remote sensing

MDPI

Article

Road Extraction from High-Resolution Remote Sensing Images via Local and Global Context Reasoning

Jie Chen [1], Libo Yang [1], Hao Wang [1], Jingru Zhu [1], Geng Sun [1], Xiaojun Dai [2], Min Deng [1] and Yan Shi [1,*]

[1] The School of Geosciences and Info-Physics, Central South University, Changsha 410083, China; cj2011@csu.edu.cn (J.C.)
[2] The School of Civil Engineering and Geomatics, Southwest Petroleum University, Chengdu 610500, China
* Correspondence: csu_shiy@csu.edu.cn

Abstract: Road extraction from high-resolution remote sensing images is a critical task in image understanding and analysis, yet it poses significant challenges because of road occlusions caused by vegetation, buildings, and shadows. Deep convolutional neural networks have emerged as the leading approach for road extraction because of their exceptional feature representation capabilities. However, existing methods often yield incomplete and disjointed road extraction results. To address this issue, we propose CR-HR-RoadNet, a novel high-resolution road extraction network that incorporates local and global context reasoning. In this work, we introduce a road-adapted high-resolution network as the feature encoder, effectively preserving intricate details of narrow roads and spatial information. To capture multi-scale local context information and model the interplay between roads and background environments, we integrate multi-scale features with residual learning in a specialized multi-scale feature representation module. Moreover, to enable efficient long-range dependencies between different dimensions and reason the correlation between various road segments, we employ a lightweight coordinate attention module as a global context-aware algorithm. Extensive quantitative and qualitative experiments on three datasets demonstrate that CR-HR-RoadNet achieves superior extraction accuracy across various road datasets, delivering road extraction results with enhanced completeness and continuity. The proposed method holds promise for advancing road extraction in challenging remote sensing scenarios and contributes to the broader field of deep-learning-based image analysis for geospatial applications.

Keywords: remote sensing; image segmentation; road extraction; deep learning; convolutional neural network (CNN)

Citation: Chen, J.; Yang, L.; Wang, H.; Zhu, J.; Sun, G.; Dai, X.; Deng, M.; Shi, Y. Road Extraction from High-Resolution Remote Sensing Images via Local and Global Context Reasoning. *Remote Sens.* **2023**, *15*, 4177. https://doi.org/10.3390/rs15174177

Academic Editors: Qiqi Zhu, Danfeng Hong, Ce Zhang and Pedram Ghamisi

Received: 29 June 2023
Revised: 21 August 2023
Accepted: 22 August 2023
Published: 25 August 2023

1. Introduction

Roads are essential artificial objects and serve as fundamental geographic information. The extraction of road information from remote sensing images holds immense significance across various domains, such as urban planning, land management, traffic management, automatic navigation, route analysis, and emergency response [1–4]. In recent years, remote sensing images have witnessed a notable trend toward vast volumes, multiple sources, and high-resolution capabilities, making them a convenient, dependable, and high-quality data source for high-precision road extraction tasks [5,6]. In high-resolution remote sensing images, roads are characterized by narrow straight lines composed of interconnected homogeneous regions. Distinguishing roads from backgrounds primarily relies on attributes such as spectrum, texture, and topology. However, real-world geographic scenes encompass complex background information, and different roads may exhibit significant variations in appearance, material, and structure [3,7–9], which significantly hinders the accurate identification and positioning of roads. Meanwhile, the problem of road occlusion remains a formidable challenge in high-resolution remote sensing image-based road extraction tasks, as depicted in Figure 1. Various factors, such as trees, vehicles, buildings, or shadows, occlude roads, impacting their spectral, color, and texture consistency to varying

degrees. This directly results in incomplete and discontinuous extraction results [10–15]. The omnipresence of road occlusion presents a significant challenge in road extraction: how to ensure the completeness and continuity of roads during the extraction process and effectively enhance the model's anti-occlusion capability. As a result, achieving efficient, high-precision, and automated road extraction while ensuring road continuity has consistently remained a major challenge in the field of remote sensing.

Figure 1. Examples of occlusion in remote sensing images. Each image covers an area of 256 m × 256 m.

