Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 9410381, 18 pages
https://doi.org/10.1155/2022/9410381

Research Article

River Extraction from Remote Sensing Images in Cold and Arid Regions Based on Attention Mechanism

Hailong Wang,[1] **Yu Shen**,[1] **Li Liang**,[1] **Yubin Yuan**,[2] **Yuan Yan**,[1] **and Guanghui Liu**[1]

[1] *School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China*
[2] *School of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China*

Correspondence should be addressed to Hailong Wang; 12201738@stu.lzjtu.edu.cn

Received 29 June 2022; Revised 14 July 2022; Accepted 3 August 2022; Published 22 August 2022

Academic Editor: Kuruva Lakshmanna

The extraction of rivers in cold and arid regions is of great significance for applications such as ecological environment monitoring, agricultural planning, and disaster warning. However, there are few related studies on river extraction in cold and arid regions, and it is still in its infancy. The accuracy of river extraction is low, and the details are blurred. The rapid development of deep learning has provided us with new ideas, but with lack of corresponding professional datasets, the accuracy of the current semantic segmentation network is not high. This study mainly presents the following. (1) According to the characteristics of cold and arid regions, a professional dataset was made to support the extraction of rivers from remote sensing images in these regions. (2) Combine transfer learning and deep learning, migrate the ResNet-101 network to the LinkNet network, and introduce the attention mechanism to obtain the AR-LinkNet network, which is used to improve the recognition accuracy of the network. (3) A channel attention module and a spatial attention module with residual structure are proposed to strengthen the effective features and improve the segmentation accuracy. (4) Combining dense atrous spatial pyramid pooling (DenseASPP) with AR-LinkNet network expands the network receptive field, which can extract more detailed information and increase the coherence of extracted rivers. (5) For the first time, the binary cross-entropy loss function combined with the Dice loss function is applied to river extraction as a new loss function, which accelerates the network convergence and improves the image quality. Validation on the dataset shows that, compared with typical semantic segmentation networks, the method performs better on evaluation metrics such as recall, intersection ratio, precision, and $F1$ score, and the extracted rivers are clearer and more coherent.

1. Introduction

The semantic segmentation of remote sensing images is very widely used. Rivers are an important part of the ecosystem, and accurately extracting river information from remote sensing images has extensive applications in resource exploration, early disaster warning, and agricultural planning. Cold and arid regions are widely distributed in China. As the birthplace of many rivers, the ecological environment is fragile. Therefore, it is important to find an accurate and fast river-extraction method on the basis of characteristics of cold and arid regions [1–3].

Traditional river-extraction methods include the threshold [4], water body index [5], and decision tree [6] methods. The threshold method is simple in principle and fast in extraction speed [7]. However, this method often cannot distinguish between ground objects with similar reflectivity to water bodies, such as other water bodies, shadows, and vegetation, which leads to low accuracy of water body information extraction. A water body index can more accurately distinguish water body and vegetation information, but a water body and building shadows are still easily confused. The decision tree method has higher extraction accuracy and many applications, but also slower extraction speed. In

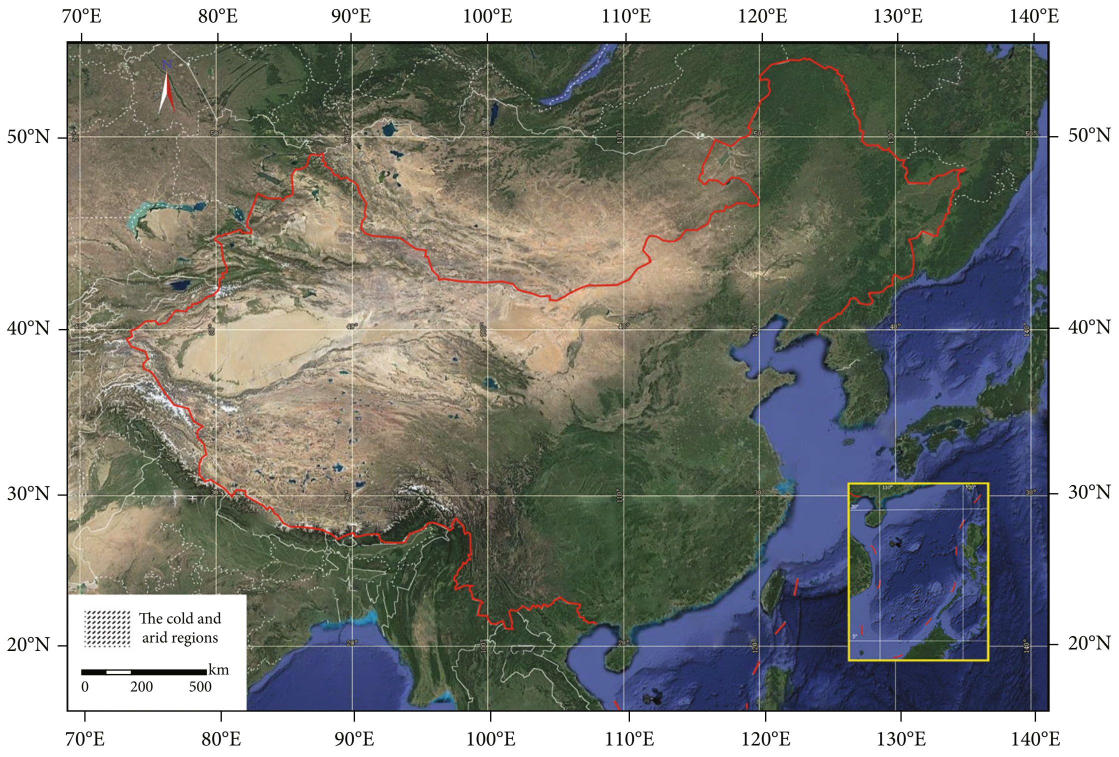


Figure 1: Experiment area: the cold and arid regions of China, as shown in the shaded area.

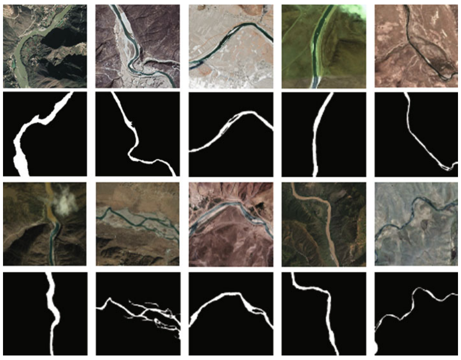


Figure 2: Dataset for extracting river water from remote sensing images in cold and arid regions.

addition, some newer methods have emerged. For instance, combined RGB remote sensing images with the characteristics of different refractive indices of various objects for river extraction, reducing the confusion of building shadows [8]. Combine the shadow–water index with the threshold and water body index methods to extract rivers in arid regions [9].

With the progress of artificial intelligence, deep learning has become a hot spot in scientific research, and great progress has been made in applying deep learning methods to the river extraction of remote sensing images [10–13]. Among them, convolutional neural networks (CNN), as an important development direction of deep learning, have had great achievements in semantic segmentation [14] and image classification [15]. With a CNN, deep features were directly extracted from the input image layer by layer, and the spectral–spatial laws of the input data were extracted. To improve the accuracy of water segmentation, combine a deep convolutional neural network with a new effective learning system, the broad learning system (BLS), which achieved good results in the classification of remote sensing images [16].

FCN learns pixel-to-pixel mapping. The input and output are pictures. Combine a full convolutional neural network with sample mining, and use the mined samples to capture different intraclass features, effectively reducing the burden of network training and achieving the rapid segmentation of water bodies [17]. Build a Unet network on the basis of FCN, introduce shortcuts, use Transposed-conv as its upsampling structure, and build a connection between the network encoder and decoder parts. Low-level information is fused with high-level information [18]. This makes the network recover more spatial information during the upsampling process, which is of great significance for fine-grained segmentation. Combine the Unet network with transfer learning and train the network using weakly supervised training methods. With only 100 pieces of data, the superior classification performance of the neural network was used to obtain pixel-level segmentation effects [18]. The ResNet network [19] introduced residual learning into deep learning, which greatly improved the accuracy of image-classification extraction and obtained superextensive applications in remote sensing image extraction [20]. In

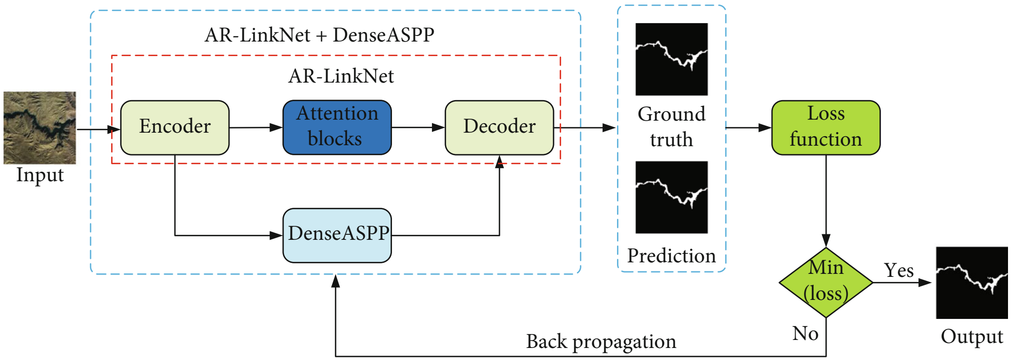


Figure 3: Network structure.

