial weight matrix W_v, so that the interference of the background can be removed to obtain the location information of the salient region of each feature map. In the whole process, two attention branches are applied to the feature matrix, and the composite operation of attention on the feature matrix is completed. Finally, in the form of residuals, the results are injected into the feature matrix Z, and the feature matrix $Z\prime$ with attention is obtained. The whole process can be expressed as follows:

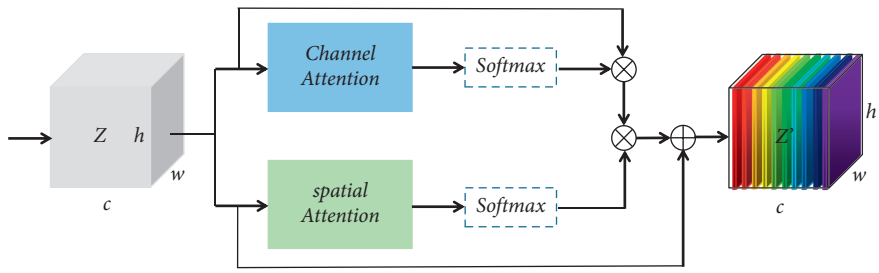

Figure 3: Proposed composite attention mechanism network structure.

$$\begin{aligned} Z' &= W_v * (W_u * Z) + Z \\ &= \sigma(V) * (\sigma(U) * Z) + Z. \end{aligned} \tag{1}$$

In formula (1), Z represents the input feature matrix and $Z\prime$ represents the output result. U and V represent channel and spatial feature matrix, respectively. W_u and W_v represent channel and spatial weight matrix, respectively. σ represents the softmax function, assigning the weight parameters of 0-1 to ensure that the sum of all weights on this dimension is 1.

$$\begin{cases} W_u = \left[W_{u_1}, W_{u_2}, \ldots, W_{u_c}\right], \quad W_{u_i} = \sigma(U_i) = \dfrac{e^{U_i}}{\sum_{i=1}^{c} e^{U_i}}, \\ 0 < W_{u_i} < 1, \qquad \sum_{i=1}^{c} W_{u_i} = 1, \end{cases} \tag{2}$$

$$\begin{cases} W_v = \left[W_{v_1}, W_{v_2}, \ldots, W_{v_{h\times w}}\right], \quad W_{v_i} = \sigma(V_i) = \dfrac{e^{V_i}}{\sum_{i=1}^{h\times w} e^{V_i}}, \\ 0 < W_{v_i} < 1, \qquad \sum_{i=1}^{h\times w} W_{v_i} = 1 \end{cases} \tag{3}$$

In formulas (2) and (3), c represents the number of channels, while $h \times w$ stands for the size of each feature map.

2.2.2. Channel Attention Mechanism. In the traditional methods of channel attention, such as squeeze and excitation networks (SeNet) [30] and BAM [29], the average pooling is adopted to compress the spatial dimension, which fails to extract the texture features fully. However, CBAM [28] directly adds the global average pooling and max pooling results of the input feature matrix, making the combination too simple. To fully retain the river texture information in the segmentation task, the author of this study adopts the method of splicing the two pooling results, as shown in Figure 4.

The function of the branch is to allocate the weight of the input feature matrix $Z \in \mathbb{R}^{h\times w\times c}$ in dimension C according to the importance of each feature map. The original feature matrix is compressed and mapped from space $h \times w \times c$ to space $1 \times 1 \times c$ by adopting the global average pooling and max pooling. The pooling results of the two are spliced, and feature map U with dimension $1 \times 1 \times c$ is obtained to remove the interference of spatial location information. Since the channel number of the original input feature map is C, it needs to go through two 1×1 convolution kernels to reduce the number of channels to further extract the channel features, and r represents the channel compression ratio. In addition, the $F_{\mathrm{BR}}(\cdot)$ transformation is performed between the two convolution kernels, as shown in the following formula:

$$F_{\mathrm{BR}}(\cdot) = \mathrm{Relu}(BN(\cdot)). \tag{4}$$

That is, using normalization and ReLU activation function successively. The attention branch of the above channel can be expressed as follows:

$$U = \mathrm{Conv}_{1\times1}(\mathrm{ReLU}(BN(\mathrm{conv}_{1\times1}(\mathrm{concat}(\max\mathrm{pooling}(Z), \mathrm{avgpooling}(Z)))))). \tag{5}$$

In formula (5), U represents the channel feature matrix and BN is normalization, while ReLU is the activation function. After the channel feature matrix $U \in \mathbb{R}^{1\times1\times c}$ passes through softmax, the weight matrix $W_u \in \mathbb{R}^{1\times1\times c}$ is obtained.

2.2.3. Spatial Attention Mechanism. Spatial attention is designed to remove the interference of image background information, such as the algorithm CBAM [25] using the pooling method to compress the channel in the spatial branch, BAM [26] using serial convolution, and dilated convolution to realize channel compression. To better eliminate the background information interference and obtain more abundant information about river features, the author of this study adopts the two-way parallel convolution structure when compressing the channel, as shown in Figure 5.

$$V_1 = \mathrm{Conv}_{1\times1}(\mathrm{ReLU}(\mathrm{BN}(\mathrm{conv}_{1\times1}(Z)))), \tag{6}$$

$$V_2 = \mathrm{Conv}_{1\times1}(\mathrm{ReLU}(\mathrm{BN}(\mathrm{conv}_{1\times3}(\mathrm{Conv}_{3\times1}(Z))))), \tag{7}$$

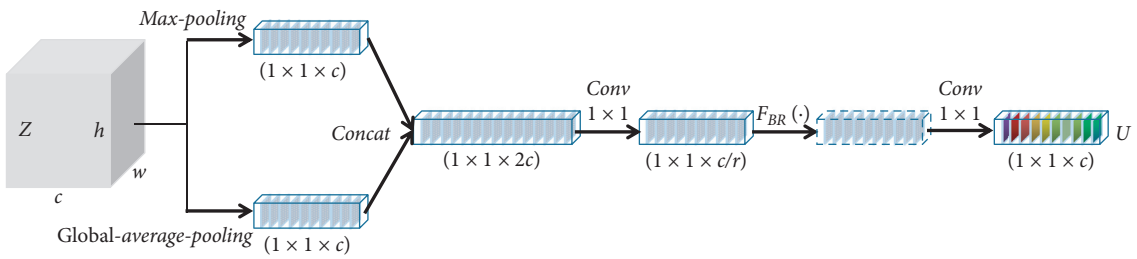

Figure 4: Proposed channel attention branch network.