Prior to the emergence of deep learning, the mainstream road extraction method involved manually designing effective features for road properties, such as spectrum, geometry, color, texture, and topology. Machine learning algorithms, like clustering and classification, were then employed to distinguish roads from the background [4]. These methods can be categorized as pixel based or object based depending on the analytical scale, and feature based or classification based depending on how features are represented and learned. However, in recent years, deep convolutional neural networks have taken center stage in road extraction tasks, gradually becoming the dominant technology. Most deep-learning-based road extraction methods are based on the encoder–decoder structure, effectively extracting road semantic features from complex scenes and handling highly differentiated roads with robust processing capabilities [5,6]. Some studies aim to optimize the model's internal structure and enhance road feature representation through effective feature extraction modules, thus improving the accuracy of road extraction [12–16]. Other studies employ multi-task learning methods, extracting road surfaces, centerlines, and boundaries simultaneously, to enhance road feature representation through constraints among multiple tasks [17,18].

However, the existing research's feature representation mode, based on a local receptive field, faces challenges in effectively establishing the topological relationship between road segments separated by occlusions [12,19]. Consequently, some studies employ context information to enhance the road semantic features of occluded parts, ensuring road completeness and continuity. Context information extraction algorithms utilize either multi-scale feature representation [20–25] or attention mechanisms [12,23,26–30]. While multi-scale features can model dependencies between geo-objects and the background, attention mechanisms can model correlations between homogeneous geo-objects. However, there are concerns regarding the insufficient coupling of multi-scale feature modules with the feature-learning process and the large number of parameters and computations associated with the self-attention mechanism when applied to high-resolution feature maps. Furthermore, encoder–decoder networks may suffer from the loss of narrow road information because of downsampling, and the skip connections between visual and semantic features may introduce irrelevant low-level noise information.

To address the challenges posed by road occlusion in high-resolution remote sensing images, we investigate strategies to enhance the completeness and continuity of road extraction results and propose a context-reasoning high-resolution road extraction network, CR-HR-RoadNet. Specifically, we leverage a road-adapted high-resolution network as the fundamental feature encoder to effectively preserve narrow road information and spatial details, thus enhancing the model's feature representation capability and improving road boundary extraction accuracy.

To better utilize multi-scale features, we introduce a multi-scale feature representation module, which couples multi-scale features into the feature-learning process to enhance

local context reasoning. This module effectively models dependencies between roads and their backgrounds, enhancing the semantic features of occluded roads.

Addressing the computation concerns of the self-attention mechanism, we employ a lightweight coordinate attention module for global context reasoning. This module generates effective channel and spatial attention weights, enhancing the model's ability to reason about correlations between homogeneous road objects and improving the semantic features of occluded roads.

In summary, the main contributions of this paper are as follows:

(1) We address the loss of narrow road information caused by downsampling and irrelevant low-level noise from skip connections by using a road-adapted high-resolution network as the feature encoder. This approach effectively retains narrow road information and spatial details, enhancing the model's feature representation ability and improving road boundary extraction accuracy.

(2) To improve the utilization of multi-scale features, we propose a multi-scale feature representation module that integrates multi-scale features into the feature-learning process, enhancing the model's local context reasoning ability. This facilitates effective modeling of dependencies between roads and their backgrounds and enhances the semantic features of occluded roads.

(3) To address computation concerns related to the self-attention mechanism, we introduce a lightweight coordinate attention module for global context reasoning. This module generates effective channel and spatial attention weights, enhancing the model's ability to reason about correlations between homogeneous road objects and improving the semantic features of occluded roads.

2. Related Work

2.1. Traditional Road Extraction Methods

Traditional road extraction methods primarily rely on manually designed features, such as spectrum, texture, and geometry, to distinguish roads from the background. We review existing research conducted at two different analytical scales: pixel-based methods and object-based methods [23]. These studies employ feature-based and classification-based approaches at varying analytical scales.

Pixel-based methods focus on extracting spectral and texture features at the pixel level and utilize algorithms like classification algorithms to identify road areas. For instance, Song and Civco [31] proposed a two-stage model for road extraction from remote sensing images. They initially classify all pixels into road and non-road groups using the support vector machine algorithm based on their spectral features. Subsequently, the road group is further refined using a segmentation algorithm to generate accurate road areas. Jing et al. [32] presented a road centerline extraction method based on multi-scale joint features. This method effectively integrates spectral, geometric, and texture features of roads in high-resolution images to produce multi-scale unified features. Pixel-based methods can successfully extract roads with clear boundaries and straightforward backgrounds. However, the results may suffer from salt-and-pepper noise, necessitating sophisticated post-processing methods for refinement.