2017, Chaurasia et al. raised a LinkNet deep learning network model that uses a U-shaped structure [21, 22]. It has more shortcuts, which can make the deeper layers obtain shallower information, and achieves pixel-level image semantic segmentation. During the execution process, excessive parameter increase is avoided, and operation efficiency is improved. To resolve the contradiction between feature map resolution and receiving field, the DeepLab series [23–25] introduced expanded the convolution and pyramid pooling structure (ASPP). ASPP uses multiscale information to further improve the segmentation effect, but in the face of high-resolution input, ASPP needs a sufficiently large expansion rate to obtain a sufficiently large receptive field; however, as the expansion rate increases ($d > 24$), volume expansion decreases. Combine ASPP in the DeepLab series with dense connections in DenseNet to form DenseASPP [26]. The new module has a larger receiving field and denser sampling points [27–29]. The attention mechanism can make the network focus on the segmentation target and improve the segmentation accuracy of the network [30, 31].

In view of the characteristics of thin rivers and mountain shadows in cold and arid regions, this paper proposes an AR-LinkNet network constructed by LinkNet and ResNet networks and attention mechanism and combines DenseASPP to extract remote sensing river images. Because of network training and loss function in river extraction, it is easy to cut off the river and reduce the recognition accuracy. Therefore, we combined two loss functions with better performance and introduced the superposition of Dice loss [32] and binary cross-entropy loss (BCE loss) [33] as the semantic segmentation network of loss function in the training process. Avoid river discontinuity and improve the accuracy of river extraction.

2. Experiment Area and Data

Our data are from China's cold and arid regions, as shown in Figure 1. The coldest monthly average temperature in this area is less than −3.0°C, the monthly average temperature is greater than 10°C for less than 5 months, and average annual rainfall is less than 500 mm. It is a typical cold and arid area. Problems, lack of unified management and scheduling, and inadequate irrigation systems and technologies worsen the ecological environment in cold and arid areas by each day. At the same time, the origin of many rivers is in this area. If these problems cannot be solved in time, ecological problems in the cold and arid regions could increase further, and it would be difficult to achieve the sustainable development of the ecological environment. Therefore, this paper produced corresponding data for river extraction in cold and arid regions and provided strong support for subsequent research.

This paper collected 100 images containing rivers through Google Maps with a size of 1024×1024, with multiple and manually labeled semantic maps. The label data are the result of binarizing water data and nonwater bodies from original image data. Water bodies were labeled as 1 and nonwater bodies as 0. The collected images include multiple river scales and various representative disturbances, such as mountain shadows, cloud occlusion, road disturbances, different river-sediment content, dry-riverbed disturbances, and image-stitching changes. There was high coverage of multiple features in cold and arid regions; some sample pictures are shown in Figure 2.

3. Network Architecture

Figure 3 shows the overall river-extraction scheme used in this paper. A remote sensing image first passes through a semantic-segmentation network that combines AR-LinkNet and DenseASPP. DenseASPP is located in the middle of an encoder and a decoder of the AR-LinkNet, which is to increase the size of the feature map and network receptive field. Through the loss function, the prediction map output by the decoder is contrasted with the ground truth until the minimal loss-function value is obtained. If it is not the minimal value, the back propagation parameter is adjusted to obtain the final output semantic map. Each section is described in detail below.

3.1. AR-LinkNet. LinkNet introduces ResNet on the basis of a U-shaped full convolutional neural network to achieve pixel-level image semantic segmentation. ResNet-18 is the encoder of the original LinkNet, which has the disadvantages of low accuracy and weak characterization ability and belongs to a lightweight network. Therefore, this paper used

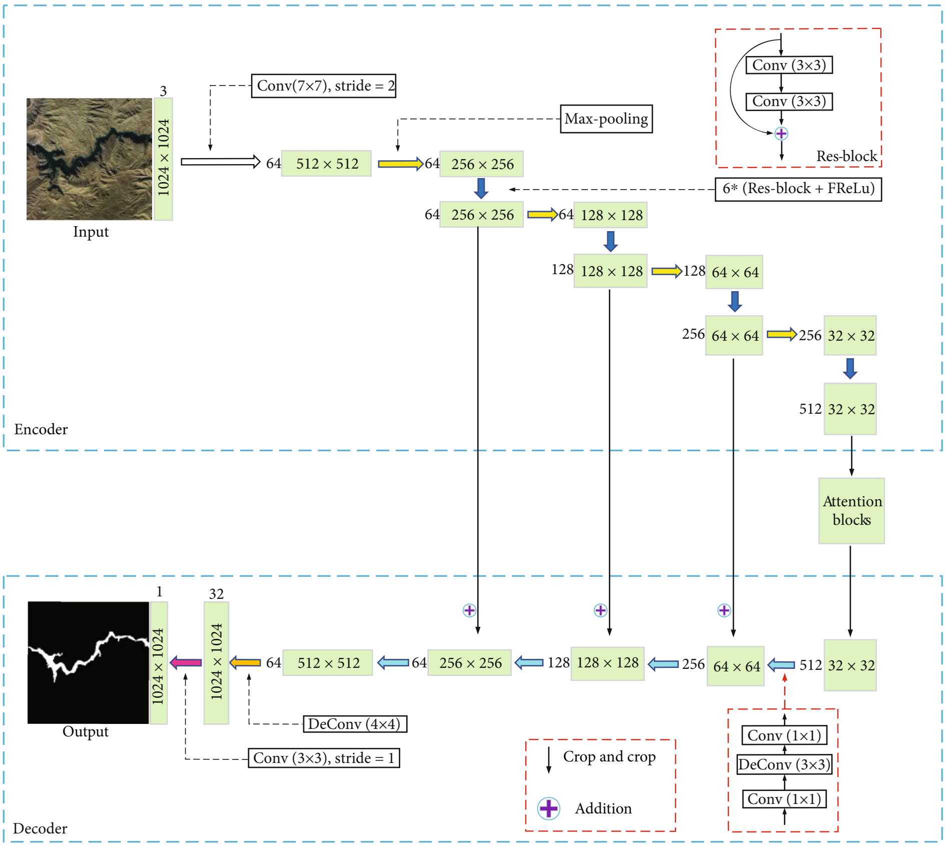


Figure 4: AR-LinkNet network structure.

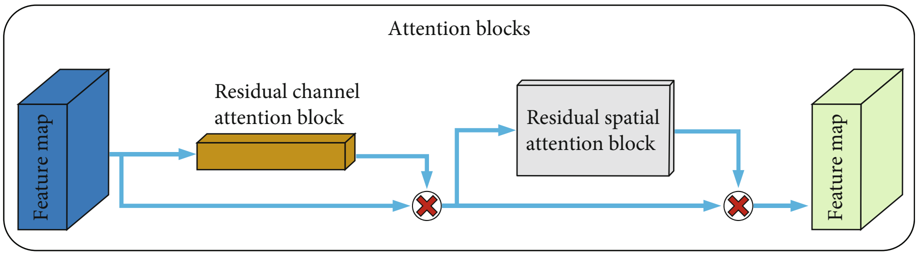

Attention blocks
Feature map
Residual channel attention block
Residual spatial attention block
Feature map

Figure 5: Structure diagram of attention block.

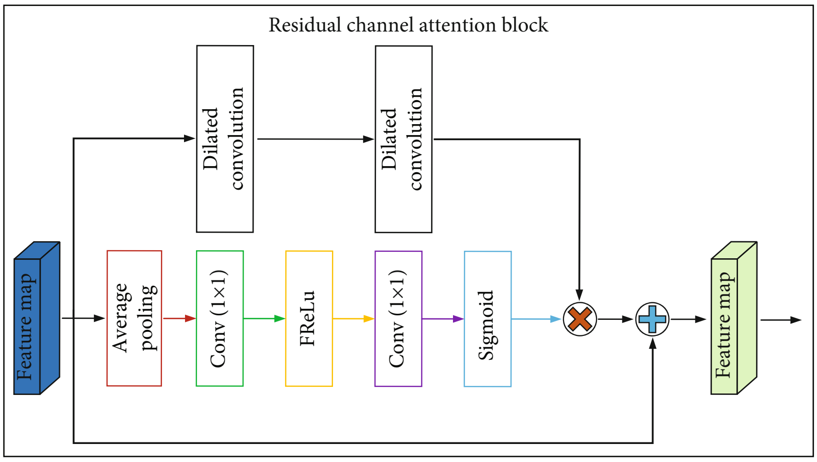


Figure 6: Structure diagram of residual channel attention module.

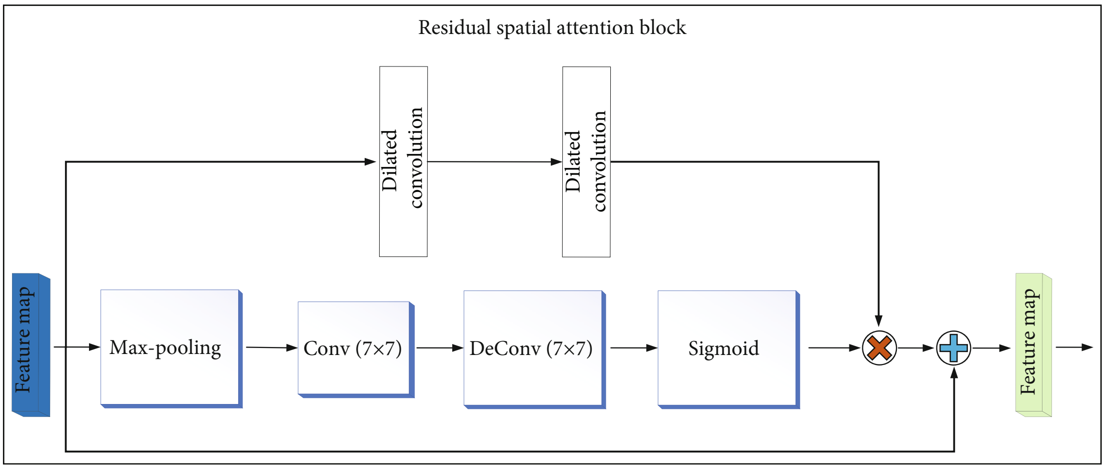


Figure 7: Structure diagram of residual spatial attention module.