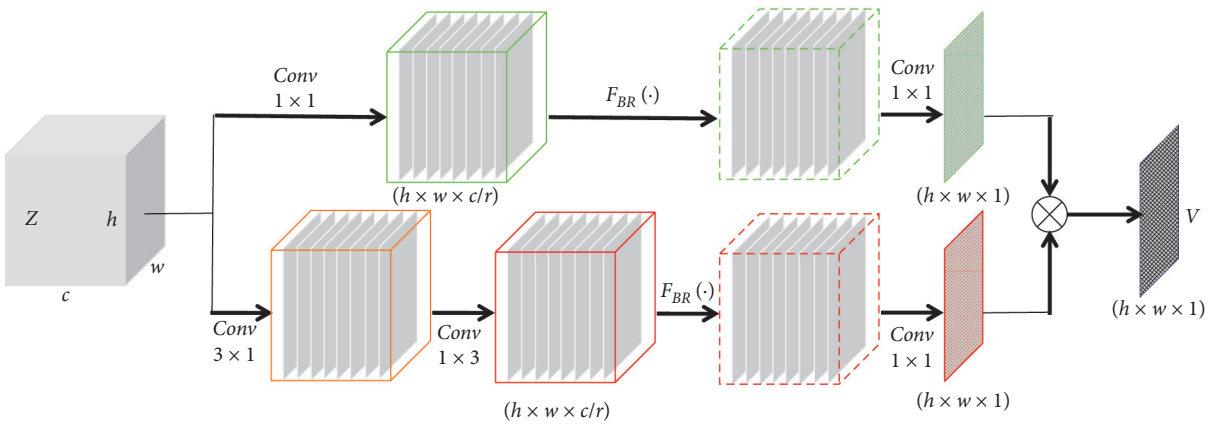

Figure 5: Proposed spatial attention branch network.

$$V = V_1 \otimes V_2, \tag{8}$$

The two parallel branches apply 1×1 and 3×3 convolution kernels to realize feature extraction of the input feature matrix $Z \in \mathbb{R}^{h \times w \times c}$ to obtain diversified feature information. To reduce the calculation amount and save the calculation cost, the 3×3 convolution kernel is decomposed into 3×1 and 1×3 convolution kernels. Based on the results, the two branches are transformed by $F_{\mathrm{BR}}(\cdot)$ and convoluted to map the feature information to space $h \times w \times 1$. For feature descriptors V_1 and V_2, multiplying the corresponding elements, the spatial feature matrix $V \in \mathbb{R}^{h \times w \times 1}$ is obtained by the feature fusion of the two matrices to obtain more abundant spatial information. The above spatial attention branch can be expressed as follows:In the formula, V_1 is the feature descriptor obtained by the upper branch in Figure 5; V_2 is the feature descriptor obtained by the lower branch; and V stands for spatial feature matrix, while $\otimes$ represents the multiplication of the corresponding elements of the matrix.

In the river segmentation task, we visualize the weights of the final outputs of attention network. Figure 6 shows that the network with attention mechanism is more expressive in the river region. The yellow region with higher brightness in the attention network indicates that the features are classified as rivers with higher reliability, and it is easier to identify the location of the river.

2.3. Network Framework Diagram. The attention mechanism is used in neural networks, usually in the form of encoder-attention-decoder. According to the D-LinkNet segmentation network structure shown in Figure 2, an attention mechanism can be introduced in the central area to improve network performance. Since D-LinkNet is an end-to-end neural network structure without complicated learning parameters and no need for a large number of redundant calculations, the attention module can be directly added to the network. Based on the above analysis, the author of this study adds the proposed composite attention model to the network, as shown in Figure 7, aiming at reducing complicated background interference in the river segmentation task, enhancing river feature information, and effectively improving the accuracy of river segmentation.

In the structure frame diagram with the attention composite module shown in Figure 7, the pretrained ResNet34 is connected to the left side of the attention module, which is used as an encoder to fully activate the network's representational ability. Connected to it on the right are the five branches added to the dilated convolution operation module, which forms the central area of the network together. As a special pooling operation, the dilated convolution has the advantage of increasing the range of the receptive field without losing feature information, which can output richer feature information in the task of river segmentation with the composite attention module. The green arrow in the figure indicates the network depth, which is 4, 3, 2, 1, and 0 (0 means identity mapping), and the corresponding receptive field size is 15, 7, 3, 1, and 0. This structure can complete multi-depth and multi-size feature fusion without loss of resolution. Subsequently, a 1×1 convolutional layer, sigmoid function, and ReLU activation