Object-based methods, on the other hand, identify road objects as a whole, which helps to mitigate salt-and-pepper noise and spectral outliers, providing robust noise immunity and applicability. Maboudi et al. [33] introduced an object-based method for road extraction from high-resolution images, integrating spatial, spectral, and textural descriptors and using object-based image analysis and the ant colony algorithm to extract road regions. Chen et al. [34] proposed a two-stage approach that combines region and boundary information. The first stage involves performing connection analysis on discrete line features with direction consistency to extract potential road objects, and the second stage employs shape features to further refine the results. However, these methods heavily rely on the segmentation outcomes of objects in the image, and confusion can easily occur between adjacent objects with similar spectra.

2.2. Deep-Learning-Based Road Extraction Methods

Given that most road features are artificially designed and cannot adequately represent roads under various conditions, traditional road extraction methods suffer from several shortcomings, including low automation, complicated and time-consuming operations, and limited generalization [19,23]. In recent years, deep convolutional neural networks (DCNN) have garnered significant attention in road extraction because of their powerful nonlinear feature-learning capabilities. DCNN methods have become the mainstream technology for road extraction tasks.

Long et al. [35] proposed a Fully Convolutional Neural Network (FCN), which restores the size of the feature image by upsampling, allowing it to not only identify the cat-egory of every pixel but also restore the position of pixels in the original image. Given that the FCN can realise the pixel-level classification of the image, many methods have been improved and optimised based on the FCN, such as SegNet [36], UNet [37] and DeepLab [38]. Currently, most road extraction research is based on encoder–decoder architectures, which have shown excellent performance and rapid development [5,6]. The encoder gradually reduces the resolution of feature images through downsampling to expand the receptive field and extract deep semantic road features. Meanwhile, the decoder gradually restores the resolution of feature maps through upsampling, achieving end-to-end pixel-level road extraction.

Researchers have improved road extraction accuracy by optimizing existing model structures, such as loss functions, feature-learning units, and multi-level feature modules. For instance, Zhang et al. [16] combined residual learning and U-Net to design a deep residual U-Net network for road extraction. Yang et al. [39] proposed a new recurrent convolutional neural network U-Net with an RCNN module that effectively utilizes spatial semantic information and rich visual features to address noise and complex background issues.

Furthermore, the road's narrow and elongated appearance, ranging from several meters to tens of meters in width, often occupies a small area in remote sensing images. Continuous downsampling operations can cause some narrow roads to disappear in feature maps, leading to missing information that is challenging to recover accurately during upsampling and adversely affecting the extraction accuracy. Additionally, many studies have used skip connection architectures to obtain spatially detailed information. However, it has been observed that low-level features contain more noise because of problems like occlusion and spectral outliers, leading to uneven road boundaries and interruptions when using skip connections between shallow and deep features. Thus, employing multiple skip connection operations in the encoder–decoder structure is unnecessary for road extraction [40].

To establish an effective topological relationship between road segments separated by occlusions and ensure a continuous road extraction result, researchers have focused on enhancing the feature representation of occluded parts by incorporating road context information to improve the model's anti-occlusion ability [19,20,23,25]. Context information encompasses dependencies between geo-objects and the background, as well as correlations between homogeneous geo-objects. Currently, most studies capture effective context information using multi-scale features, attention mechanisms, and self-attention mechanisms.

Multi-scale feature techniques, such as image pyramids, pyramid pooling, skip connections, and atrous convolutions, effectively capture local context and have found widespread use in road extraction tasks. For example, D-LinkNet [21] is a UNet-like network that incorporates multi-scale atrous convolution modules, enabling the model to capture road features at multiple scales. Other studies, like those of Gao et al. [12], Wu et al. [22], Zhu et al. [23], and Tan et al. [24], have proposed customized multi-scale modules to enhance accurate road edge perception and feature representation. Despite the effectiveness of these multi-scale modules in capturing local context, there remains room for improvement in terms of the interaction between multi-scale modules and feature learning in the encoder, which affects the model's overall feature representation ability for road extraction.

