Hindawi
Journal of Sensors
Volume 2022, Article ID 6391238, 11 pages
https://doi.org/10.1155/2022/6391238

Research Article

BSIRNet: A Road Extraction Network with Bidirectional Spatial Information Reasoning

Hai Tan[1], **Hao Xu**[2,3] **and Jiguang Dai**[2,3]

[1]*Land Satellite Remote Sensing Application Center, Ministry of Natural Resources, Beijing 100048, China*
[2]*School of Geomatics, Liaoning Technical University, Fuxin 12300, China*
[3]*Institute of Spatiotemporal Transportation Data, Liaoning Technical University, Fuxin 12300, China*

Correspondence should be addressed to Hai Tan; tanh@lasac.cn

Received 22 September 2021; Accepted 17 December 2021; Published 12 January 2022

Academic Editor: Wei Zhang

Automatic extraction of road information from remote sensing images is widely used in many fields, such as urban planning and automatic navigation. However, due to interference from noise and occlusion, the existing road extraction methods can easily lead to road discontinuity. To solve this problem, a road extraction network with bidirectional spatial information reasoning (BSIRNet) is proposed, in which neighbourhood feature fusion is used to capture spatial context dependencies and expand the receptive field, and an information processing unit with a recurrent neural network structure is used to capture channel dependencies. BSIRNet enhances the connectivity of road information through spatial information reasoning. Using the public Massachusetts road dataset and Wuhan University road dataset, the superiority of the proposed method is verified by comparing its results with those of other models.

1. Introduction

Roads play an important role in urban planning, traffic navigation, map updating, and other fields [1]. With the rapid development of remote sensing satellites and sensors, it is becoming increasingly easy to collect very high-resolution (VHR) satellite imagery, which can provide sufficient data sources for road extraction. Therefore, extracting road information from VHR satellite imagery has become a popular topic of research. To date, researchers have developed many different road extraction methods [2], which can be generally divided into traditional methods and deep learning methods.

Traditional road extraction methods rely on road image features and the construction of a theoretical model [3]. For example, Song and Civco used a shape index and density features to extract road features [4], Valero et al. proposed the use of directional morphological operators that can flexibly fit straight and slightly curved structures for road extraction [5], and Dai et al. used a multiscale directional histogram and sector descriptor to extract road information through heuristic tracking [6]. However, the image features used to extract roads in these methods are manually designed and lack an automatic learning process. Consequently, traditional road extraction methods have the disadvantages of low automation, complex operation, and high time consumption.

Deep learning methods rely on a hierarchical feature expression framework to mathematically model specific problems in the real world and then use the resulting models to solve similar problems [7]. Different from traditional road extraction methods, deep learning methods have the characteristics of high automation and a strong learning ability [8], allowing them to better handle occlusion and shadows on roads. For example, a road structure refined convolutional neural network (CNN) was proposed by Wei et al. [9], which incorporates the geometric information of the road structure in the network learning process; Kestur et al. used a U-shaped fully convolutional network (U-FCN) that combines shallow fine-grained layers with a final-score layer to extract roads [10]; and Zhang et al. proposed a deep residual U-Net model [11], which combines the advantages of residual learning and U-Net [12, 13], for the road extraction task.

Although deep learning methods have achieved good results in automatic road extraction, these methods still often produce discontinuous road segments, which cause great difficulties in practical applications. There are two main reasons for this: (1) the occlusions and shadows caused by trees, buildings, etc., may cause a deep learning model to fail to correctly capture the information of occluded and shadowed roads. (2) The texture of a road may be very similar to that of the surrounding ground features, causing the model to be unable to extract clear road boundaries and locations. Both of these situations will lead to incomplete road extraction results, resulting in discontinuous roads.

At present, to address the discontinuity problem in road extraction, researchers have proposed various network models to directly improve the road extraction results. For example, He et al. proposed a road extraction network that relies on an encoder–decoder structure [14]. In the decoder component of this network, the spatial resolution of the feature maps is gradually restored by upsampling, but the details of the road edges are also lost. To prevent the loss of road edge details, Zhou et al. proposed a boundary and topologically-aware neural network (BT-RoadNet) [15]. This network extracts both rough features and fine features and then fuses them to improve the road extraction results. However, each channel contains a specific semantic feature, and BT-RoadNet ignores the relationships between the channels. Lu et al. proposed a globally aware road detection network with multiscale residual learning (GAMSNet) [16]. This network uses global average pooling to process channels and then uses a fully connected layer to establish the relationships between channels to improve the accuracy of road extraction. Although this network considers the relationships between channels, a large amount of road information will inevitably be lost when the channels are subjected to global average pooling. The above semantic segmentation network undeniably improves the accuracy of road extraction to a certain extent and improves the overall effectiveness on this task. However, the above network will still produce road discontinuities during the road extraction process.

To better solve the problem of road discontinuity, a road extraction network with bidirectional spatial information reasoning (BSIRNet) is proposed in this paper. In BSIRNet, a spatial reasoning perception module (SRPM) is established to capture spatial context dependence, a channel reasoning perception module (CRPM) is established to capture interchannel dependencies, and a multiscale skip connection structure is used to capture more semantic information.

The major contributions of this research are summarized as follows:

(1) A road extraction network with bidirectional spatial information reasoning (BSIRNet) is proposed. BSIRNet captures the dependencies of the road information in the spatial dimension and the channel dimension simultaneously. Moreover, BSIRNet extracts multiscale features of roads and integrates them to capture more road information. The BSIRNet method proposed in this research enhances the information reasoning ability applied in the road extraction process to solve the problem of discontinuous road extraction

(2) A spatial reasoning perception module (SRPM) and a channel reasoning perception module (CRPM) are proposed. The SRPM is aimed at capturing spatial context dependence such that at each location, the characteristics of the neighbourhood can be adaptively inferred to expand the receptive field. The CRPM is aimed at establishing the relationships between channels and capturing the dependencies between channels. Together, the SRPM and CRPM can solve the problem that road information cannot be captured due to occlusion and shadows

(3) A multiscale skip connection structure is used to extract multiscale semantic features and perform feature fusion processing. Feature maps of different scales contain different road information. A low-level feature map captures rich spatial information and can highlight the road boundaries, while a high-level semantic feature map reflects the road location information [17]. Our multiscale skip connection structure can solve the problem of unclear road boundaries caused by similar textures of roads and background features

The BSIRNet model proposed in this paper was verified on road datasets from Massachusetts and Wuhan University. The experimental results show that this method is superior to the existing deep-learning-based road extraction methods.

The rest of the paper is structured as follows. Section 2 introduces related network architectures. Section 3 describes the details of BSIRNet. Section 4 describes the datasets used in the experiments, the experimental setup, and the experimental results and presents a comprehensive analysis. Section 5 discusses the methods and advantages of BSIRNet. Finally, we summarize our conclusions in Section 6.