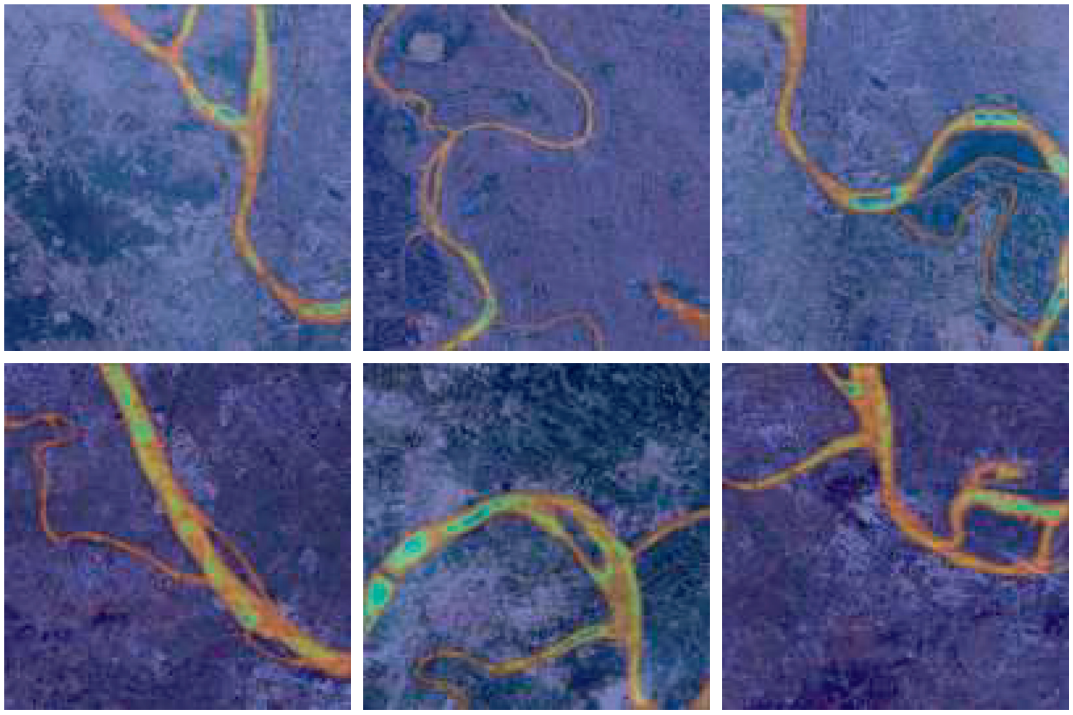

Figure 6: Visual map of attention network.

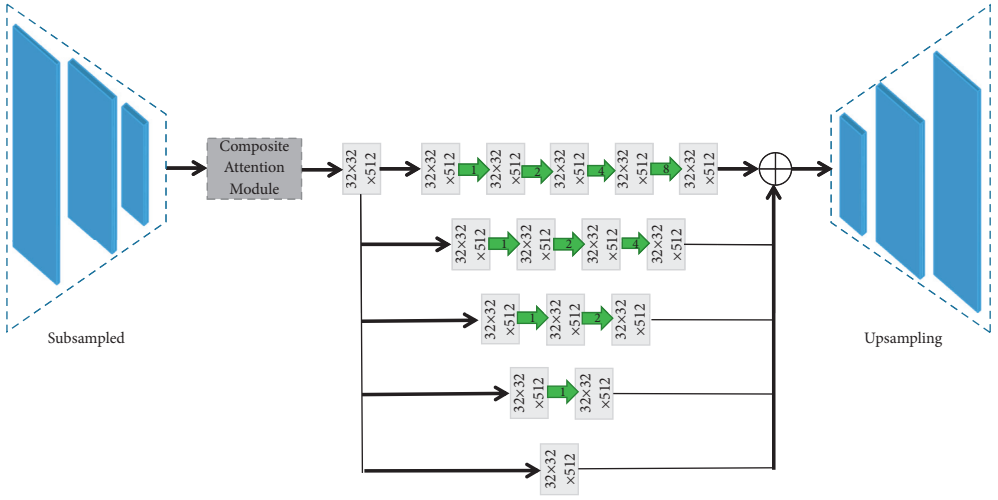

Figure 7: Structure frame diagram with attention composite module.

function are used to obtain the probability map of river prediction, and the binary image of the river segmentation prediction of remote sensing images is obtained according to the preset threshold.

2.4. Construction of Loss Function 2. The loss function is an important part of the deep learning network, which is used to calculate the differences between the network prediction result and the real label to optimize the network parameters through back propagation update. At present, in the field of deep learning semantic segmentation, the most widely used loss function is cross-entropy [20]. D-LinkNet uses the Dice coefficient (DICE) loss and binary cross-entropy (BCE) loss functions [22] to constitute the form of DICE + BCE loss function, among which BCE satisfies the "maximum entropy principle" to optimize the network output so that the predicted output is consistent with the real label, while DICE is a set similarity measurement function used to measure the similarity between two samples. The author presents a new weighted loss function based on the original DICE + BCE loss function. By weighting the two losses, the performance of the network model for river segmentation and prediction is optimized:

$$\text{Loss} = \mu\left[\frac{2\times\sum_{n=1}^{N}\left|P_i\cap Y_i\right|}{\sum_{n=1}^{N}\left(\left|P_i\right|+\left|Y_i\right|\right)}\right]+\lambda\left[\sum_{n=1}^{N}\text{BCELoss}\left(P_i,Y_i\right)\right], \quad (9)$$

where

$$\text{BCELoss}\left(P,Y\right) = -\sum_{i=1}^{W}\sum_{j=1}^{H}\left[y_{ij}\cdot\log p_{ij}+\left(1-y_{ij}\right)\cdot\log\left(1-p_{ij}\right)\right], \quad (10)$$

$$\mu+\lambda = 2. \quad (11)$$

In formula (9), μ and λ represent the weight parameter of DICE and BCE, respectively, n is the serial number of the current iteration sample, N stands for the batch size, P is the prediction probability graph of the output, and Y is the real label.

3. Experimental Results and Analysis

3.1. Data Set. In this study, the Landsat 8 satellite image is selected to make experimental data set. Landsat 8 satellite was successfully launched by the National Aeronautics and Space Administration (NASA) on February 11, 2013. The satellite carries two sensors, namely the operational land imager and the thermal infrared sensor. Landsat 8 is consistent with Landsat 1–7 in terms of spatial resolution and spectral characteristics. The satellite has 11 bands, with a spatial resolution of 30 meters in bands 1–7 and 9–11, and a full-color band with a resolution of 15 meters. The satellite can achieve global coverage every 16 days. The images studied in this study are all real color images synthesized in 4, 3, and 2 bands, which are close to the real color of the ground objects. The images are flat and gray, which can be used in river segmentation. To make the algorithm more general, as shown in Figure 8, data acquisition possesses the following characteristics including wide distribution of regions, diversity of geomorphic features, and diversity of river morphology. In the process of acquisition, to apply the collected image to the segmentation network proposed in this study, the size of each sample is 1024×1024. 18000 remote sensing satellite images are gathered as a river segmentation data set, and we further split it into a training set with 10800 training images, a validation set with 3200 validation images, and 3200 test images for performance testing. Therefore, in this study, the ratio of training/validation/test samples is 6: 2: 2.