Attention mechanisms play a vital role in road extraction by enhancing feature representation. These mechanisms can be categorized into channel attention and spatial attention. Spatial attention focuses on establishing attention weights on spatial locations through large-scale kernel convolutions, thereby improving road extraction by considering the local texture and morphological structure of the road. On the other hand, self-attention mechanisms establish long-distance spatial or channel dimension attention. The use of self-attention has shown promise in improving the segmentation accuracy of roads in remote sensing images by capturing global context and enhancing the completeness of road regions [23,30].

However, it should be noted that spatial attention mechanisms based on convolution operations can only capture local relations and may not adequately model desired long-range dependencies. Similarly, self-attention mechanisms, although effective in capturing global context, can introduce a large number of parameters and computations, leading to increased time and memory costs during model training and inference. This issue becomes particularly challenging when applied to high-resolution feature maps.

To address these challenges and improve the accuracy and completeness of road extraction results, this work proposes a high-resolution road extraction network (CR-HR-RoadNet) based on context reasoning. The network aims to enhance feature representation by effectively utilizing multi-scale context information while addressing the issue of road occlusion.

3. Methods

To address the practical problem of incompleteness and discontinuity caused by occlusions in remote sensing images, we propose a CR-HR-RoadNet by using the feature enhancement effect of prior contextual information. On the one hand, the feature representation ability of the model is enhanced during the feature-learning process. On the other hand, the road information of the occluded part is mined. The specific model structure is shown in Figure 2, which includes two main parts: a road-adapted high-resolution backbone network and a local and global context reasoning module. The local and global context reasoning modules include the multi-scale feature representation module and the coordinate attention module. In particular, the multi-scale feature representation module, as the main feature-learning module, exists in the entire feature-learning process and is used to reason local context information. The coordinate attention module is between different feature-learning stages and is used to reason global context information. The two modules influence each other. The richer the multi-scale road features captured by the multi-scale feature representation module, the more effective the subsequent coordinate attention module will be, and vice versa.

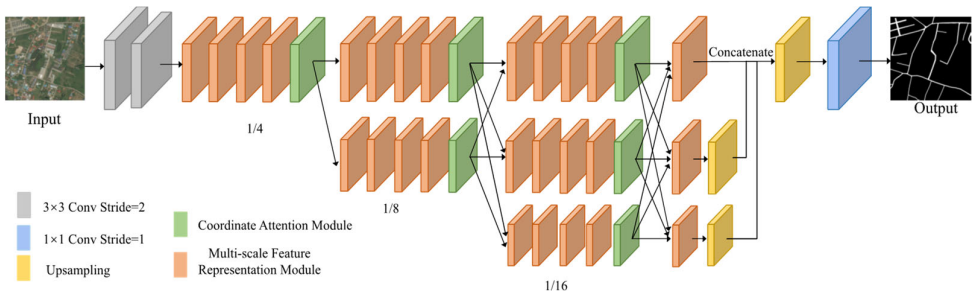

Figure 2. Specific structure of the CR-HR-RoadNet.

3.1. Road-Adapted High-Resolution Network

The continuous downsampling operation in the encoder–decoder network will reduce the resolution of the feature map, causing some narrow roads to disappear in the low-resolution feature maps. Skip connections may also bring irrelevant noise information, which seriously affects the effect of road extraction. We aim to ensure that the road information is not lost, whilst enabling the network to extract deep semantic features and

capture rich spatial details. On the basis of [41], we use a high-resolution network to replace the encoder–decoder network as the backbone network to ensure that the feature maps are maintained at a high resolution. Specifically, this network is able to not only retain the complete road information but also ensure that the road has rich spatial detail information. The specific structure is shown in Figure 2.

First, we use two standard convolutions of 3×3 with a stride of 2 to process the input image and downsample the image resolution to the quarter of the original image. The result is used as the high-resolution input of the next module to reduce model computation and preserve complete road information and valid spatial details. Then, the branch with $4\times$ downsampling is used as the first stage in the multi-resolution branch structure. The model will gradually add a new branch according to the resolution from high to low to generate a new structure. Specifically, the parallel branch of the later stage is composed of all the branches of the previous stage and a new branch with a lower resolution. Then, feature fusion is performed on the feature maps of all branches in the output part of the model. A fusion feature map with $4\times$ downsampling is obtained by merging the outputs of the three branches. Finally, the bilinear interpolation operation is used to restore the size of the fusion feature map to the original image size, and the final prediction map is obtained through the standard convolution of 1×1.