2. Related Architectures

As shown in Figure 1, the basic network used in BSIRNet is DeepLabV3+ [18], which, in turn, uses the improved Xception network as its backbone [19]. In the entry flow of the improved Xception network, all max pooling layers are changed to depthwise separable convolutions with stride = 2. In the middle flow, the residual blocks are repeated 16 times instead of 8 times. The atrous spatial pyramid pooling (ASPP) unit includes 5 different convolution operations, which extract different feature maps, and concatenation is then applied for multiscale feature fusion. DeepLabV3+ uses depthwise separable convolution to reduce the number of parameters to improve its computational efficiency.

A spatial information inference structure (SIIS) enables multidirectional message passing between pixels when it is integrated to a typical semantic segmentation framework [20]. Since the spatial information could be propagated

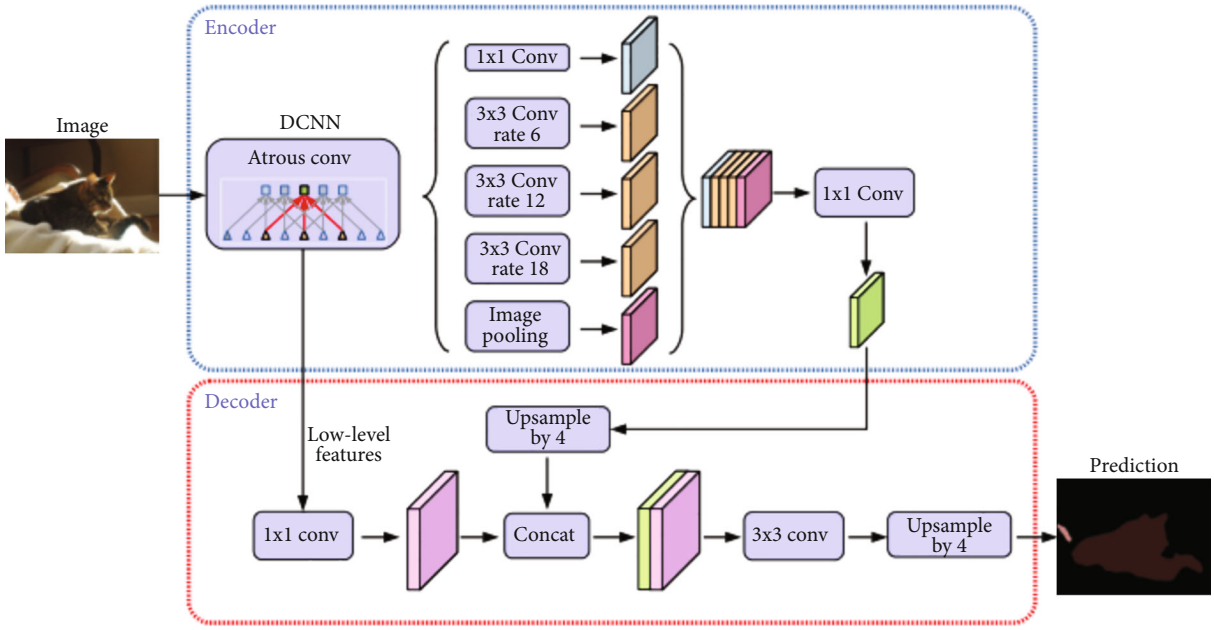

Figure 1: The architecture of DeepLabV3+.

and reinforced via interlayer propagation, SIIS can learn both the local visual characteristics of the road and the global spatial structure information. We also take inspiration from the advantages of SIIS. As shown in Figure 2, in SIIS, each feature map is divided into blocks by rows or columns, and information processing units with a recurrent neural network are sequentially applied to establish the semantic context [21].

Although DeepLabV3+ is an efficient semantic segmentation network, its convolutions can process only one local neighbourhood at a time, and it cannot effectively capture the long-range dependencies of the road information in the road extraction task. Similarly, although SIIS can establish contextual semantic relations, its convolutions can handle only one local neighbourhood at a time, and SIIS does not consider the relationships between channels.

Consequently, DeepLabV3+ and SIIS will both result in road discontinuities to varying degrees. To address the shortcomings of these existing architectures, we propose our road extraction network with bidirectional spatial information reasoning (BSIRNet).

3. Method

Figure 3 shows the overall flow chart of BSIRNet. BSIRNet is based on DeepLabV3+ and consists of Xception, ASPP, SIIS, SRPM, and CRPM components and a multiscale skip connection structure. The SRPM is used to capture the spatial context dependence, the CRPM is used to capture the dependence between channels, and the multiscale skip connection structure is used to capture more semantic information. The detailed architecture of BSIRNet is described in the following.

As shown in Figure 3, four outputs are generated from the input image after the deep convolutional neural network (DCNN): three low-level feature maps of different scales and one high-level semantic feature map. The three low-level feature maps of different scales correspond to the output_stride of the entry flow in Xception, which takes values of 4x, 8x, and 16x. These low-level feature maps represent the extracted spatial information, boundary information and location information of the roads in the input image.

The high-level semantic feature map is the output of the exit flow and serves as the input to the ASPP structure. The high-level semantic feature map undergoes five different convolution operations in the ASPP structure to yield outputs corresponding to five different scales. The five different convolution operations include a 1×1 convolution, three dilated convolutions with different dilation rates, and an image pooling operation. Among the three dilated convolutions with different dilation rates, the 3×3 dilated convolution with a dilation rate of 6 yield features with a smaller receptive field and clearer boundaries (fine features). In contrast, the 3×3 dilated convolution with a dilation rate of 18 yield features with a larger receptive field and blurred boundaries (rough features). The image pooling operation consists of global average pooling of the input features, followed by upsampling to the original size.

In summary, the road boundaries in the fine features are clear, but the receptive field is small. The road boundaries in the rough features are fuzzy, but the receptive field is large. If the fine feature pixels can be associated with the neighbourhood pixels corresponding to the rough features, not only can a reasoning relationship be established between each location and its neighbourhood but the receptive field in each region can also be expanded.