With the deepening of the network, the parameters that need to be learned also increased, which will easily lead to the overfitting of the network. To solve this problem, the author of this study uses data enhancement methods to increase the amount of data, so as to enable these parameters to work normally and improve the network generalization performance. As shown in Table 1, the existing data are appropriately translated, flipped (horizontal and vertical), rotated, randomly clipped, and operated by changing the HSV saturation. Figure 9 shows the results after using the data enhancement methods.

3.2. Experimental Environment and Setting of Hyperparameter. To verify the river segmentation effect of the proposed network model, an objective evaluation is made, and a control experiment is established, and the software and hardware environment configuration of all experiments in this study is shown in Table 2.

The input of the network is the image with 1024×1024 pixel. The initial learning rate is set as 0.0002, the momentum factor as 0.9, and the batch size as 8. The learning rate is adjusted four times, and the optimization algorithm adopts stochastic gradient descent (SGD). The training process is iterated 150 times in total, and the learning rate is multiplied by 0.8 every 30 epochs. After that, the training model is saved as .pth file.

3.3. Evaluation Index. In the experiment, four evaluation indexes, namely pixel accuracy (PA), mean pixel accuracy (MPA), mean intersection over union (MIoU), and frequency weight intersection over union (FMIoU), are used as the reference basis to evaluate the river segmentation effect. The calculation formula is as follows:

$$\begin{aligned}\text{PA} &= \frac{\sum_{i=1}^{k}P_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k}p_{ij}},\\ \text{MPA} &= \frac{1}{K+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}},\\ \text{MIoU} &= \frac{1}{K+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji-}p_{ii}},\\ \text{FMIoU} &= \frac{1}{\sum_{i=0}^{k}\sum_{j=0}^{k}p_{ij}}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji-}p_{ii}}.\end{aligned} \quad (12)$$

Among the formulas, $k+1$ represents the number of categories (k target classes and 1 background class) and p_{ij} indicates the number of pixels with correct classification; that is, the real number p_{ij} indicates the number of pixels of category misjudged as category j, while p_{ji} is on the contrary as p_{ij} and p_{ij} indicate false positive and false negative, respectively.

3.4. Experiment on Weight of Loss Function. To limit the value of μ and λ to a certain range, the limiting condition is added that the sum of μ and λ is 2, as shown in formula (11). By studying the gradient of DICE and BCE, it is found that the loss of DICE is greater than that of BCE. Thus, $\mu > \lambda$ is set as the previous condition, and the best weight value can be obtained through experiments. The weights of DICE and BCE in formula (8) are set according to five prior values of (1.0, 1.0), (1.2, 0.8), (1.4, 0.6), (1.6, 0.4), and (1.8, 0.2), and the performance of network river segmentation under different weights is evaluated by various evaluation indexes. As shown in Table 3, when the values of μ and λ are 1.6 and 0.4, respectively, the evaluation indexes are higher than other

Figure 8: Diversified data set.

Table 1: Parameters of different data augmentation methods.

Augmentation method	Parameter
Translation	±15%
Random rotation	±15°
Random horizontal flip	50%
Random vertical flip	50%
HSV saturation	±50%

values, and the best performance for network prediction is obtained.

3.5. Training Process of CoANet Network. Figures 10(a) and 10(b), respectively, show the line graphs of the various evaluation indexes of CoANet and D-LinkNet proposed in this study with the iterative number after 150 epochs.

It can be seen from the figure that as the network iteration becomes stable, the indicators of CoANet are higher than those of D-LinkNet. PA changes from 0.9691 to 0.9743, MPA from 0.9467 to 0.9556, and MIoU from 0.9407 to 0.9502, while FMIoU from 0.9641 to 0.9728, and the four indicators are all improved in varying degrees. It is not difficult to see that the river segmentation ability of CoANet is better than that of D-LinkNet, and the prediction effect is also enhanced.

3.6. Comparative Analysis of Different Networks. To verify the accuracy and effectiveness of CoANet proposed in this study, five kinds of network, FCN-8s, U-Net, SegNet, BiSeNet, and D-LinkNet, are selected and compared with the network proposed in this study. As can be seen from Figure 11 and Table 4, CoANet has certain advantages compared with other image semantic segmentation networks as the training process can converge quickly and all indexes are superior.

For details such as river boundaries, small buildings along the bank, and bridges, it is still unable to predict correctly. In all the four images, there are discontinuous small tributaries, unsegmented bridges, and rough riverbank edges. The segmentation effect of BiSeNet is similar to that of SegNet, and the segmentation accuracy and main river channel segmentation are slightly improved, but there are still defects in the processing of details such as small tributaries and bridges. BiSeNet can obtain a relatively complete prediction effect of river edge, which is the same as SegNet. There are still defects in detail prediction, and the segmentation effect obtained by River 2 and River 3 is the most obvious. As a classic segmentation network, U-Net also shows its advantages when dealing with river segmentation, as shown in Figure 12(d), which can separate the river from the background well. In addition, the performance of riverbank segmentation is better than the above three networks. For some details, such as the "L"-shaped building on the riverbank in River 3, the segmentation effect is fairly good, but it does not show enough accuracy in the bridge segmentation. D-LinkNet has a good overall segmentation effect due to the addition of dilated convolution layer in the central region. For the details of bridges and tributaries, the segmentation is relatively complete, but at the edge of riverbank, there are detail segmentation errors, such as the missegmentation of riverbank edge in River 2 and the discontinuous segmentation of the "L"-shaped building on