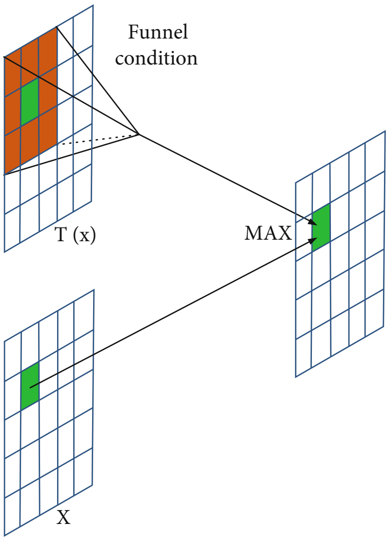


Figure 8: Two-dimensional FReLU activation function.

ResNet-101 as the encoder of the network and adds the attention blocks with residual structure between the encoder and decoder to strengthen the effective features and improve the segmentation accuracy and convergence speed of the network. At the same time, the FReLU function was the most active function to form the AR-LinkNet network. The network structure is shown in Figure 4, which included 2 convolution layers, 1 transconvolution layer, 4 encoding layers, and 4 decoding layers. Each coding layer contained 6 Res-blocks, and each decoding layer contained 2 convolution layers and 1 deconvolution layer.

Forward propagation is a process in which training data pass information through multiple hidden layers, and the prediction data are then obtained at the output layer. During the training of the river-information-extraction model, training data with a size of 1024 × 1024 pixels containing 3 feature channels were first input. The convolution kernel of the first convolution layer was set to a size of 7 × 7, and the number of steps was 64, growing to 2. After the first convolution layer, image size became 512 × 512, and the number of feature channels was 64. The specific convolution was calculated as

$$X_j^l = f\left(\sum_{i=1}^{d} X_i^{l-1} \cdot K_{ij}^l + B_j^l\right), \tag{1}$$

where X_j^l is the j-th feature map output by the l-th layer, $f(\cdot)$ is the activation function FReLU; X_i^{l-1} is the i-th channel image of layer $l-1$, K_{ij}^l is the convolution kernel matrix of the l layer, "$\cdot$" is the convolution operation, B_j^l is the bias of the j-th feature map after the l-th layer convolution, and d is the number of convolution kernels.

After the convolution operation, each pixel in the image was fused with its own original information and the information of each pixel in the 3 × 3 neighborhood. To raise network robustness, a pooling layer was set after convolution. The pooling way mainly includes taking, for example, the maximal, average, and random values. AR-LinkNet uses maximal pooling. The maximal pooling window of this network was 3 × 3, and the step was set to 2. After the pooling operation, the image size became 256 × 256, and the feature channel number was 64.

In the coding structure, feature channels of the four-time output image were 64, 128, 256, and 512, respectively. At this point, the number of feature channels of the image was expanded up to 8 times. In the decoding structure, 2 convolutional layers and 1 transposed convolutional layer were used. In this structure, the number of feature channels of the output image after four decoding instances was 512, 256, 128, and 64, respectively. At this time, the feature channels of the image were reduced to their original number.

After the encoding–decoding structure, the image enters the deconvolution layer. This process is equivalent to an upsampling operation that reduces computational complexity and maintains the spatial-position information of the image data. After transposing the convolutional layer, the

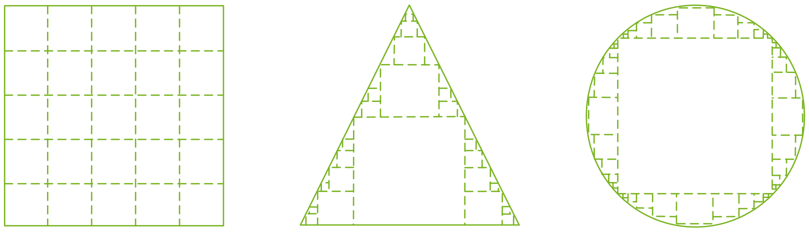


Figure 9: Description diagram of pixel-level modeling capabilities for funnel conditions.

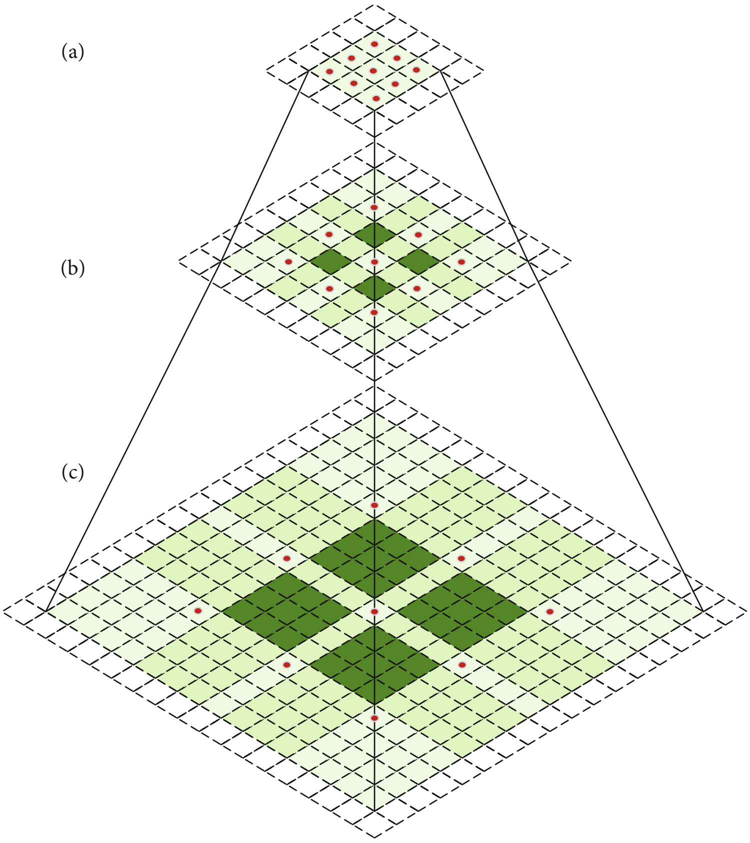


Figure 10: Dilated convolution.

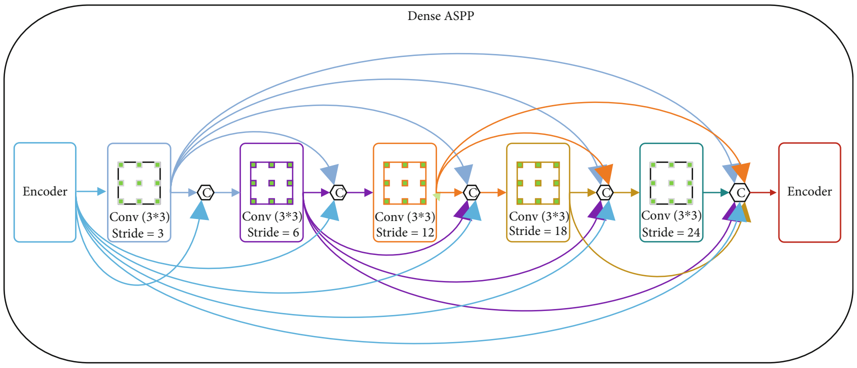


Figure 11: Dense atrous spatial pyramid pooling (DenseASPP), where *c* represents concatenation.

		Prediction	
		Position	Negative
Actual	Negative	TP	FN
	Position	FP	TN

Figure 12: Confusion matrix.

image size became 512 × 512, and the number of feature channels was 32. Finally, the image entered a convolution layer, the image size became 1024 × 1024, and the number of channels became 1. At this time, the extracted prediction data by the river were output.

3.1.1. Attention Blocks. In the task of semantic segmentation of remote sensing images, the model needs to distinguish some objects with blurred boundaries and distinguish objects with different appearances. For example, a road object and a river object are indistinguishable in the boundary area, and the river object can be affected by the shadows of snow and mountains. Therefore, it is necessary to strengthen features that are effective for segmentation tasks. Each feature channel and feature space in high-level features contain rich semantic information, which can be regarded as the prediction of a specific category, and the semantics of different categories have a certain correlation. By modeling the interdependencies between different feature channels and feature spaces, the ability of feature maps to express specific semantics can be effectively enhanced. Inspired by this, this chapter proposes two attention modules with residual structure, including Residual Channel Attention Block (RCAB) and Residual Spatial Attention Block (RSAB). This attention module does not use a parallel structure but a serial structure design and is a structure in which the residual channel attention module is in front and the residual space attention module is behind. After actual testing, it is found that this structure has the best effect, as shown in Figure 5.