3.1. Spatial Reasoning Perception Module (SRPM). Since the convolution operations of DeepLabV3+ can process only one local neighbourhood at a time, the spatial context

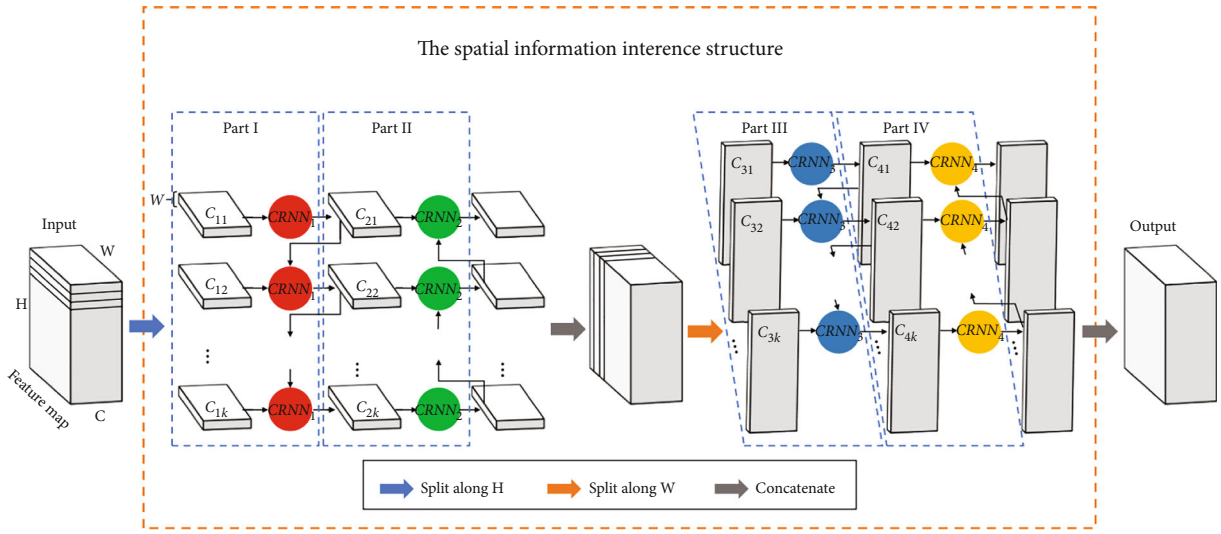

Figure 2: The architecture of the spatial information inference structure (SIIS).

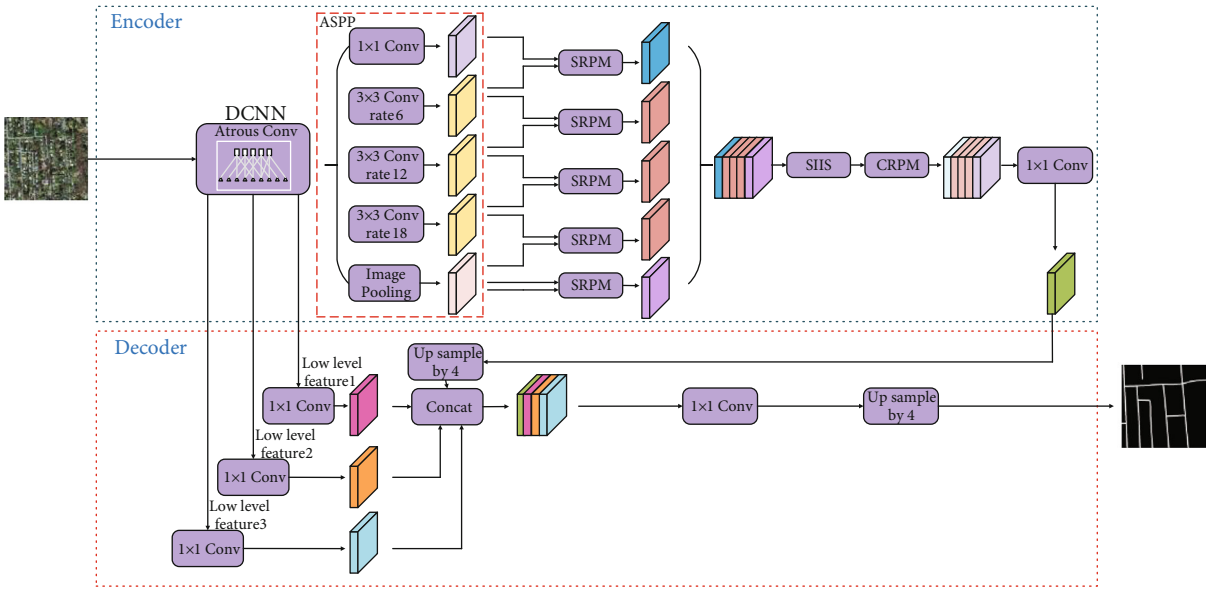

Figure 3: Overall flow chart of BSIRNet.

information obtained is limited, which prevents the network from capturing the long-range dependence of the road information. Therefore, we propose the SRPM to capture the spatial context dependence so that neighbourhood characteristics can be adaptively inferred for different locations, and the receptive field can be expanded.

In the SRPM, we use t different repeated paddings to expand the edge pixels of the feature map, such that each pixel of the feature map has a relative translation; thus, a set of feature map sequences $\{m_1, m_2, \cdots, m_t\}$ is obtained. We then apply concatenation to fuse the feature map sequence $\{m_1, m_2, \cdots, m_t\}$ to establish the inference relationship between each location and its corresponding neighbourhood and expand the receptive field of the feature map. The detailed structure of the SRPM is shown in Figure 4. The SRPM takes a rough feature map (pink) and a fine feature map (orange) as input. For the fine features, we use repeated padding (green) to fill in a layer of edge pixels in each of the four directions of the fine feature map to obtain f.

For the rough features, first use repeated padding to fill the rough feature map with one layer of edge pixels in the up and down directions and two layers of edge pixels on the right to obtain r_1. Second, we fill in two layers of edge pixels on the right and bottom sides of the rough feature map to obtain r_2. Then, we fill the rough feature map with one layer of edge pixels in the left and right directions and two layers of edge pixels on the bottom to obtain r_3. Analogously, we also fill in layers of edge pixels to obtain r_4, r_5, r_6, r_7, and r_8. Finally, concatenation is used to merge f, r_1, r_2, r_3

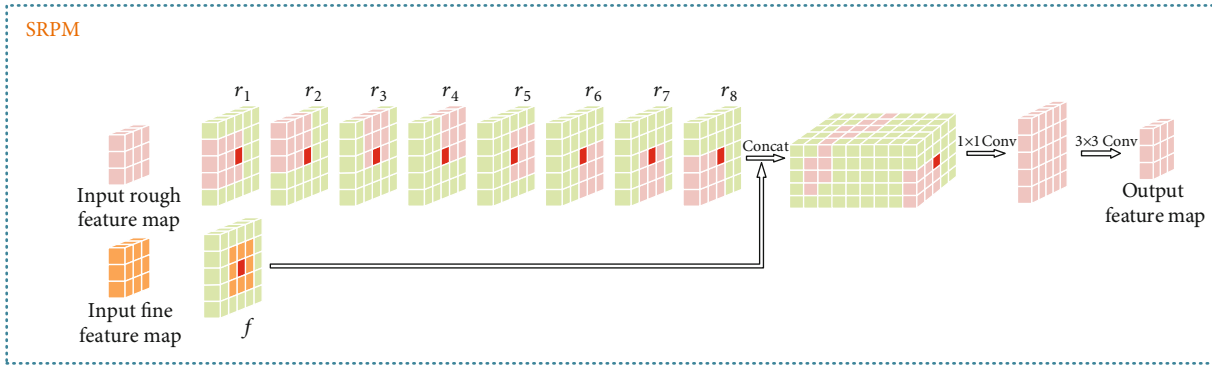

Figure 4: The architecture of the spatial reasoning perception module (SRPM).