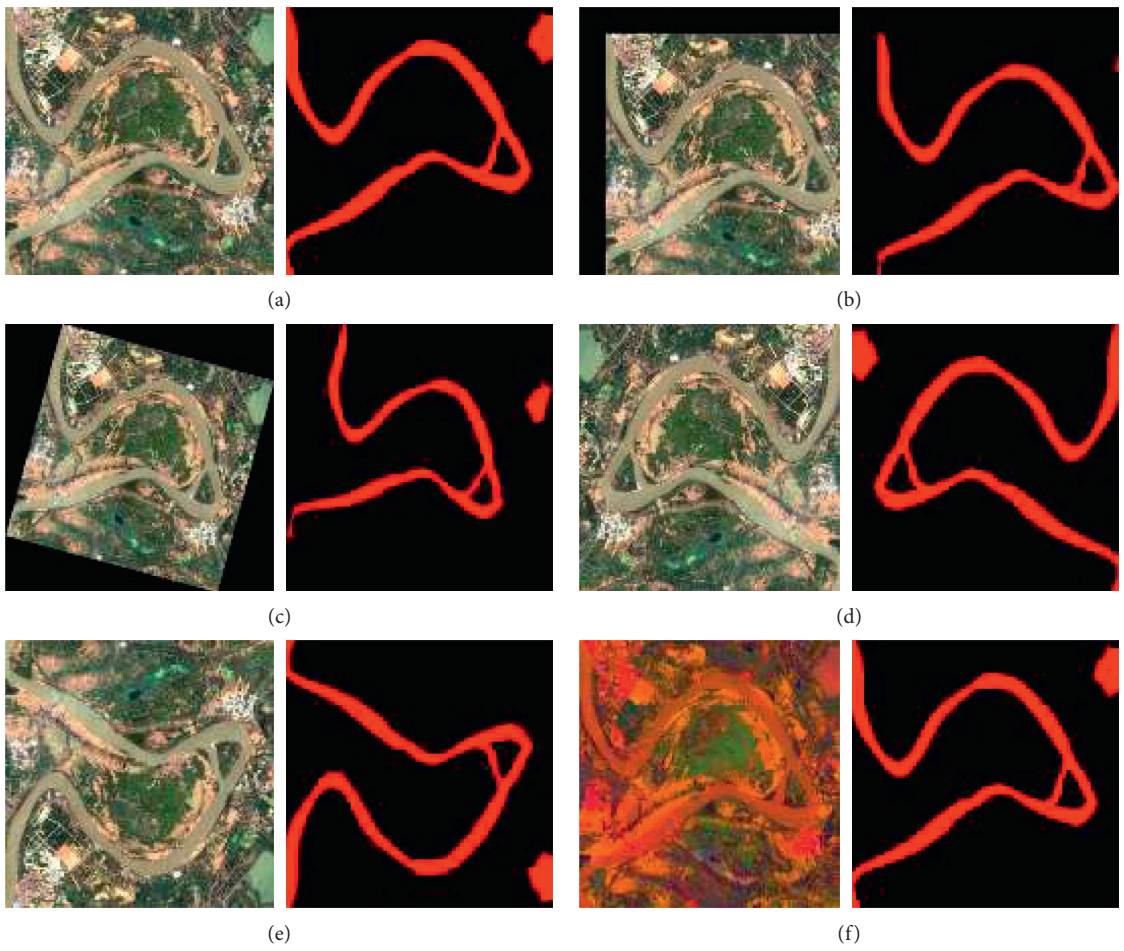

Figure 9: Diagrams of different data augmentation methods. (a) Original image. (b) Translation. (c) Random rotation. (d) Horizontal flip. (e) Vertical flip. (f) HSV saturation.

Table 2: Configuration of experiment.

	Configuration version
Operating system	64 Bit Windows 10
Processor	Intel(R) Core(™) i7-10700 CPU @ 2.90 GHz
GPU	NVIDIA GeForce RTX 3070 (8G)
CUDA	CUDA 11.0
Python	*Python* 3.8.5
Depth framework	PyTorch 1.8.1
Development tool	PyCharm 2020.3

the riverbank in River 3. CoANet is a modification of the backbone network based on D-LinkNet. The attention mechanism and the dilated convolution layer constitute the central region, which further enhances the sensitivity of the network for detail processing. As shown in Figure 12(g), not only the integrity of river segmentation but also the "L"-shaped buildings along the riverbank similar to River 3, the intersection area of River 4 double tributaries (including details of multiple bridges at the same time), and the bridges in each image all show satisfactory prediction effects. CoANet's performance in river segmentation is better than that of other comparison networks, showing its ability to deal with details.

To further verify the prediction ability of different models in river target details, local clipping is performed in Figure 12 to visually display the segmentation effect of details. As shown in Figure 13, the clipping order is the same as that shown in Figure 12: River 1, River 2, River 3, and River 4. Square frames are used to mark the location of segmentation details, and different colors are adopted to distinguish segmentation effects (yellow square frame represents better segmentation, green square frame represents poor segmentation, and blue square frame represents noise) to display the overall performance of various segmentation networks more systematically. After comparison, it is not difficult to find that the CoANet proposed in this study has the best generalization performance and shows a good processing ability for the details in each figure as other contrast segmentation networks have different degrees of error segmentation in the prediction results. Both D-LinkNet and U-Net mistakenly classify similar river areas in River 2 as rivers (green square frame), while SegNet and BiSeNet also show incoherent main river channels (green square frame). FCN has the worst segmentation effect. In addition to the above problems, noise even appears in the segmentation results of river 2 and river 3 (blue square

Table 3: The net performance comparison of different parameter weights.

(μ, λ)	PA	MPA	MIoU	FMIoU
(1.0, 1.0)	0.9703	0.9498	0.9432	0.9653
(1.2, 0.8)	0.9724	0.9537	0.9471	0.9682
(1.4, 0.6)	0.9735	0.9549	0.9489	0.9709
(1.6, 0.4)	**0.9743**	**0.9556**	**0.9502**	**0.9728**
(1.8, 0.2)	0.9728	0.9542	0.9476	0.9687

Bold indicates the maximum value of each evaluation index.