First, the features that have been subjected to the residual channel attention module are fused with the original features, and the fused features are fused with the features that have been subjected to the residual spatial attention module to obtain the final output feature map. The residual channel attention model provides the weight parameters of the channel so that the channels with rich feature information get greater weight, and the residual spatial attention module provides the spatial weight parameters of the feature map so that the key points in the feature map can be obtained, greater weight. Residual attention module, it is expected that the model can obtain the importance of each feature channel and feature space through self-learning, that is, the weight. In order to make full use of the features that are effective for the segmentation task, the respective weights are applied to each original feature channel and feature space and then selectively strengthen these features so that the subsequent processing can make full use of these features, while suppressing invalid or effect small features. The RCAB module and the RSAB module perform feature recalibration on the input features in the channel and space dimensions, which strengthens the features that are effective for segmentation tasks and helps to improve the segmentation accuracy.

The residual channel attention module draws on the idea of SENet [34] and adopts the model of first compression and then expansion, and its structure is shown in Figure 6. First, the feature map extracted by the encoder is compressed using Average Pooling in the spatial dimension to obtain the global feature information of each channel, and then, the Bottleneck structure composed of two 1 × 1 Conv layers is used to compress the features. The interchannel dependency modeling is aimed at fitting the complex interchannel dependencies and reducing the amount of parameters and computational overhead. Then, use the Sigmoid function to obtain the weight of each feature channel, and then, apply the weight to the feature map processed by dilated convolution through the scale operation to obtain the feature map with channel attention.

The idea of the residual spatial attention module [35] is similar to that of the residual channel attention module, and it also adopts the idea of compressing and then expanding. Its structure is shown in Figure 7. The feature map is first max-pooled at the channel level, because this saves the most obvious feature points. Then, after a 7 × 7 (experiments have proved that in the spatial attention model, the 7 × 7 convolution kernel sampling is better than the 3 × 3 convolution kernel sampling) convolution is downsampled to obtain the compressed feature map, then pass. The 7 × 7 deconvolution upsampling restores the same size as the original feature map. Finally, the weight parameter matrix is obtained through the Sigmoid function, which is multiplied by the channel and the feature map after hole convolution to obtain the fused output feature map.

3.1.2. Visual Activation Function: FReLU. The nonlinear activation function is a necessary part of the convolutional neural network to provide good nonlinear modeling ability. Common activation functions mainly include ReLU and its evolved PReLU. However, in the field of computer vision, these activation functions cannot extract finer pixel-level spatial modeling capabilities, so the visual task activation function Funnel ReLU (FReLU) [36] semantic segmentation network proposed by Hong Kong University of Science and Technology and Megvii Technology in 2020 is used for accuracy compensation. Obtain richer spatial context semantic information. FReLU is a two-dimensional funnel-shaped activation function specially proposed for computer vision tasks. By adding a funnel condition $T(X)$ to the one-dimensional ReLU activation function to expand it to a two-dimensional space (as shown in Figure 8), only a small amount of computation and risk of overfitting is introduced to improve the vision task by activating the spatially insensitive information in the network, which is expressed as

$$f\left(x_{i,j,k}\right) = \max\left(x_{i,j,k}, T\left(x_{i,j,k}\right)\right), \tag{2}$$

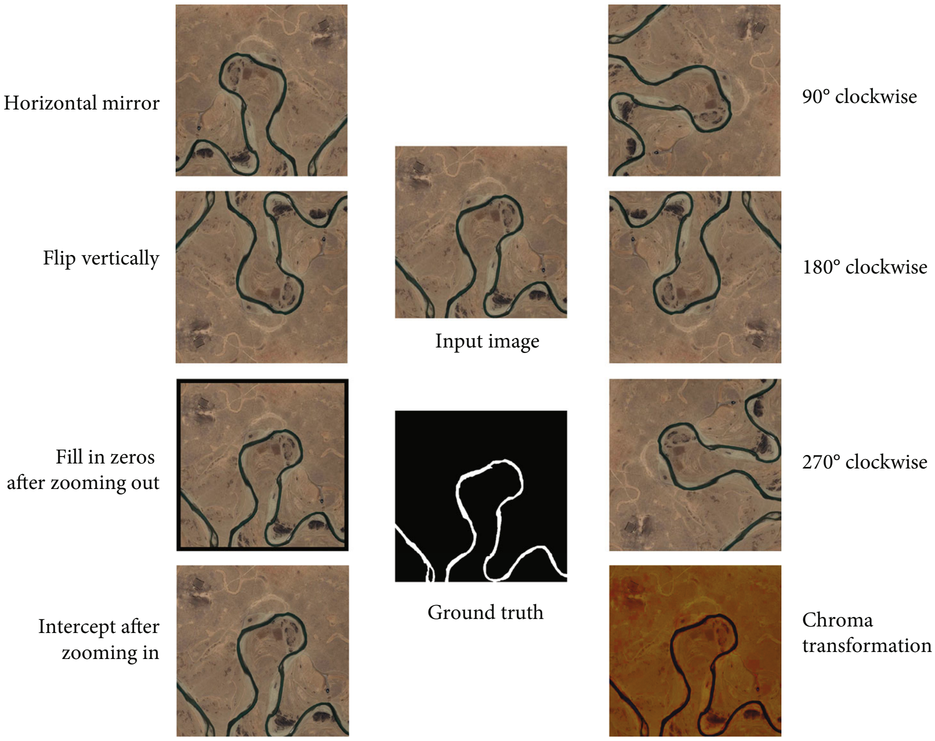


Figure 13: Database extension. An image is rotated, scaled, and transformed into 9 images.

Table 1: Super parameter setting.

Parameter	Setting
Batch	4
Epoch	100
Optimizer	Adam
Learning rate	0.01
Rate scheduler	Poly
Weight decay	0.0001
Momentum	0.9

$$T\left(x_{i,j,k}\right) = x_{i,j,k}^{\omega} \cdot p_i^{\omega}, \tag{3}$$

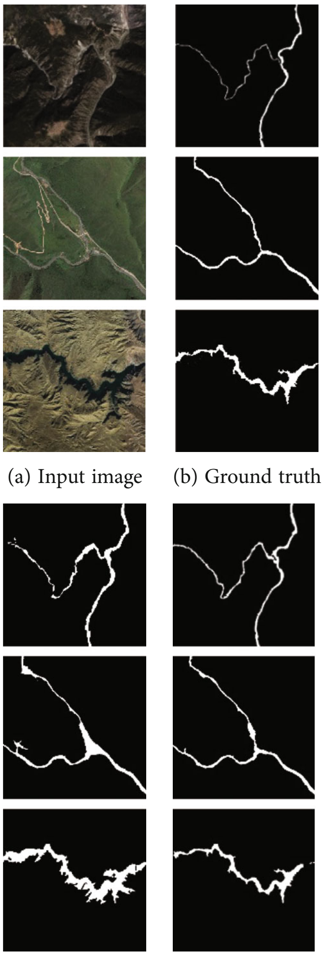
where $x_{i,j,k}$ is the two-dimensional spatial position of the nonlinear activation function $f(\cdot)$ of the i-th channel, the function $T(\cdot)$ is the funnel condition, $x_{i,j,k}^{\omega}$ is the parameterized pooling window on $x_{i,j,k}$, p_i^{ω} is the shared coefficient on the common channel, and $(\cdot)$ is the point multiplication operation.

Its funnel condition is a square sliding window with preset parameters, which is realized by depthwise separable convolution and data normalization (BN), which can improve the spatial dependence between pixels and activate spatially insensitive information to obtain rich spatial context information to improve pixel-level spatial modeling capabilities. The graphical description of the pixel-level modeling capability of the funnel condition is shown in Figure 9, which introduces only a small number of parameters and minimal complexity. Considering that in natural objects, in addition to vertical and horizontal directions, oblique lines and circular arcs are also common, so squares of different sizes are used to represent the pixel spatial information extracted by different activation layers. The slash and arc activation domains are formed through extreme approximation thinking, so as to avoid the insufficient modeling ability caused by only using ordinary horizontal and vertical activation domains.

3.2. DenseASPP. Compared with a traditional convolution operator, dilated convolution can expand the receptive field while keeping the number of kernel parameters unchanged. The size of the feature map generated by dilated convolution remains unchanged, but the receptive field of each neuron is expanded, so it can encode higher-level semantics, and each output of the convolution has a larger range of information. The principle is shown in Figure 10. Figure 10(a) represents a common convolution kernel that can also be understood as dilated convolution with a dilation rate = 1, which is a special form of dilated convolution; Figure 10(b) is dilated convolution with a dilation rate = 2 that expands a 3 × 3 convolution kernel to 7 × 7 by adding holes with a weight of 0 around 9 points on the basis of an ordinary convolution kernel. The 7 × 7 convolution kernel increased the receptive field, but only 9 original points in the figure had weights to participate in the convolution operation, and the other values were all 0. In Figure 10(c), dilation rate = 4, which expanded the receptive field range to 15 × 15.

Table 2: Results of precision and IoU values in test dataset.

Network model	ResNet-101	Attention blocks	ReLU	FReLU	Precision	IoU
LinkNet	✓	✗	✓	✗	0.802	0.589
AR-LinkNet	✓	✗	✗	✓	**0.837**	**0.611**



(a) Input image (b) Ground truth

(c) LinkNet (d) AR-LinkNet

Figure 14: River segmentation results on the test dataset.