The HRNet-w32 is selected as the main backbone model and adapts the network structure in the original paper for the task of road extraction. The original $32\times$ downsampling branch is deleted to prevent the disappearance of road semantic information caused by a considerably low resolution. Therefore, the proposed road-adapted high-resolution network has a total of three parallel branches, and the corresponding image resolutions are $4\times$, $8\times$, and $16\times$ downsampling. The proposed multi-scale feature representation module is used as the basic feature-learning module in all branches, thereby improving the feature representation ability of the backbone model. The higher-resolution branch in the multi-branch structure enables the model to always retain accurate spatial detail information and complete narrow road information. The low-resolution branch enables the model to extract sufficiently effective deep semantic features. Thus, the multi-branch structure can achieve strong semantic information learning and precise location information capture. Considering that the model has multiple branches, the number of feature channels in the model must be reduced to minimize the scale of the model and prevent the amount of model parameters and computation from being considerably large.

After each stage of feature learning is completed, a deep information interaction occurs between different branches, namely, the feature fusion process, as shown in Figure 3. In the case of three branches in parallel, (a) represents the $1/4$ branch fuse feature information from the $1/8$ and $1/16$ branches, (b) represents the $1/8$ branch fuse feature information from the $1/4$ and $1/16$ branches, and (c) represents the $1/16$ branch fuse feature information from the $1/4$ and $1/8$ branches. The upsampling operation is mainly realized by bilinear interpolation, and the downsampling operation is realized by standard convolution with a stride of 2. The feature fusion aims to exchange information between multi-resolution representations. Each branch can receive feature information from other branches to supplement the information loss caused by the reduction in the number of feature channels and effectively enhance the feature representation ability of the model.

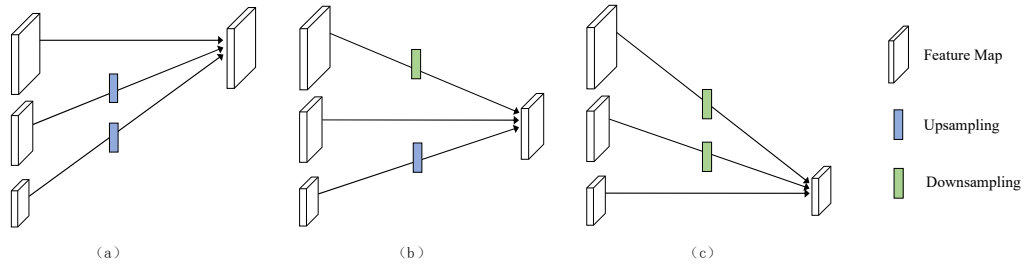

Figure 3. Multi-branch feature fusion process. (**a**) represents the 1/4 branch fuse feature information from the 1/8 and 1/16 branches, (**b**) represents the 1/8 branch fuse feature information from the 1/4 and 1/16 branches, and (**c**) represents the 1/16 branch fuse feature information from the 1/4 and 1/8 branches.

3.2. Multi-Scale Local Context Reasoning

The multi-scale feature representation module combines multi-scale convolution and residual learning units [42]. This module aims to realize the effective representation and aggregation of the local context information with multiple scales, thereby improving the feature representation ability of the encoder and enhancing the feature representation of the occlusion parts by reasoning the dependence between the road and the background environment. The module is embedded in each branch of the backbone network. Accordingly, the multi-scale feature representation is fused in the whole feature-learning process, and the coupling degree between the two parts is effectively improved. The specific module structure is shown in Figure 4. According to the different types of residual learning units, the corresponding multi-scale feature representation modules are also different: (a) denotes the multi-scale feature representation module based on the BasicBlock module, and (b) denotes the multi-scale feature representation module based on the BottleNeck module.