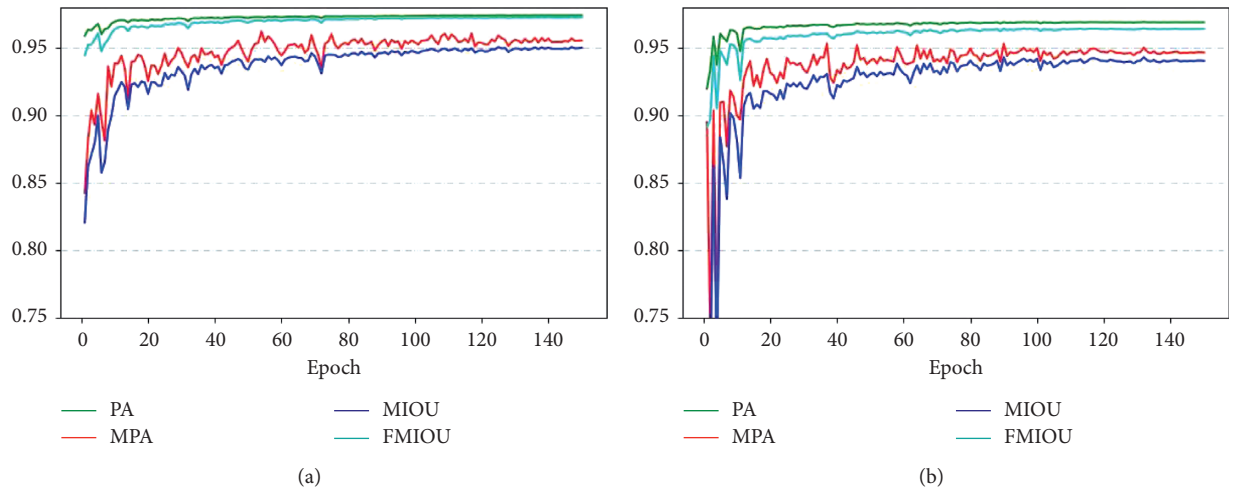

Figure 10: Each evaluation index iteration line diagram. (a) Each evaluation index iteration line diagram of CoANet. (b) Each evaluation index iteration line diagram of D-LinkNet.

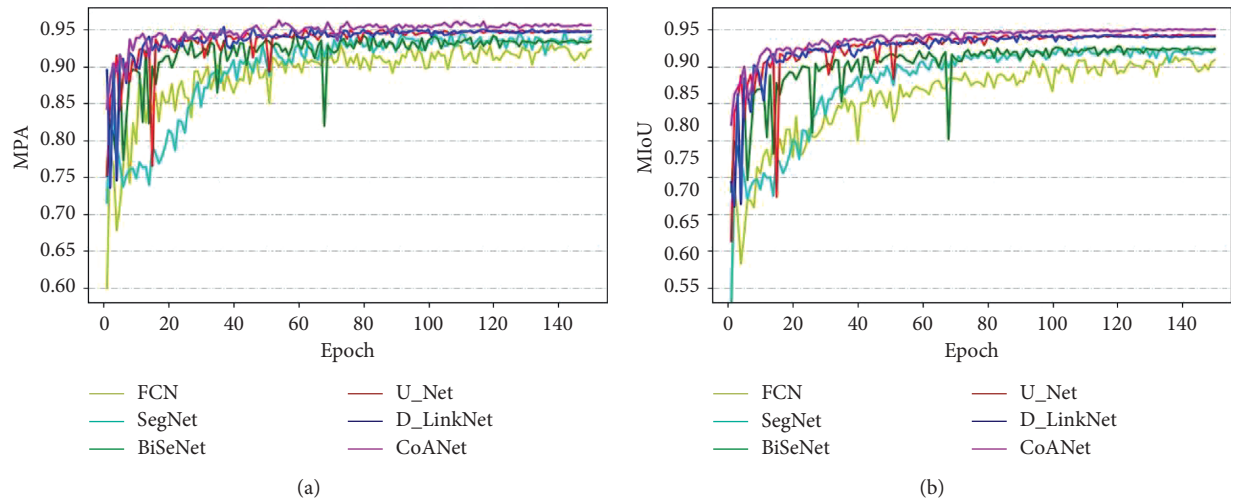

Figure 11: Variation curves of evaluation index for different networks. (a) MPA variation curves of different models. (b) MIoU variation curves of different models.

Table 4: Comparison of evaluation indexes of different river segmentation networks.

Network	PA	MPA	MIoU	FMIoU
FCN-8s	0.9524	0.9236	0.9086	0.9462
SegNet	0.9602	0.9311	0.9172	0.9548
BiSeNet	0.9617	0.9335	0.9234	0.9603
U-Net	0.9682	0.9472	0.9418	0.9674
D-LinkNet	0.9691	0.9467	0.9407	0.9641
CoANet	**0.9743**	**0.9556**	**0.9502**	**0.9728**

Bold indicates the maximum value of each evaluation index.