Table 3: Attention model comparison test results.

Network	ResNet101	FReLU	Attention blocks	Precision	IoU	Average time (ms)
AR-LinkNet	✓	✓	✗	0.837	0.611	91.3
			Add only residual channel attention model	0.849	0.625	92.1
			Add only residual spatial attention model	0.843	0.619	91.8
			Two attention models in parallel	0.852	0.628	92.3
			Space first then channel model	0.856	0.634	93.1
			Channel first then space model	**0.864**	**0.645**	**92.9**

Although hole convolution to solve the contradiction between the accepted domain size characteristic map resolution, all neurons in the feature-map output by the hole convolution have the same size as the acceptance domain, that is, the semantic-mask-generation process does not take advantage of features on multiple scales. However, multiscale information helps to resolve ambiguities and produce more robust extraction results.

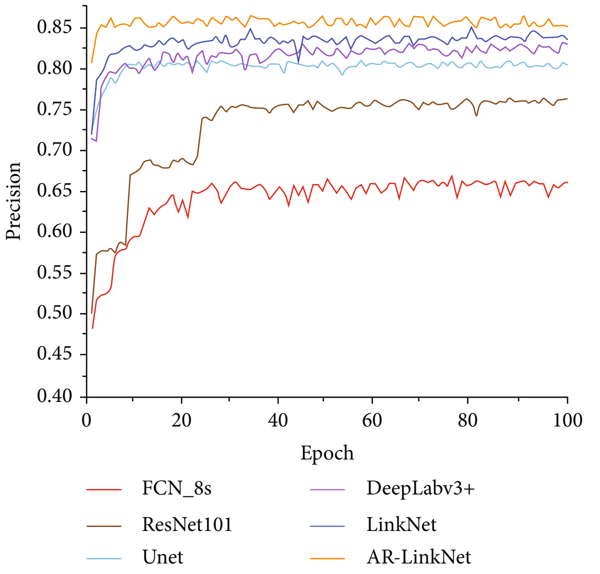


Figure 15: Curves of precision during each network test.

Table 4: Evaluation indicators of each network.

Network model	Recall	*F*1 score	IoU	Precision
FCN_8s	0.653	0.647	0.428	0.641
ResNet101	0.711	0.728	0.489	0.745
Unet	0.801	0.797	0.519	0.793
DeepLabv3+	0.826	0.821	0.575	0.817
LinkNet	0.817	0.823	0.604	0.829
AR-LinkNet	**0.839**	**0.851**	**0.645**	**0.864**

DenseASPP consists of a basic network and a series of stacked convolutional layers, and it combines the advantages of using convolutional layers in parallel and cascade to produce more scale features over a larger range. Through a series of feature connections, all neurons on the feature map of each layer are interconnected to encode semantic information from multiple scales, and different neurons encode multiscale information at different scales. Through series and parallel dilated convolutions, the receptive fields of neurons at later levels increase and avoid the problem of ASPP nuclear degradation. Finally, DenseASPP's final output feature map densely covers a large range of semantic information. Figure 11 shows the structure of DenseASPP. The feature map output by the encoder passed through five convolutional layers, the convolution kernel size of each convolutional layer was 3, and the stride was 3, 6, 12, 18, and 24. The output of each stage was concatenated with each subsequent stage, used as the input of the subsequent stage, and finally input to the decoder.

3.3. Loss Function. The loss function was used to evaluate the difference between the semantic map output by the trained model and the real semantic map. The smaller the value was, the more suitable the model was for river extraction.

The most widely used loss function in the semantic segmentation problem is the Dice coefficient loss function (Dice loss), which can truly reflect the overlap between the predicted image and the real semantic map. However, in the extreme case where the predicted image and the real semantic map are very small, the unstable phenomenon of training is easy to occur. Therefore, this paper introduces the two-class cross-entropy loss function (BCE loss) add L_N as a new loss function to avoid training instability.

Dice loss introduced in the V-Net paper, Dice loss is used to calculate the overlap between the prediction and ground-truth classes. The Dice coefficient (value range [0, 1]) is shown in

$$\mathrm{DIC} = \frac{2|\mathrm{GT} \cap P|}{|\mathrm{GT}| + |P|}, \tag{4}$$

where $|\mathrm{GT} \cap P|$, intersection between real semantic map and prediction map; $|\mathrm{GT}|$ and $|P|$, numbers of elements of GT and P, respectively. Factor 2 in the numerator is because the denominator had the repeated calculation of common elements between GT and P. Its formula is equivalent to the intersection ratio of the prediction result area and the ground truth area, so it calculates loss by taking all pixels of a category as a whole. Because Dice loss directly uses the segmentation effect evaluation index as loss to supervise the network and it also ignores a large number of background pixels when calculating Intersection over Union (IoU), which solves the problem of imbalance of positive and negative samples, so the convergence speed is fast.

Our goal was to maximize the overlap between the predicted real class and the base truth class (i.e., maximize the Dice coefficient). Therefore, we usually minimize $(1 - D)$ to achieve the same goal because most machine-learning libraries provide only the minimized option, and the expanded form is shown in

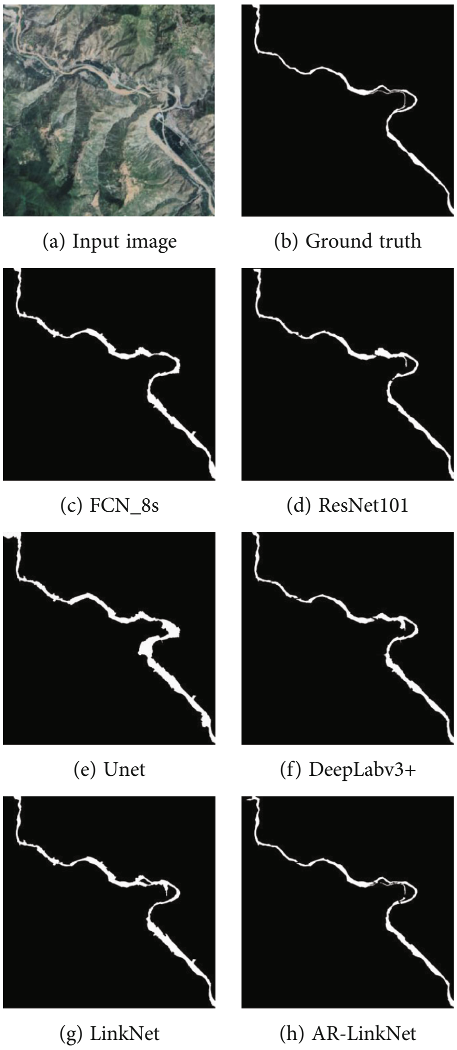
$$\mathrm{DiceLoss} = 1 - \frac{2|\mathrm{GT} \cap P|}{|\mathrm{GT}| + |P|}. \tag{5}$$

Entropy is used to measure the chaos of a system and represents the sum of the information in the system; the larger the entropy value, the greater the uncertainty of the system. The river extraction problem can be regarded as a binary classification problem, so the binary classification cross-entropy loss function can well represent the stability of the system.

The expanded form of BCE loss is shown in

$$\mathrm{BCELoss}(\mathrm{GT}, P) = -\sum_{i=1}^{W}\sum_{j=1}^{H}\left[\mathrm{GT}_{ij} \cdot \log P_{ij} + \left(1 - \mathrm{GT}_{ij}\right) \cdot \log\left(1 - P_{ij}\right)\right], \tag{6}$$

where GT is corresponding position label on semantic-segmentation map, P is the probability value predicted by the network, and H and W are the input-image height and width, respectively.



(a) Input image (b) Ground truth

(c) FCN_8s (d) ResNet101

(e) Unet (f) DeepLabv3+

(g) LinkNet (h) AR-LinkNet

Figure 16: Comparison results of each network extraction (1).

The loss function used the superposition of Dice and BCE losses, as shown in

$$L_N = \text{BCELoss} + \text{DiceLoss}. \tag{7}$$

4. Experiment Result Analysis

4.1. Evaluation Index. In this paper, we used recall, pixel accuracy (Acc), $F1$ score, and Intersection over Union (IoU) as evaluation indicators. These are the most commonly used indicators for evaluating river-extraction results. River extraction is considered an application of semantic segmentation; river pixels are set to 1, and other types of pixels are set to 0. Predictions are divided into four types: false negative (FN), false positive (FP), true negative (TN), and true positive (TP). The confusion matrix is shown in Figure 12. The first T/F indicates the prediction is right or wrong, and the second P/N indicates the prediction result. FN represents the number of river pixels that were mispredicted as other types. FP represents the number of other types of pixels that were mispredicted as rivers. TN represents the number of other types of pixels that were correctly predicted. TP represents the number of river pixels correctly predicted, which is consistent with the true value.