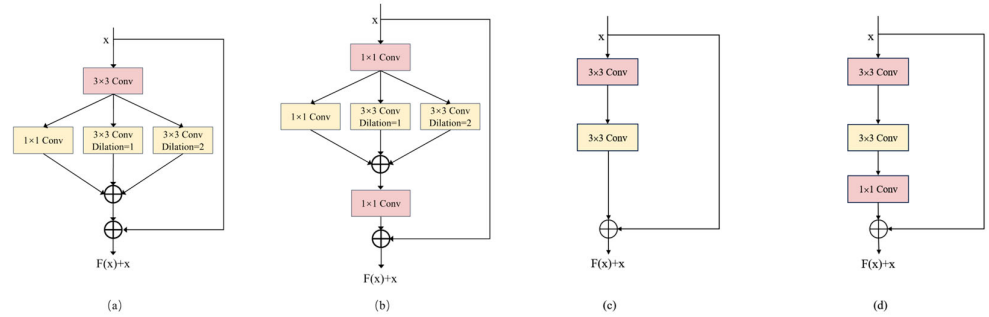

Figure 4. Structure of the multi-scale feature representation module. (**a**) denotes the module based on the BasicBlock unit, (**b**) denotes the module based on the BottleNeck unit, (**c**) denotes the original BasicBlock unit, and (**d**) denotes the original BottleNeck unit.

In Figure 4, we modify the original residual learning unit and replace the standard convolution of 3×3 with multi-scale convolution. We use atrous convolution as the main technique to extract multi-scale local context [38,43–46] and control the size of the dilation rate to realize receptive fields of different sizes. The multi-scale feature representation module mainly uses three convolution kernels of different sizes to extract road features of different spatial scales: that is, the standard convolution kernel of 1×1, the dilated convolution of 3×3 with dilation rate of 1, and the dilated convolution kernel of 3×3 with dilation rate of 2. The standard convolution kernel of 1×1 is used to extract the features of the road itself, whilst the other two dilated convolutions are utilized to capture local road context information. The feature representation of the occlusion parts is enhanced by the reasoning local context information at different scales. The module inputs the feature maps into the three convolutional layers for feature extraction and uses the addition operation to

fuse the output feature maps of the three scales. Then, the fusion result is inputted into the subsequent residual learning process.

3.3. Coordinate Attention-Based Global Context Reasoning

On the basis of [47], we use the coordinate attention module as the main method to capture long-range dependence between different roads. The goal of this mechanism is to enable the network to learn effective global context information to enhance the feature representation of the occlusion parts by capturing the feature correlations between homogeneous road geo-objects. The coordinate attention module can effectively capture the global attention in the feature channel and space location and has a low amount of computation and parameters compared with the other attention mechanisms. This mechanism is a lightweight module and can be well embedded anywhere in the model.

(1) Coordinate information embedding: global average pooling is often used for channel attention to encode spatial information globally, but it compresses global spatial information into channel descriptors. Accordingly, the location information is difficult to preserve. Location information is the key to capture the spatial structure in vision tasks. Therefore, accurate spatial location information must be retained, and the global feature information must be captured during feature compression.

During the coordinate information embedding, the 2D global average pooling operation is decomposed into two 1D global average pooling operations to encourage the attention module to capture long-range spatial interactions and precise location information. This module performs feature compression along the x direction (horizontal) and y direction (vertical) to generate a pair of feature tensors with different spatial information, namely, the X Avg Pool layer and the Y Avg Pool layer in Figure 5. Specifically, given the input $X \in R^{C\times H\times W}$, two 1D average pooling kernels, (1, W) and (H, 1), are used for each channel of the feature map along the horizontal and vertical dimensions, respectively. After information compression, two feature tensors, $f_x \in R^{C\times H\times 1}$ and $f_y \in R^{C\times 1\times W}$, that aggregate different spatial information are obtained. The output of the c_{th} channel at height h or width w can be expressed as follows:

$$f_x^c(h) = \frac{1}{W}\sum_{0\le i<W} X^c(h,\, i), \tag{1}$$

$$f_y^c(w) = \frac{1}{H}\sum_{0\le j<H} X^c(j,\, w) \tag{2}$$

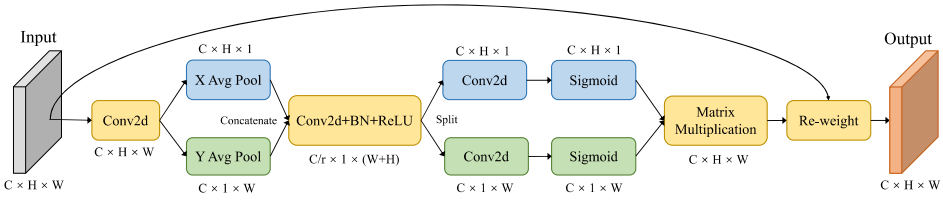

Figure 5. Structure of the coordinate attention module. For a detailed explanation, please refer to Section 3.3.