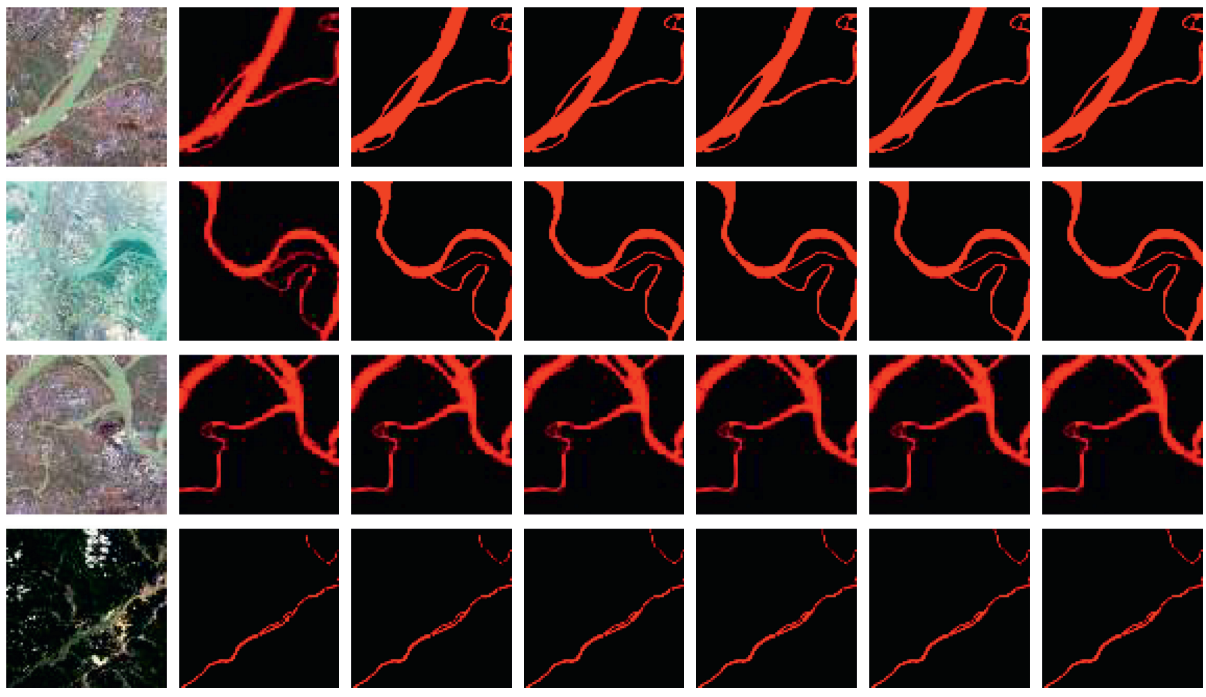

Figure 12: Comparison of effects of different networks. (a) Real image. (b) FCN-8s. (c) SegNet. (d) BiSeNet. (e) U-Net. (f) D-LinkNet. (g) CoANet.

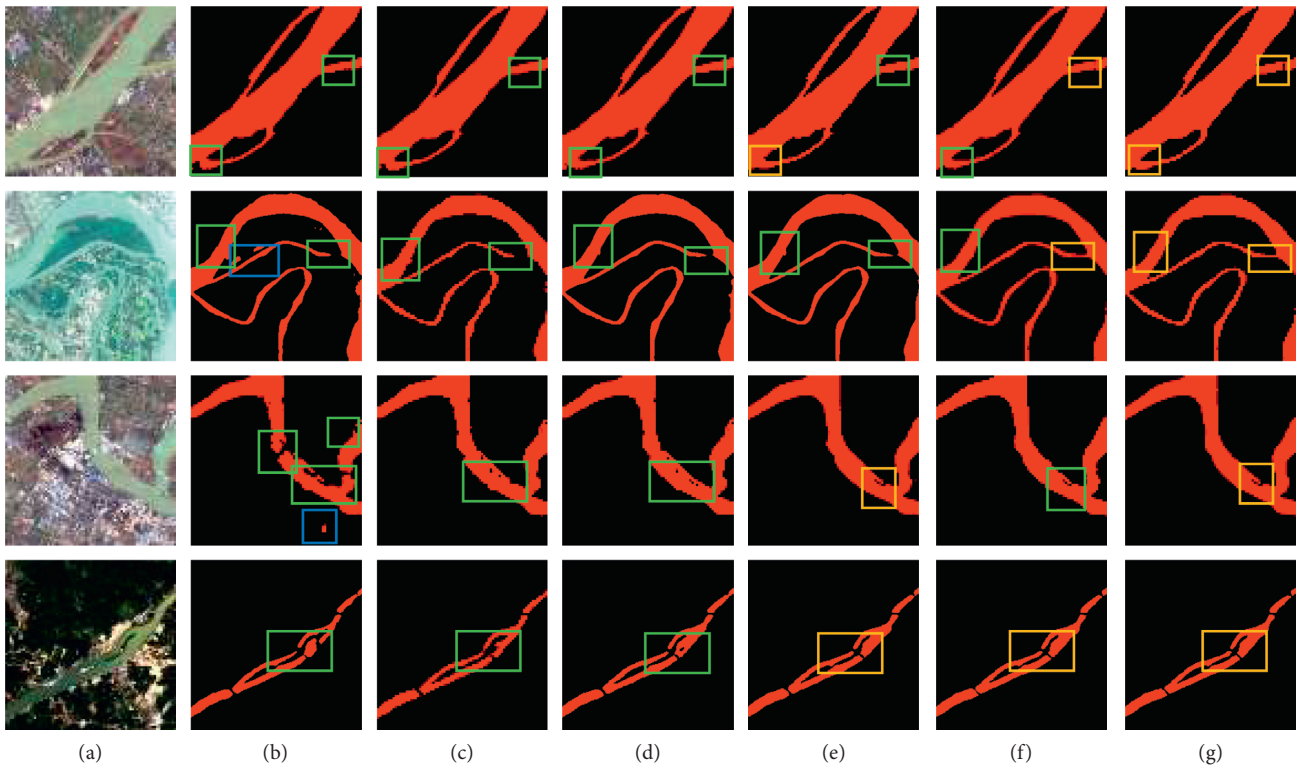

Figure 13: Partial diagrams of different segmentation networks. (a) Real image. (b) FCN-8s. (c) SegNet. (d) BiSeNet. (e) U-Net. (f) D-LinkNet. (g) CoANet.

frame). CoANet adopts the composite attention mechanism to extract the texture and location information of the river target more pertinently through the two branches of channel and space, so that the network has the ability to resist interference in the complicated background and further improves the accuracy of processing details.