Recall definition is shown in

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

The definition of precision is shown in

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{9}$$

The definition of $F1$ score is shown in

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{10}$$

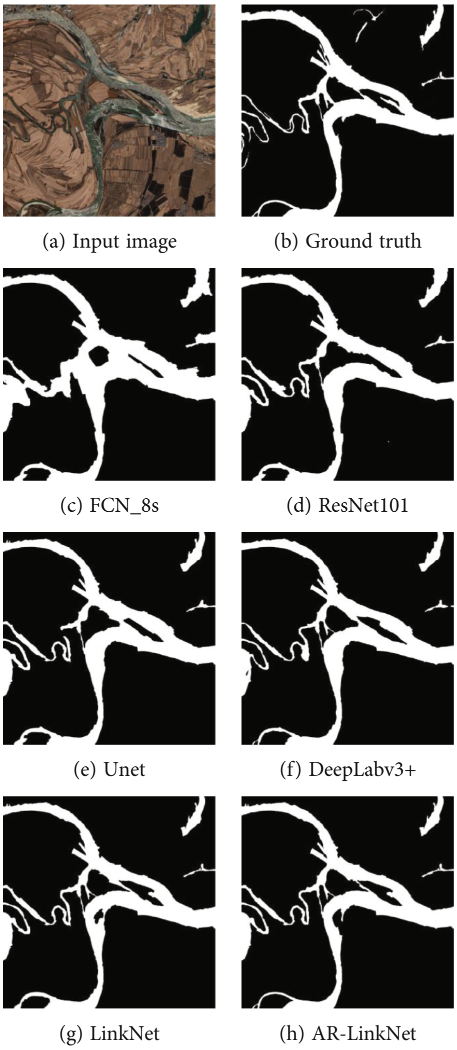

(a) Input image
(b) Ground truth
(c) FCN_8s
(d) ResNet101
(e) Unet
(f) DeepLabv3+
(g) LinkNet
(h) AR-LinkNet

FIGURE 17: Comparison results of each network extraction (2).

The definition of IoU is shown in

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP}}. \tag{11}$$

4.2. Pretreatment. Due to the difference of photographing time, different remote sensing images have light changes, and the ground cover in the image has great differences. A more aggressive color-amplification method was therefore used. Because remote sensing image shooting is in bird-eye view, most objects can remain semantically unchanged after being zoomed in and out. Because remote sensing images are isotropic, there is no such thing as direction difference, and the purpose of adding data is achieved by spinning the image. Through spatial geometric inversion and chromatic transformation, the acquired images were enlarged 9 times to 900 pieces. The schematic is shown in Figure 13. The test data collected 20 water-containing images with a side length of 10 kilometers in cold and arid regions and included rivers of various scales, including mountain shadows, road disturbances, image stitching, and river color changes.

4.3. Experiment Settings. The experimental hardware platform in this paper is Windows10 operating system, NVIDIA GeForce RTX 3080 (16G) GPU, i9-11980HK CPU, 64 GB memory. The network is built under the Pytorch framework, and other super parameters are shown in Table 1.

4.4. Experimental Results. In order to verify the effectiveness of the visual activation function FReLU, ResNet-101 was used as the pretraining model to migrate to the LinkNet network, and then, the original nonlinear activation function ReLU in the network was replaced with the visual activation function FReLU. The experimental results of the Intersection over Union and precision are shown in Table 2.

According to the experimental results in Table 2, compared with the original LinkNet network, the IoU value of the improved AR-LinkNet network is increased by 0.035, the precision value is increased by 0.022, and the

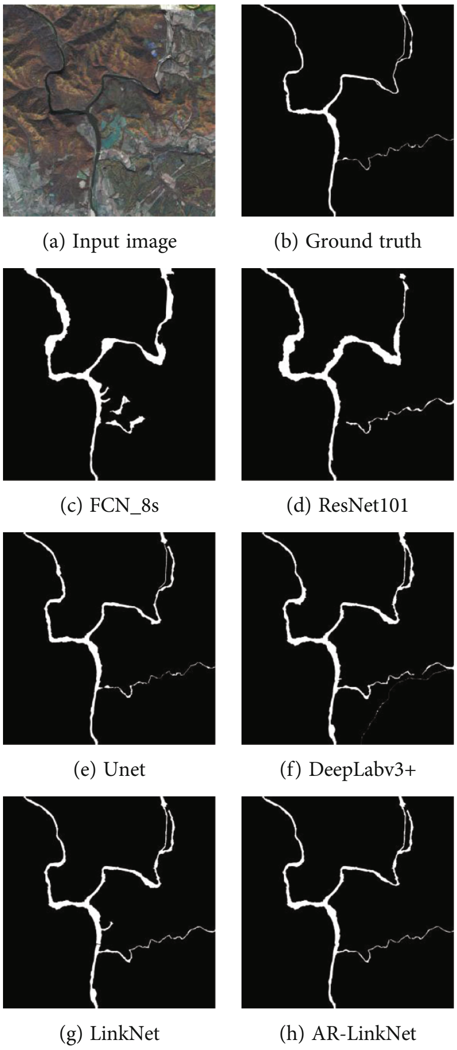
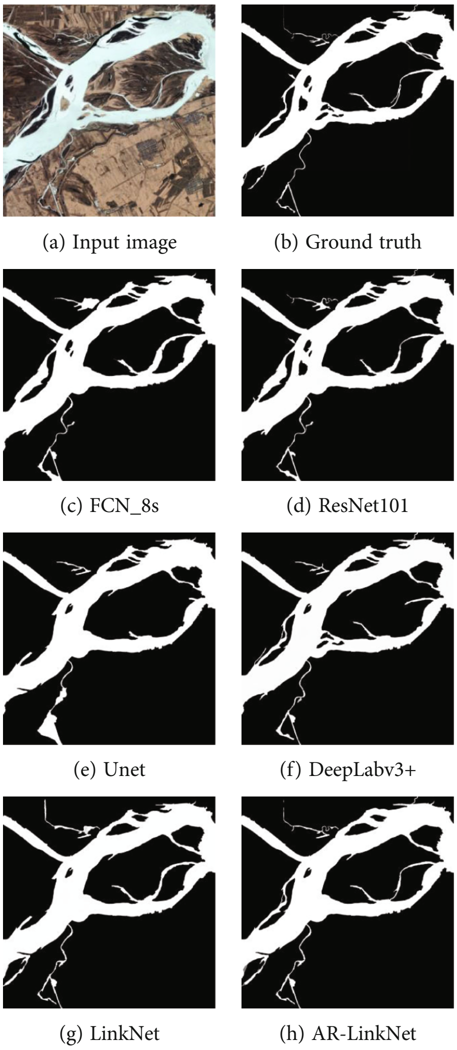

(a) Input image (b) Ground truth (c) FCN_8s (d) ResNet101 (e) Unet (f) DeepLabv3+ (g) LinkNet (h) AR-LinkNet

Figure 18: Comparison results of each network extraction (3).

segmentation accuracy of most objects is improved, which proves that the FReLU activation function can optimize the network effectiveness. The visual comparison diagram is shown in Figure 14. It can be seen from the recognition results that in the test image, the LinkNet network can segment wide rivers quite clearly, but there is an obvious lack of segmentation of small river targets that are occluded by mountain shadows. It can be seen that the AR-LinkNet network after using the FReLU activation function for accuracy compensation has better semantic segmentation ability for small river targets. In the second line of test images, the LinkNet network segmentation results have the problem of blurred target boundaries and wrongly segmented river targets. By comparison, it can be seen that the AR-LinkNet network segmentation results after using the FReLU activation function for accuracy compensation have higher boundary accuracy. It can effectively reduce the problem of misidentifying segmentation targets.

Next, the attention model is first verified. Here, the effect of adding the residual attention model to AR-LinkNet and the effect of the order of the two residual attention models on the network are mainly tested. The test results are shown in Table 3. Compared with the AR-LinkNet network without the attention module, when only adding the residual channel attention model, the accuracy rate is increased by 0.012 percentage points, and the IoU is increased by 0.014 percentage points, which is better than adding only residual space attention. The 0.006 and 0.008 improvements of the force model are relatively significant, indicating that the residual channel attention model has a greater impact on the accuracy. In terms of test time, the average test time increases by 0.5 ms due to the addition of the two models, and the residual channel attention model takes longer, mainly because of its larger amount of computation. When both attention models are added, we compare the three sequential structures of the two attention models in parallel, space-first-channel and channel-first-space, and the three structures are better than just adding. The residual channel or residual space model has a certain improvement, which also shows that the effect of using the two attention models together is better.



(a) Input image (b) Ground truth

(c) FCN_8s (d) ResNet101

(e) Unet (f) DeepLabv3+

(g) LinkNet (h) AR-LinkNet

Figure 19: Comparison results of each network extraction (4).

However, in terms of average test time, the parallel model is slightly faster, and the space-first-channel model is slightly slower, but both are more than 1 ms slower than the original network. After comprehensive comparison, the segmentation effect of the residual channel attention model and the residual spatial attention model is the best.

To prove the effectiveness of the AR-LinkNet network, without the addition of DenseASPP, we made an experiment comparison with popular semantic-segmentation networks. The selected compared networks were FCN_8s, ResNet101, DeepLabv3+, Unet, and LinkNet. The parameter initialization methods all use the MSRA initialization method [36]; considering only the number of inputs, the weight initialization follows a Gaussian distribution with a mean value of 0 and a variance of $2/n$ (n is the number of inputs). In this experiment, the loss function of all networks uses Dice loss. The change curve of the precision in the network test is shown in Figure 15, and the evaluation indicators are shown in Table 4.