, r_4, r_5, r_6, r_7 and r_8 obtained via the above operations. Let us consider the middle pixel of the fine feature map (red in f) as an example. When f is combined with r_1, a reasoning relationship is established between the middle pixel of the fine feature map and the pixel to the right in the corresponding rough feature map (red in r_1). Then, as the result continues to be merged with r_2, a reasoning relationship is established between the middle pixel of the fine feature map and the pixel to the lower right in the corresponding rough feature map (red in r_2). As the result continues to be merged with r_3, a reasoning relationship is established between the middle pixel of the fine feature and the pixel below in the corresponding rough feature map (red in r_3). Finally, r_4, r_5, r_6, r_7, and r_8 are sequentially merged to establish reasoning relationships between the middle pixel of the fine feature map and each pixel in its neighbourhood in the corresponding rough feature map. Because the receptive field of the rough feature map is larger, not only are reasoning relationships established between the fine feature map and the corresponding neighbours in the rough feature map but the receptive field of the fine features is also increased.

As shown in Figure 3, the SRPM is used to establish spatial reasoning perception relations for the five different scales of the ASPP output features. Since the output of the last image pooling operation does not have a corresponding rough feature map, the image pooling output can be used as both rough features and fine features to establish spatial reasoning perception relations.

3.2. Channel Reasoning Perception Module (CRPM). Each channel contains a specific semantic feature, but existing deep learning networks do not consider the relationships between channels when performing road extraction [16]. Therefore, we propose the CRPM to capture the interdependencies between channels to mitigate the discontinuity of roads. The proposal of the CRPM is also inspired by SIIS [20]. Although SIIS can efficiently establish semantic context relationships, it does not consider the transfer of information among recurrent neural network information processing units, easily leading to gradient disappearance or gradient explosion. Therefore, based on the above consideration, the number of skip connections in the CRPM is increased relative to SIIS to prevent gradient disappearance or gradient explosion.

The detailed structure of the CRPM is shown in Figure 5. As shown in part I of Figure 5, the input to the CRPM is a tensor with dimensions of $C \times H \times W$ consisting of SIIS output, where C, H, and W represent the numbers of channels, rows and columns, respectively. The tensor is first divided into k chunks along C, with the thickness of each chunk being $w = C/k$. Then, each chunk in the obtained sequence $S_1 = \{C_{11}, C_{12}, \cdots, C_{1k}\}$ is sent into CRNN_1 one by one. To prevent the gradient from disappearing or exploding, a skip connection is added after every four chunks. CRNN_1 is the first information processing unit of the CRPM. The main structure of a convolutional RNN (CRNN) unit is shown in Figure 6. This unit takes a three-dimensional tensor as input and produces an output in the same form to establish the reasoning relationships between channels. Specifically, the first chunk C_{11} is optimized by CRNN_1 to generate a new chunk C_{21} of equal size. When CRNN_1 optimizes the second chunk C_{12}, the most recent new chunk C_{21} will also be taken as input to provide the channel information. When the skip connection is optimized for the fifth chunk C_{15}, the generated chunks C_{21} and C_{24} are used as inputs to provide channel information. This process continues until the last chunk C_{k1} is updated, and during this process, the channel information is continuously transmitted downward.

In part II of Figure 5, the new chunks $C_{21}, C_{22}, \cdots, C_{2K}$ form a sequence $S_2 = \{C_{2K}, \cdots, C_{22}, C_{21}\}$ from bottom to top, which is then sent into CRNN_2 for optimization in the same way as in part I to produce k new chunks. These new chunks are then connected in the C dimension to form a complete tensor with dimensions of $C \times H \times W$, which is returned as the output tensor of the CRPM.

3.3. Multiscale Skip Connection Structure. Combining multiscale features is important for achieving accurate segmentation because feature maps of different scales contain different information. Low-level feature maps capture rich spatial information and can highlight the boundaries of roads, while high-level semantic feature maps reflect the location information of roads [17]. DeepLabV3+ operates at a single scale in the improved Xception. Therefore, in this work, multiscale features are extracted from Xception to extract more road information. Our multiscale skip connection structure fuses feature maps of different scales and then

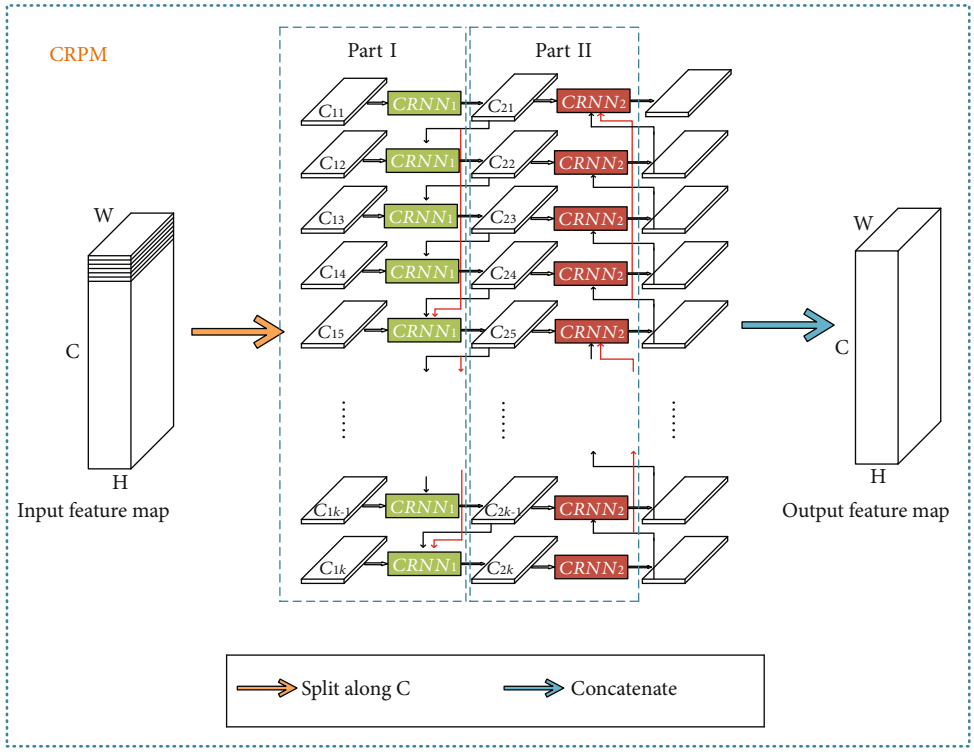

Figure 5: The architecture of the channel reasoning perception module (CRPM).