4. Conclusion

Aiming at the problems of river segmentation, this study proposes an efficient extraction method based on composite attention mechanism. Based on the D-LinkNet segmentation network and aiming at the phenomena of "fake detection," "missing detection," and "false detection," which are easy to appear in the process of segmentation, a central area combining the attention mechanism and the dilated convolution layer is formed to effectively improve the accuracy of river extraction. In the training process, the original loss function is improved, a weight parameter loss is constructed, and a priori weight parameter value is preset to obtain the best effect through experiments. The experiments show that compared with the mainstream image semantic segmentation network, the proposed CoANet module has better performance in river segmentation. The future work will focus on the algorithm of the model to further optimize the segmentation performance of the network model. In addition, a lightweight backbone network also has a very important application value for the study of river segmentation.

Data Availability

The codes used in this paper are available from the corresponding author upon request (zhiyongfan1981@163.com).

Conflicts of Interest

The authors declare that there are no conflicts of interest.

References

[1] G. Schumann, R. Hostache, C. Puech et al., "High-resolution 3-D flood information from radar imagery for flood hazard management," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 6, pp. 1715–1725, 2007.

[2] G. Kaplan and U. Avdan, "Object-based water body extraction model using Sentinel-2 satellite imagery," *European Journal of Remote Sensing*, vol. 50, no. 1, pp. 137–143, 2017.

[3] C. Y. Lu, C. Y. Ren, Z. M. Wang et al., "Monitoring and Assessment of Wetland Loss and Fragmentation in the Cross-Boundary Protected Area: a Case Study of Wusuli River Basin," *Remote Sensing*, vol. 11, no. 21, 2019.

[4] M. O. Sghaier, S. Foucher, and R. Lepage, "River Extraction From High-Resolution SAR Images Combining a Structural Feature Set and Mathematical Morphology," *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, pp. 1–14, 2016.

[5] P. Yousefi, H. A. Jalab, R. W. Ibrahim, N. F. Mohd Noor, M. N. Ayub, and A. Gani, "River segmentation using satellite image contextual information and Bayesian classifier," *The Imaging Science Journal*, vol. 64, no. 8, pp. 453–459, 2016.

[6] K. Yang, M. Li, Y. Liu, L. Cheng, Q. Huang, and Y. Chen, "River Detection in Remotely Sensed Imagery Using Gabor Filtering and Path Opening," *Remote Sensing*, vol. 7, pp. 8779–8802, 2015.

[7] Z. Tian, C. Wu, D. Chen, and X. Yu, "A Novel Method of River Detection for High Resolution Remote Sensing Image Based on Corner Feature and SVM," in *Proceedings of the 9th international conference on Advances in Neural Networks - Volume Part II*, Springer, Berlin, Germany, July 2012.

[8] B. Han, Y. Wu, and Y. Song, "A Novel Active Contour Model Based on Median Absolute Deviation for Remote Sensing River Image segmentation," *Computers & Electrical Engineering*, vol. 62, 2017.

[9] R. Xu, Y. Tao, Z. Lu, and Y. Zhong, "Attention-Mechanism-Containing Neural Networks for High-Resolution Remote Sensing Image Classification," *Remote Sensing*, vol. 10, no. 10, Article ID 1602, 2018.

[10] J. Liang, Y. Deng, and D. Zeng, "A Deep Neural Network Combined CNN and GCN for Remote Sensing Scene Classification," *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, no. 99, p. 1, 2020.

[11] P. Deng, H. Huang, and K. Xu, "A Deep Neural Network Combined With Context Features for Remote Sensing Scene Classification," *IEEE Geoscience and Remote Sensing Letters*, no. 99, pp. 1–5, 2020.

[12] Z. Miao, K. Fu, H. Sun, and M. Yan, "Automatic Water-Body Segmentation From High-Resolution Satellite Images via Deep Networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, pp. 1–5, 2018.

[13] F. Mohammadimanesh, B. Salehi, M. Mahdianpari, E. Gill, and M. Molinier, "A new fully convolutional neural network for semantic segmentation of polarimetric SAR imagery in complex land cover ecosystem," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 151, pp. 223–236, 2019.

[14] M. Xia, X. Zhang, W. Liu, L. Weng, and Y. Xu, "Multi-stage Feature Constraints Learning for Age Estimation," *IEEE Transactions on Information Forensics and Security*, vol. 15, no. 1, pp. 2417–2428, 2020.

[15] M. Xia, W. Liu, K. Wang, W. Chen, and Y. Li, "Non-intrusive load disaggregation based on composite deep long short-term memory network," *Expert Systems with Applications*, vol. 160, Article ID 113669, 2020.

[16] M. Xia, W. Liu, K. Wang, X. Zhang, and Y. Xu, "Non-intrusive load disaggregation based on deep dilated residual network," *Electric Power Systems Research*, vol. 170, pp. 277–285, 2019.

[17] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2015.

[18] O. Ronneberger, P. Fischer, and T. Brox, *U-net: Convolutional Networks for Biomedical Image Segmentation*, Springer, New York, NY, USA, 2015.

[19] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, p. 1, 2017.

[20] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation," in *Proceedings of theEuropean Conference on Computer Vision*, Springer,Cham, Munich, Germany, September 2018.

[21] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation,"

in *Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP)*, December 2018.

[22] L. Zhou, C. Zhang, and M. W. D-LinkNet, "LinkNet with Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2018.

[23] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Springer, Cham, Munich, Germany, September2018.

[24] K. Hu, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*, July 2016.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[26] H. Jie, S. Li, S. Gang, S. Albanie, and E. Wu, "Squeeze-and-Excitation Networks," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[27] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," *Advances in Neural Information Processing Systems*, vol. 28, pp. 2017–2025, 2015.

[28] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional Block Attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19, Munich, Germany, September 2018.

[29] J. Park, S. Woo, J. Y. Lee, and I. S. Kweon, "Bam: Bottleneck Attention module," 2018, https://arxiv.org/abs/1807.06514.

[30] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, Salt Lake City, UT, USA, June 2018.