From the experiment results in Figure 15 and Table 4, AR-LinkNet had certain advantages compared with current mainstream image semantic-segmentation methods, which could quickly converge during training, and compared with FCN_8s, Resnet101, Unet, DeepLabv3+, and the original LinkNet network, the precision has been improved by 22.3%, 11.9%, 7.1%, 4.7%, and 3.5%, and the IoU has increased by 21.7%, 15.6%, 12.6%, 7%, and 4.1%. Figures 16–19 show the segmentation effect in various situations of the dataset. Figures 10(a)–10(h) show the input images, ground truth, FCN_8s, ResNet101, Unet, DeepLabv3+, LinkNet, and AR-LinkNet network extraction results, respectively. FCN_8s caused many mountain features, such as mountain shadows, roads, and vegetation, to be similar to spectral features of the water body. The extraction results were messy and unsatisfactory, and it was prone to interruption in some small rivers. ResNet101 and DeepLabv3+ were greatly improved, but it also incorrectly indicated shadows, roads, and vegetation; however, it had a lower false-lift rate than that of FCN_8s. The effect of the river extracted by the LinkNet network is better, and its object-based extraction method eliminates a lot of noise interference, but the details of the riverside are not handled

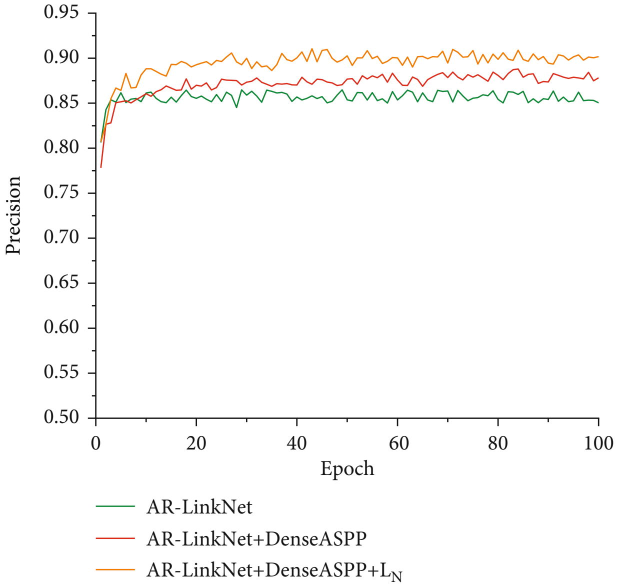


Figure 20: Precision change curve in network test.

Table 5: Network evaluation indicators.

Network	Recall	$F1$ score	IoU	Precision
AR-LinkNet	0.839	0.851	0.645	0.864
AR-LinkNet+DenseASPP	0.871	0.874	0.682	0.877
AR-LinkNet+DenseASPP+L_N	0.894	0.897	0.713	0.901

well, and the shadow effect is unavoidable. Compared with the comparison network, the river information extracted by AR-LinkNet is less affected by roads and vegetation and does not completely avoid shadow occlusion, but the detail segmentation of river edges is greatly improved.

To prove the effectiveness of DenseASPP and L_N loss functions on the river extraction of remote sensing images, we compared AR-LinkNet, AR-LinkNet+DenseASPP, and AR-LinkNet+DenseASPP+L_N. The precision changes are shown in Figure 20, and various evaluation indicators are shown in Table 5.

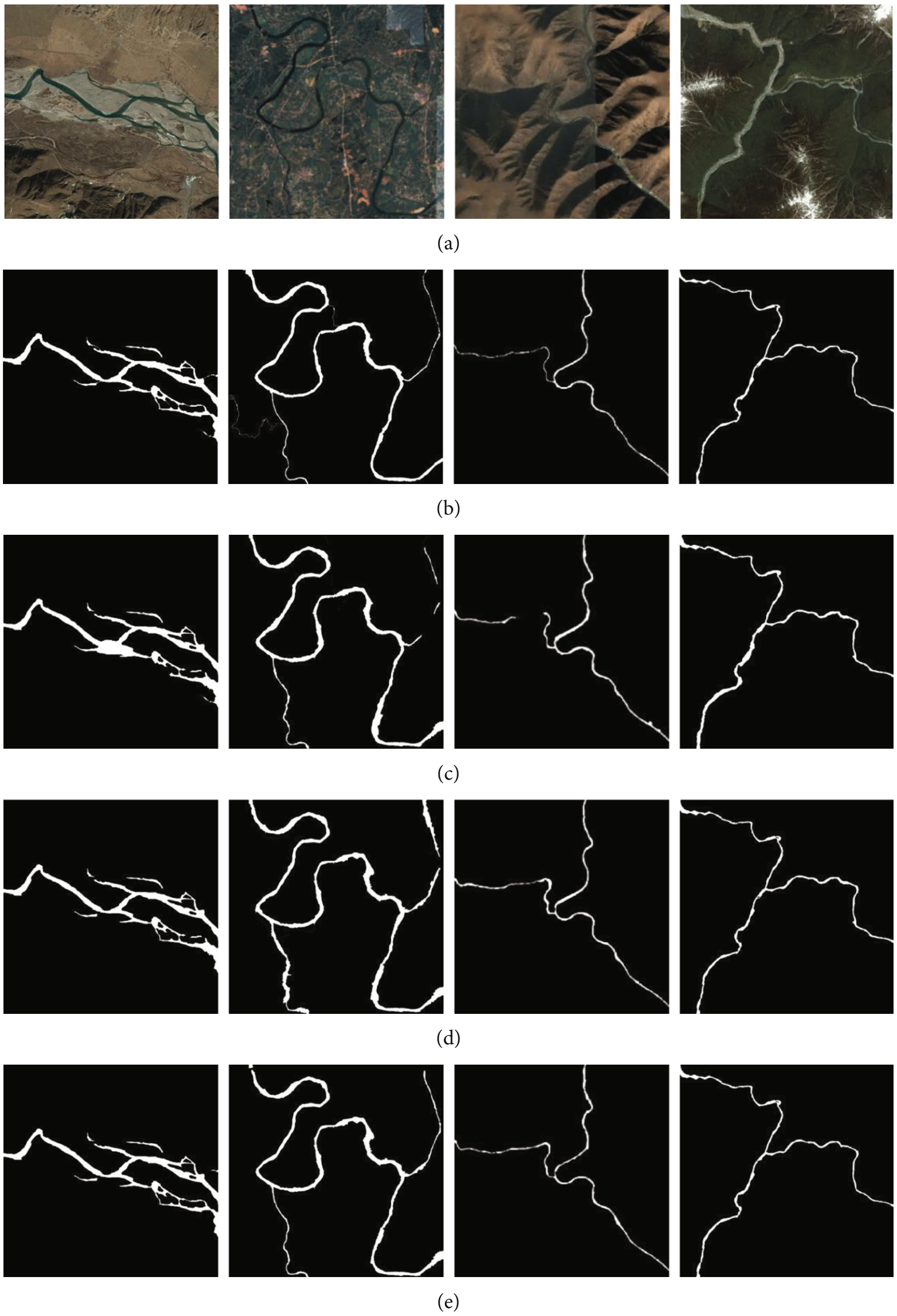
As shown in Figure 20 and Table 5, after adding DenseASPP, the precision is increased by 1.3%, and the IoU is increased by 3.7%. After using the new loss function, the precision and the IoU are increased by 2.4% and 3.1%, respectively. It can be seen from Figure 20 that DenseASPP can effectively improve the accuracy of river extraction without affecting the convergence of the entire network, which proves the effectiveness of DenseASPP and L_N.

Figure 21 shows the river extraction effect of each network on the dataset. The continuity of river extraction is improved, small rivers are more coherent, and edge detail extraction is more accurate. Using L_N as the loss function can effectively optimize the water edge details, and the river network extraction is complete.

The original remote sensing image river dataset consists of 100 pairs of training images. In order to improve the generalization ability of the network and avoid overfitting, we performed data enhancement through operations such as rotation, reduction, enlargement, mirroring, and chromaticity transformation. Compared with the 100 pairs of training samples before data enhancement, the training set is enlarged by 9 times and consists of 900 pairs of training samples after data enhancement. The effects of different training scales, including precision rate, recall rate, $F1$ value, and IoU are in the internal evaluation index shown in Figure 22. Obviously, with the expansion of the training scale, the four evaluation indicators have been improved by 2.5%, 1.3%, 1.9%, and 3.4%, respectively. Therefore, a larger training scale can improve the generalization ability of the network and avoid the network overfitting, especially for networks with many parameters.

5. Conclusion

Aiming at the problem of river identification in remote sensing images in cold and arid regions, an efficient identification method is proposed. Firstly, according to the characteristics of remote sensing images in cold and arid regions, a dataset of river identification in remote sensing images in cold and arid regions is made. Second, by combining transfer learning and deep neural network, ResNet-101 is migrated to LinkNet, and the FReLU function is used as the activation function to design and introduce an attention mechanism with residual structure to form an AR-LinkNet network. The network can completely restore the resolution of the image, while ensuring the continuity of river extraction. Combined with DenseASPP, the accuracy of river extraction is effectively improved; in the training process, the BCE loss function and the Dice loss function are added. The loss function effectively improves the image quality. The experimental results show that compared with the mainstream image semantic segmentation network, the AR-



(a)

(b)

(c)

(d)

(e)

Figure 21: Comparison results of multiple network extractions. Columns (a–e) show the input images, ground truth, AR-LinkNet, AR-LinkNet+DenseASPP, AR-LinkNet+DenseASPP+L_N, respectively.

LinkNet proposed in this paper has a greater performance improvement in river extraction. Combining AR-LinkNet with DenseASPP, the extracted river network is more coherent. L_N is used as the loss function, and the details of the river information extracted by the network after training are more abundant.