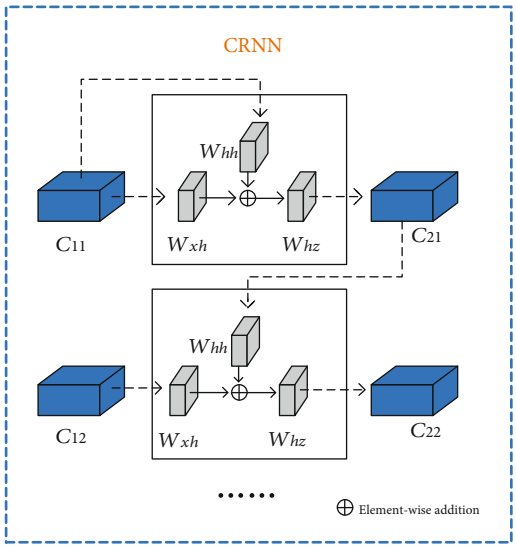

Figure 6: The architecture of a CRNN information processing unit.

learns a hierarchical representation of the aggregated multiscale feature maps.

In the multiscale skip connection structure, we use concatenation to fuse the three low-level feature maps of different scales with the processed high-level semantic feature map to capture more road information and improve the accuracy of road extraction. Then, 1×1 convolution and upsampling are used to process the fused multiscale features to obtain the final output.

4. Experiments and Analysis

4.1. General Details of the Experiments

4.1.1. Datasets. In our experiments, we use the Massachusetts road dataset and the Wuhan University road dataset. The Massachusetts road dataset consists of 1171 images, of which 1108 images are used for training, 14 images are used for validation, and 49 images are used for testing [22]. The size of each image is 1500×1500 pixels, and the resolution is 120 cm/pixel. The Wuhan University road dataset contains images from Boston and its surrounding cities in the United States, Birmingham in the United Kingdom, and Shanghai in China [16].

4.1.2. Data Preprocessing. For the Massachusetts road dataset and Wuhan University road dataset, since there are more background (nonroad) pixels in each image than road pixels, erroneously predicting road pixels to be background pixels is the main source of loss. Optimizing the semantic segmentation network can reduce the overall loss, but with the optimized semantic segmentation network, there is a high probability that uncertain pixels will be misidentified as background instead of as road pixels. To solve this problem, we adopt a simple and effective data preprocessing strategy known as the category ratio cropping (CRC) method [20].

As shown in Figure 7, we take an image I in the training set and its corresponding ground-truth label image L as an example. First, we use the same stride s and a $w \times w$

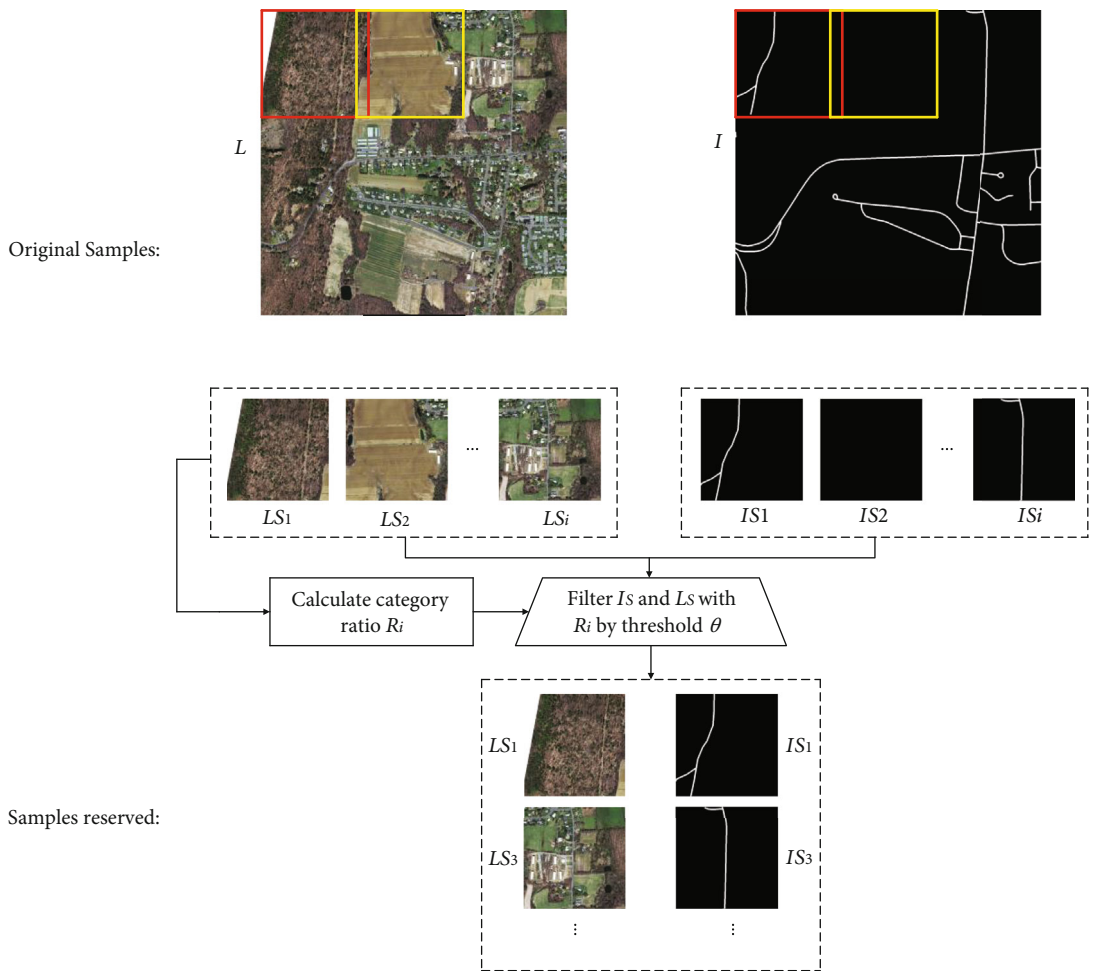

Figure 7: The process of applying the category ratio cropping (CRC) method to a typical sample from the Massachusetts road dataset.

cropping window to perform sliding cropping on $\{I, L\}$ to obtain a set of subimages and corresponding sublabels $\{I_{si}, L_{si}\}$, where $s = 494$ and $w = 512$. Second, L_{si} is used to calculate the ratio $R_i = [n1/ns, n2/ns, \cdots, nc/ns]$, where n_c denotes the number of pixels belonging to class C in L_{si} and n_s denotes the total number of pixels in L_{si}. Finally, the minimum value in R_i, min (R_i), is compared against a specified threshold θ. For pairs of I_{si} and L_{si}, only images with min (R_i) greater than the threshold θ are retained. We set the ratio threshold θ, which is a user-defined constant, to 0.01 [20].

After CRC data preprocessing, the imbalance in the numbers of road pixels and background pixels is effectively alleviated, thereby improving the efficiency of model training. In the end, we obtain 8988 images from the Massachusetts road dataset for training, 124 images for validation, and 386 images for testing. Similarly, we obtain 4568 images from the Wuhan University road dataset for training, 30 images for validation, and 127 images for testing.

4.1.3. Analysis of Experimental Parameters. To prevent excessively regular data from causing network overfitting or nonconvergence, we preprocess the training dataset using a data enhancement method that disrupts the order of the data. Second, considering that resizing will result in the loss of detailed image information, all images are used in their original size (512×512) to train the network. All models are trained with the same parameter settings and in the same environment. Specifically, we use the Adam optimizer for model training on a Windows 10 computer. The computer is equipped with an NVIDIA GeForce RTX 2080 Ti graphics card (with 11 GB of memory), allowing a batch size of 2 images. The learning rate for the Massachusetts road dataset is initially set to $1e-3$ and reduced by 0.85 every three epochs. The learning rate for the Wuhan University road dataset is initially set to $1e-4$ and reduced by 0.85 every three epochs. On these two datasets, the proposed network (BSIRNet) converges within only 50 epochs.

4.1.4. Evaluation Metrics. To evaluate the performance of a road extraction method, we adopt the following three evaluation metrics:

(1) The $F1$ score is an evaluation metric defined as the harmonic mean of the precision (P) and recall (R), and it is calculated as shown in

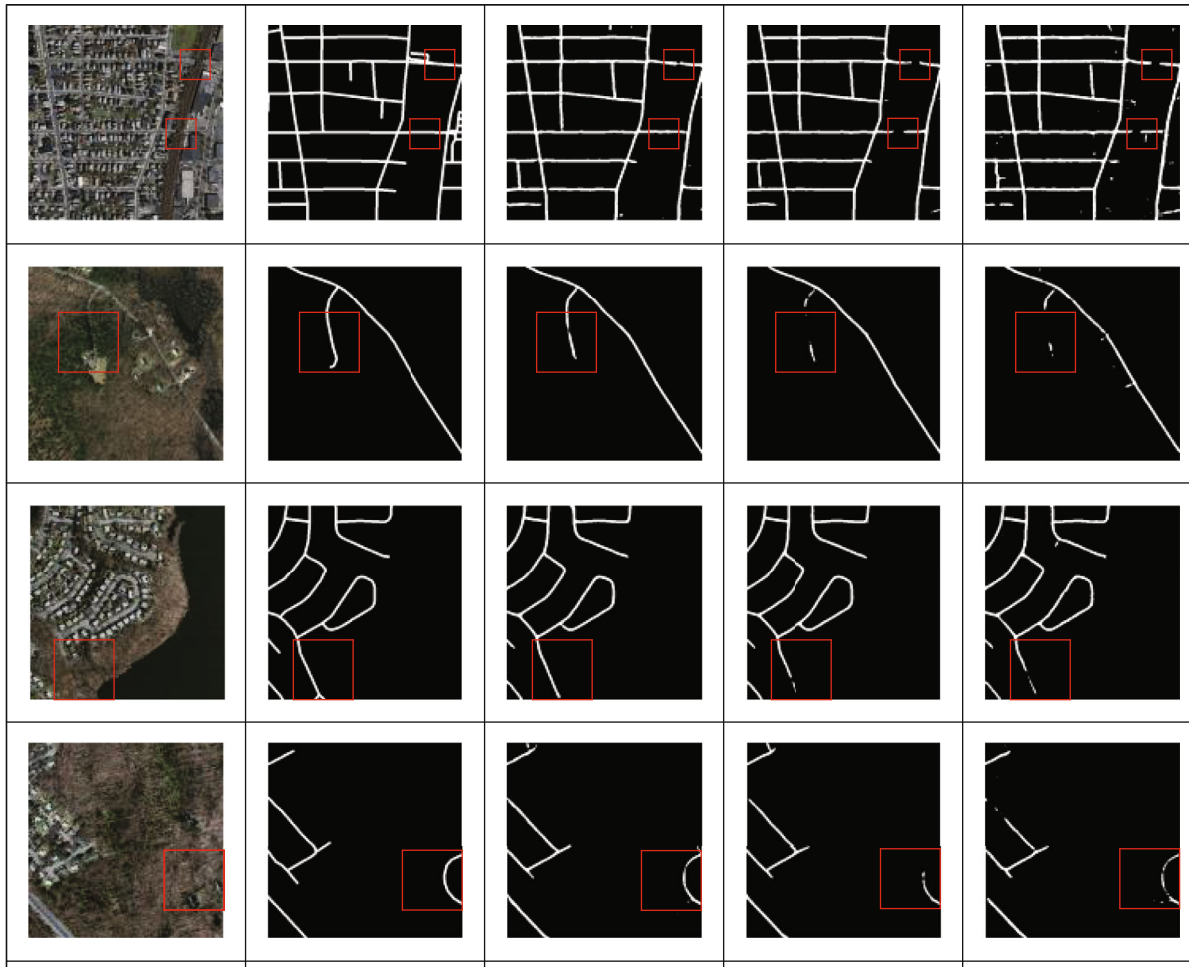

Figure 8: Road extraction results on the Massachusetts road dataset.

$$F1 = 2 \times \frac{p \times R}{P + R}. \tag{1}$$

(2) The intersection over union (IoU) is a comprehensive metric. It is defined as the ratio of the overlap area to the union area between the ground-truth map and the predicted map, and it is calculated as shown in

$$\text{IoU} = \frac{\text{TP}}{\text{FN} + \text{FP} + \text{TP}}, \tag{2}$$

where TP, FN, and FP denote the numbers of true positives, false negatives, and false positives, respectively. True positives are correctly identified road pixels; false negatives are road pixels incorrectly identified as nonroad pixels, and false positives are nonroad pixels incorrectly identified as road pixels

Table 1: Quantitative evaluation results on the Massachusetts road dataset.