The following aspects will be explored in the subsequent research:

(1) Further improving results of network edge segmentation: the loss function used in this paper could ensure that the gradient of back propagation was more balanced for objects of different scales so that targets of small rivers could also be identified. However, this loss function did not clearly monitor the edges of the object. Therefore, there is still much potential for advancement on the edge segmentation of large rivers

(2) Weakly supervised learning for network training: the cost of manually labeling a large number of data is relatively high, so the weakly supervised learning

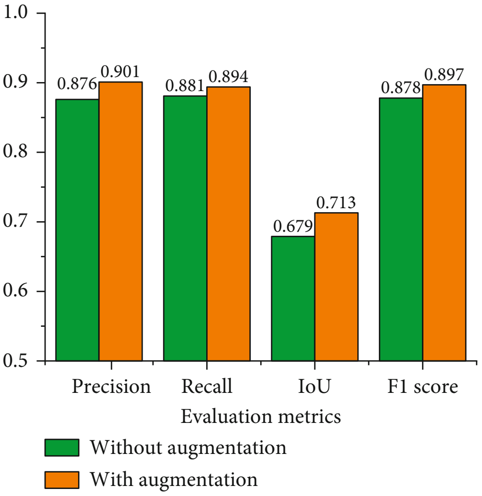


Figure 22: Effect of data augmentation on the precision, recall, $F1$ score, and IoU of our model.

method trains large numbers of unlabeled data with labeled data, which can reduce the cost of manual labeling. This could at present avoid bothering with few data

(3) The predicted semantic-segmentation graph was more like an annotation graph, and the forecast map was a kind of probability map. When dealing with some small rivers, discontinuities often occur. Some articles used the idea of adversarial training to make the prediction map and label map output by the network more similar. This idea can be used to solve the problem of identifying small rivers

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors stated that there are no conflicts of interest in the paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China, grant number 61861025, by the Intelligent Tunnel Supervision Robot Research Project (China Railway Research Institute (Scientific Research) Word 2020-KJ016-Z016-A2), by the Young Doctoral Fund Project of Higher Education Institutions in Gansu Province, China (2021QB-49), and by the Research Project of Higher Education Institutions in Gansu Province, China (2016A-018).

References

[1] Z. Lin and J. Lv, "Monitoring of natural ecological environment based on Sentinel 2A," in *2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, pp. 944–947, Nanchang, China, 2021.

[2] D. R. Charan, D. S. S. Teja, R. Subhashini, Y. B. Jinila, and G. M. Gandhi, "Convolutional neural network based water resource monitoring using satellite images," in *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, pp. 1261–1266, Coimbatore, India, 2020.

[3] J. Xin, Y. Yang, and S. Huang, "Study on regional drought monitoring based on multi-sources data in China," in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, pp. 6898–6901, Waikoloa, HI, USA, 2020.

[4] E. M. Moh Aung and T. Tint, "Ayeyarwady River regions detection and extraction system from Google Earth imagery," in *2018 IEEE International Conference on Information Communication and Signal Processing (ICICSP)*, pp. 74–78, Singapore, 2018.

[5] 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.

[6] Q. Yin, J. Cheng, F. Zhang, Y. Zhou, L. Shao, and W. Hong, "Interpretable POLSAR image classification based on adaptive-dimension feature space decision tree," *IEEE Access*, vol. 8, pp. 173826–173837, 2020.

[7] X. Huang, J. Liu, W. Zhu, C. Atzberger, and Q. Liu, "The optimal threshold and vegetation index time series for retrieving crop phenology based on a modified dynamic threshold method," *Remote Sensing*, vol. 11, no. 23, p. 2725, 2019.

[8] X. Luo, H. Xie, and X. Tong, "A water extraction method based on airborne hyperspectral images in highly complex urban area," in *2017 International Workshop on Remote Sensing with Intelligent Processing (RSIP)*, pp. 1–4, Shanghai, China, 2017.

[9] J. Wang, J. Ding, and C. Zhang, "Method of water information extraction by improved SWI based on GF-1 satellite image," *Remote Sensing for Land & Resources*, vol. 29, no. 1, pp. 29–35, 2017.

[10] R. Dong, D. Xu, L. Jiao, J. Zhao, and J. An, "A fast deep perception network for remote sensing scene classification," *Remote Sensing*, vol. 12, no. 4, p. 729, 2020.

[11] Z. Miao, K. Fu, H. Sun, X. Sun, and M. Yan, "Automatic waterbody segmentation from high-resolution satellite images via deep networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 4, pp. 602–606, 2018.

[12] X. U. E. Yuan, L. I. Dan, W. U. Baosheng, and F. U. Xudong, "Automatic extraction of small mountain river information and width based on China-made GF-1 satellites remote sensing images," *Bulletin of Surveying and Mapping*, vol. 3, pp. 12–16, 2020.

[13] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, "Semantic segmentation with context encoding and multi-path decoding," *IEEE Transactions on Image Processing*, vol. 29, pp. 3520–3533, 2020.

[14] M. J. Hughes and R. Kennedy, "High-quality cloud masking of Landsat 8 imagery using convolutional neural networks," *Remote Sensing*, vol. 11, no. 21, p. 2591, 2019.

[15] J. Y. Chiao, K. Y. Chen, and Y. K. Liao, "Detection and classification the breast tumors using mask R-CNN on sonograms," *Medicine*, vol. 98, no. 19, article e15200, 2019.

[16] Z. Liu, C. L. P. Chen, S. Feng, Q. Feng, and T. Zhang, "Stacked broad learning system: from incremental flatted structure to deep model," *In IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 1, pp. 209–222, 2021.

[17] X. Han, C. Nguyen, S. You, and J. Lu, "Single image water hazard detection using FCN with reflection attention units," *European Conference on Computer Vision*, vol. 11210, pp. 105–121, 2018.

[18] S. Wang, W. Chen, S. M. Xie, G. Azzari, and D. B. Lobell, "Weakly supervised deep learning for segmentation of remote sensing imagery," *Remote Sensing*, vol. 12, no. 2, p. 207, 2020.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *In European conference on computer vision*, pp. 630–645, Cham, 2016.

[20] G. Wang, M. Wu, X. Wei, and H. Song, "Water identification from high-resolution remote sensing images based on multidimensional densely connected convolutional neural networks," *Remote Sensing*, vol. 12, no. 5, p. 795, 2020.

[21] A. Chaurasia and E. Culurciello, "LinkNet: exploiting encoder representations for efficient semantic segmentation," in *2017 IEEE Visual Communications and Image Processing (VCIP)*, pp. 1–4, St. Petersburg, FL, USA, 2017.

[22] L. Zhou, C. Zhang, and W. Ming, "D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 182–186, Salt Lake City, USA, 2018.

[23] H. Fan, Q. Wei, D. Q. Shu, Y. Li, L. Zhang, and C. D. Yang, "An improved Deeplab based model for extracting cultivated land information from high definition remote sensing images," in *2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP)*, pp. 1–6, Chongqing, China, 2019.

[24] M. N. Mahmud, M. K. Osman, A. P. Ismail, F. Ahmad, K. A. Ahmad, and A. Ibrahim, "Road image segmentation using unmanned aerial vehicle images and DeepLab V3+ semantic segmentation model," in *2021 11th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, pp. 176–181, Penang, Malaysia, 2021.

[25] Z. Yang, X. Peng, Z. Yin, and Z. Yang, "Deeplab v3 plus-net for image semantic segmentation with channel compression," in *2020 IEEE 20th International Conference on Communication Technology (ICCT)*, pp. 1320–1324, Nanning, China, 2020.

[26] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3684–3692, Salt Lake City, USA, 2018.

[27] W. Xia, C. Ma, J. Liu et al., "High-resolution remote sensing imagery classification of imbalanced data using multistage sampling method and deep neural networks," *Remote Sensing*, vol. 11, no. 21, p. 2523, 2019.

[28] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. B. Ayed, "Hyper dense-net: a hyper-densely connected CNN for multi-modal image segmentation," *IEEE Transactions on Medical Imaging*, vol. 38, no. 5, pp. 1116–1126, 2019.

[29] H. He, D. Yang, S. Wang, S. Wang, and Y. Li, "Road extraction by using atrous spatial pyramid pooling integrated encoder-decoder network and structural similarity loss," *Remote Sensing*, vol. 11, no. 9, p. 1015, 2019.

[30] R. Li, S. Zheng, C. Zhang et al., "Multi-attention-network for semantic segmentation of fine resolution remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2020.

[31] Z. Zhao, K. Chen, and S. Yamane, "CBAM-Unet++: easier to find the target with the attention module "CBAM"," in *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)*, vol. 655, Kyoto, Japan, 2021no. 657.

[32] A. Creswell, A. Kai, and A. Anil, "On denoising autoencoders trained to minimise binary cross-entropy," 2017, https://arxiv.org/abs/1708.08487.

[33] M. Fausto, N. Nassir, and A. Seyed, "V-Net: fully convolutional neural networks for volumetric medical image segmentation," 2016, https://arxiv.org/abs/1606.04797.

[34] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2020.

[35] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," in *In Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, Munich, Germany, 2018.

[36] N. Ma, X. Zhang, and J. Sun, "Funnel activation for visual recognition," in *In European Conference on Computer Vision*, pp. 351–368, Cham, 2020.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: surpassing human-level performance on ImageNet classification," in *In Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, Santiago, Chile, 2015.