Method	$F1$ score	Kappa	IoU
BSIRNet	0.7548	0.7392	0.6062
GAMSNet	0.7454	0.7311	0.5941
SII-Net	0.7199	0.7094	0.5623

(3) The kappa coefficient is an indicator for a consistency test and can also be used to measure the accuracy of semantic segmentation [23]. It may take values in the range of -1 to 1 but usually falls between 0 and 1. The greater the value of the kappa coefficient is, the higher the accuracy. The kappa coefficient is calculated as shown in

$$\text{kappa} = \frac{\text{Pa} - \text{Pe}}{1 - \text{Pe}}, \tag{3}$$

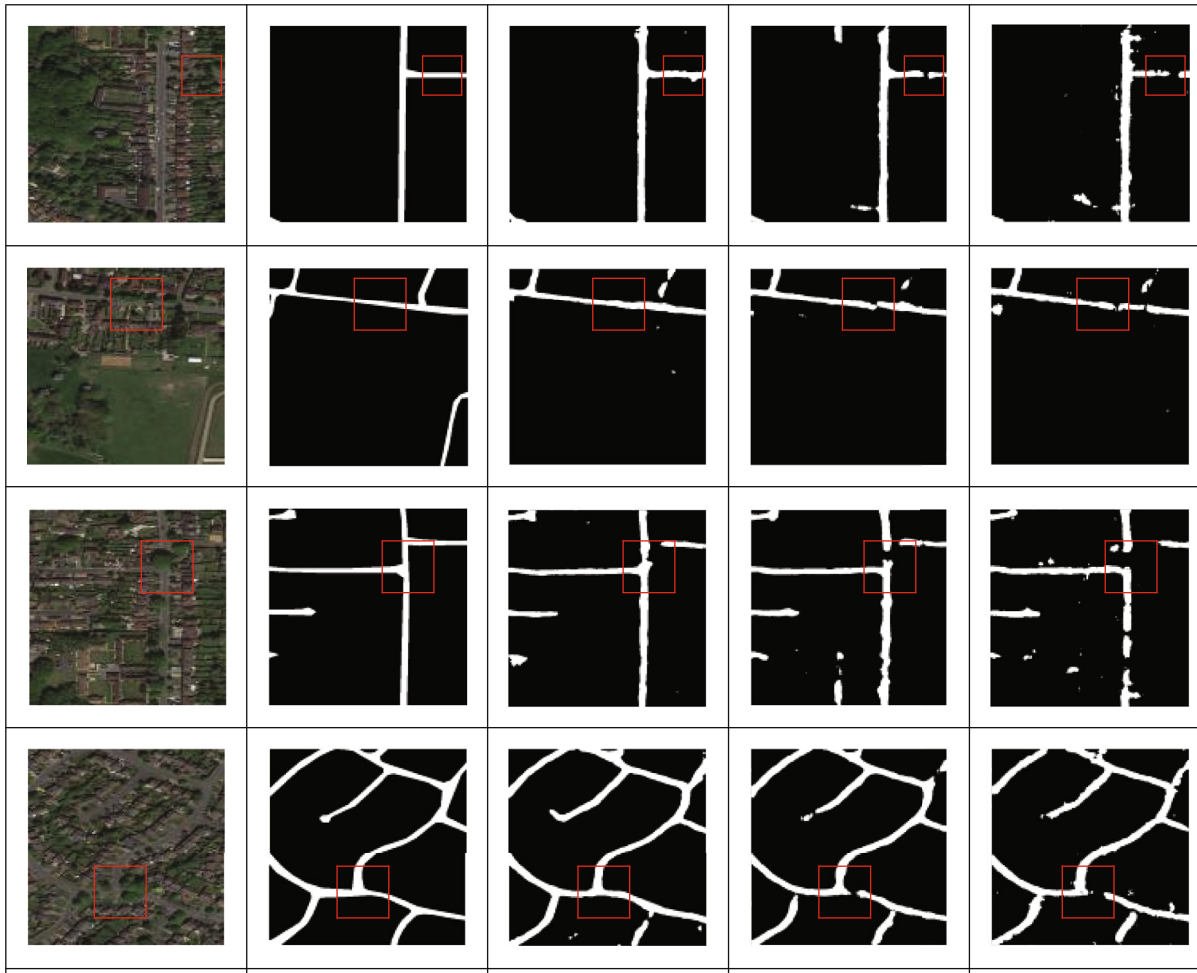

Figure 9: Road extraction results on the Wuhan University road dataset.

where Pa is the "actual agreement rate," and Pe is the "theoretical agreement rate."

4.2. Experiment Using the Massachusetts Road Dataset. In this experiment, road extraction is regarded as a semantic segmentation problem, focusing on the extraction of complete roads. We compare the proposed BSIRNet with two other road extraction methods based on semantic segmentation, namely, GAMSNet and the DeepLabV3+ network with SIIS (SII-Net). As shown in the first and second rows of Figure 8, our proposed BSIRNet completely extracts the occluded road, while the other two networks do not achieve complete extraction of this road. As shown in the third and fourth rows of Figure 8, for roads whose texture is similar to that of surrounding ground features, BSIRNet can also extract these roads completely. This shows that BSIRNet does not only solely depend on the visual characteristics of roads but also has some spatial information reasoning ability by virtue of modelling the specific context of roads. In particular, compared with GAMSNet and SII-Net, BSIRNet has a stronger spatial reasoning ability.

We also conduct a quantitative evaluation to compare the effectiveness of these methods. As shown in Table 1, the BSIRNet proposed in this study obtains an $F1$ score of 0.7548 and a kappa coefficient of 0.7392, greater than those of GAMSNet (with an F1 score of 0.7199 and a kappa coefficient of 0.7094). Compared with those of SII-Net, the F1 score and kappa coefficient of BSIRNet are increased by 3.49% and 2.98%, respectively. Because of the multiscale skip connection structure of BSIRNet, the extracted roads have clear boundaries. Compared with GAMSNet and SII-Net, BSIRNet also achieves a higher IoU. This shows that our spatial information reasoning perception network combined with a multiscale skip connection structure can effectively extract roads. The above experimental results verify the superiority of BSIRNet.

4.3. Experiment Using the Wuhan University Road Dataset. Using the Wuhan University road dataset, we further compare BSIRNet with the above two road extraction methods based on semantic segmentation. As shown in Figure 9, BSIRNet is able to completely extract roads occluded by

Table 2: Quantitative evaluation results on the Wuhan University road dataset.

Method	$F1$ score	Kappa	IoU
BSIRNet	0.7684	0.7392	0.6238
GAMSNet	0.7556	0.7261	0.6072
SII-Net	0.7343	0.7015	0.5801

trees, unlike GAMSNet and SII-Net. Moreover, as shown in the third row of Figure 9, in complex situations, the BSIRNet extraction results have stronger continuity than the results of the other two networks.

Similarly, we conduct a quantitative evaluation of the road extraction results on the Wuhan University road dataset. In Table 2, it can be clearly seen that BSIRNet outperforms the other two networks on the Wuhan University road dataset. Compared with those of GAMSNet, the $F1$ score and kappa coefficient of BSIRNet are increased by 1.28% and 1.31%, respectively. Compared with those of SII-Net, the $F1$ score and kappa coefficient of BSIRNet are increased by 3.41% and 3.77%, respectively. In addition, the extraction results of BSIRNet have stronger continuity, and the IoU value is also significantly higher than those of the other two networks. These experimental results show that our method results in fewer false extractions and missing extractions, thus further verifying that our method can alleviate the problem of discontinuity in road extraction results based on deep learning in the presence of occlusion and texture similarity.

5. Discussion

In the BSIRNet model proposed in this paper, the SRPM is established to capture spatial context dependence, the CRPM is established to capture interchannel dependencies, and the multiscale skip connection structure is used to capture more semantic information. The SRPM and CRPM together solve the problem that road information sometimes cannot be effectively captured due to occlusion and shadows. The multiscale skip connection structure solves the problem of unclear road boundaries caused by road textures similar to those of surrounding ground features. BSIRNet does not solely depend on the visual characteristics of roads but instead achieves some spatial information reasoning ability by modelling the specific context of roads, thereby solving the problem of road discontinuity caused by noise and occlusion.

In the above experiments, we use the Massachusetts road dataset and the Wuhan University road dataset to compare the BSIRNet model proposed in this paper with GAMSNet and SII-Net. The experimental results prove the effectiveness and superiority of BSIRNet, especially in overcoming the influence of occlusion to maintain the continuity of the extracted roads. The experimental results on the Massachusetts road dataset show that BSIRNet can completely extract roads affected by occlusion. Furthermore, when the road texture is very similar to the texture of surrounding ground object, BSIRNet can also extract roads with clear boundaries. As shown in Table 1, compared with those of SII-Net, the $F1$ score, kappa coefficient, and IoU of BSIRNet on the Massachusetts road dataset are increased by 3.49%, 2.98%, and 4.39%, respectively. The road extraction results of BSIRNet are also significantly better than those of GAMSNet on this dataset. Moreover, as shown in Table 2, the experimental results on the Wuhan University road dataset further prove the effectiveness and superiority of BSIRNet. Additionally, the method proposed in this paper has certain reference significance for the discontinuous extraction of other linear ground objects such as railways, power lines, pipelines, and rivers.

6. Conclusions

This paper proposes a road extraction network with bidirectional spatial information reasoning (BSIRNet), which can effectively improve the accuracy of road extraction. Roads possess natural connectivity; however, when an existing method extracts a road, discontinuity problems can easily arise due to interference from noise and occlusion. To solve this problem, we establish a spatial reasoning perception module (SRPM) to capture spatial context dependence and a channel reasoning perception module (CRPM) to capture interchannel dependence in BSIRNet. At the same time, we use a multiscale skip connection structure to capture more semantic information. Using the public Massachusetts road dataset and the Wuhan University road dataset, the superiority of the proposed method is verified by comparing its results with those of other models. When a road is occluded or shadowed by trees, buildings, etc., or the texture of the road is very similar to that of surrounding objects, BSIRNet can effectively improve the accuracy of road extraction. The quantitative results of our experimental evaluation also confirm the superiority of the proposed method. However, using very high-resolution (VHR) satellite imagery to classify road materials still presents great challenges. Therefore, in future work, we will conduct related research and propose a model for classifying road materials.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This research was funded by the Natural Resources Satellite Remote Sensing Technology System Construction Sub-Project: Satellite Data Product Quality Inspection System (AA2111-9) and the National Natural Science Foundation of China (42071428).

References

[1] Q. Zhu, Y. Zhang, L. Wang et al., "A global context-aware and batch-independent network for road extraction from VHR satellite imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 175, pp. 353–365, 2021.

[2] J. Dai, Y. Wang, Y. Du et al., "Development and prospect of road extraction method for optical remote sensing image," *Journal of Remote Sensing (Chinese)*, vol. 24, no. 7, pp. 804–823, 2020.

[3] J. Dai, T. Zhu, Y. Zhang, Y. Wang, and X. Fang, "Line segment fusion method for high-resolution optical satellite image," *Acta Geodaetica et Cartographica Sinica*, vol. 49, no. 4, pp. 489–498, 2020.

[4] M. Song and D. Civco, "Road extraction using SVM and image segmentation," *Photogrammetric Engineering & Remote Sensing*, vol. 70, no. 12, pp. 1365–1371, 2004.

[5] S. Valero, J. Chanussot, J. A. Benediktsson, H. Talbot, and B. Waske, "Advanced directional mathematical morphology for the detection of the road network in very high resolution remote sensing images," *Pattern Recognition Letters*, vol. 31, no. 10, pp. 1120–1127, 2010.

[6] J. Dai, T. Zhu, Y. Wang, R. Ma, and X. Fang, "Road extraction from high-resolution satellite images based on multiple descriptors," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 227–240, 2020.

[7] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: a meta-analysis and review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 152, pp. 166–177, 2019.

[8] X. Lu, Y. Zhong, Z. Zheng, J. Zhao, and L. Zhang, "Edge-reinforced convolutional neural network for road detection in very-high-resolution remote sensing imagery," *Photogrammetric Engineering & Remote Sensing*, vol. 86, no. 3, pp. 153–160, 2020.

[9] Y. Wei, Z. Wang, and M. Xu, "Road structure refined CNN for road extraction in aerial image," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 709–713, 2017.

[10] R. Kestur, S. Farooq, R. Abdal, E. Mehraj, O. S. Narasipura, and M. Mudigere, "UFCN: a fully convolutional neural network for road extraction in RGB imagery acquired by remote sensing from an unmanned aerial vehicle," *Journal of Applied Remote Sensing*, vol. 12, no. 1, article 016020, 2018.

[11] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.

[12] 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, 2016.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, vol. 2, pp. 234–241, Munich, Germany, 2015.

[14] H. He, S. Wang, D. Yang, S. Wang, and X. Liu, "An road extraction method for remote sensing image based on encoder-decoder network," *Acta Geodaetica et Cartographica Sinica*, vol. 48, no. 3, pp. 330–338, 2019.

[15] M. Zhou, H. Sui, S. Chen, J. Wang, and X. Chen, "BT-RoadNet: a boundary and topologically-aware neural network for road extraction from high-resolution remote sensing imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 168, pp. 288–306, 2020.

[16] X. Lu, Y. Zhong, Z. Zheng, and L. Zhang, "GAMSNet: globally aware road detection network with multi-scale residual learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 175, pp. 340–352, 2021.

[17] H. Huang, L. Lin, R. Tong et al., "UNet 3+: a full-scale connected UNet for medical image segmentation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona International Convention Centre, Spain, 2020.

[18] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Computer Vision–ECCV 2018*, pp. 833–851, Munich, Germany, 2018.

[19] F. Chollet, "Xception: deep learning with depthwise separable convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 1800–1807, Hawaii, America, 2017.

[20] C. Tao, J. Qi, Y. Li, H. Wang, and H. Li, "Spatial information inference net: road extraction using road-specific contextual information," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 158, pp. 155–166, 2019.

[21] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," 2014, https://arxiv.org/abs/1409.2329.

[22] V. Mnih and G. E. Hinton, "Learning to detect roads in high-resolution aerial images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 210–223, Heraklion, Crete, Greece, 2010.

[23] R. Fan, Y. Chen, Q. Xu, and J. Wang, "A high-resolution remote sensing image building extraction method based on deep learning," *Acta Geodaetica et Cartographica Sinica*, vol. 48, no. 1, pp. 34–41, 2019.