In summary, the coordinate attention module compresses features maps along two spatial directions through 1D global average pooling and preserves precise location information of feature maps, which helps in accurately locating regions of interest. Coordinate information embedding aims at aggregating global context information from different directions, enabling information interaction between different road areas and modeling feature connections between occlusion areas and other road areas.

(2) Coordinate attention generation: the coordinate attention generation stage aims to reason the context information aggregated in different directions, thereby enabling the model to localize the road regions of interest and generate effective spatial and channel

attention weights to indirectly enhance the road features of occlusion parts. First, the horizontal and vertical feature tensors are concatenated to generate a new feature tensor, $f \in R^{C\times 1\times (W+H)}$. Second, a shared 1×1 standard convolution is used to perform feature transformation on the feature tensor, thereby generating a dimension-reduced feature tensor, $F \in R^{C/r\times 1\times (W+H)}$, where r represents the downsampling ratio of the channel dimension. Third, the module inputs the tensor into a batch normalization layer and a nonlinear activation layer for processing and separates the dimension of the feature tensor F to obtain the feature tensors $F_x \in R^{C/r\times H\times 1}$ and $F_y \in R^{C/r\times 1\times W}$ in two different directions. Then, the module uses two 1×1 standard convolutions to perform attention calculation on the two feature tensors, thereby obtaining attention tensors $G_x \in R^{C\times H\times 1}$ and $G_y \in R^{C\times 1\times W}$ in different directions. Finally, the module uses the sigmoid function to normalize the attention tensors and limits the value to the range of zero to one. The complete global attention weight matrix $G \in R^{C\times H\times W}$ is obtained by the matrix multiplication between G_x and G_y. This attention map contains adaptive attention in the channel and spatial dimensions.

Then, the module multiplies the attention weight G by the initial input X to complete the re-weighting process, thereby achieving the attention optimization and obtaining the final output $Y \in R^{C\times H\times W}$. The detailed calculation process is shown in the following formulae:

$$F = ReLU\big(BN\big(Conv\big(\big[f_x, f_y\big]\big)\big)\big), \tag{3}$$

$$G_x = Sigmoid(Conv_x(F_x)), \tag{4}$$

$$G_y = Sigmoid\big(Conv_y\big(F_y\big)\big), \tag{5}$$

$$G = Mul\big(G_x, G_y\big), \tag{6}$$

$$Y = X * G \tag{7}$$

where $Conv(.)$ represents the convolution operation, $BN(.)$ represents the batch normalization operation, $ReLU(.)$ represents the nonlinear activation function, $Mul(.)$ represents the matrix multiplication operation, * represents the element-wise multiplication, and [,] represents the tensor stacking operation.

In summary, the coordinate attention module not only considers the importance between different channels but also pays attention to the feature encoding between different spatial locations. The elements in the attention tensors reflect whether the road region of interest exists in the corresponding row and column by paying attention to the input in both horizontal and vertical directions. In this way, the model can accurately locate the road areas in each feature channel, achieves attention optimization in different dimensions, and effectively enhances the feature representation of the roads, thereby helping the model to better extract occluded road areas.

4. Experiments and Results

4.1. Datasets

We select three high-resolution remote sensing image road extraction datasets for model evaluation, namely, the Massachusetts Roads Dataset [48], DeepGlobe Roads Dataset [49], and CH6-CUG Roads Dataset [23], to verify the extraction effect and performance of the proposed model on high-resolution remote sensing images. A specific example is shown in Figure 6.

The Massachusetts Roads Dataset [48] is an aerial remote sensing image dataset collected in Massachusetts. The dataset covers multiple geographic scenes, such as urban, suburban, and rural scenes. The dataset contains a total of 1171 images, of which 1108 images are used for model training, 14 images are employed for model validation, and 49 images are utilized for model testing. The spatial resolution of this dataset is 1.2 m, and each image is 1500×1500 pixels in size. We randomly crop the images in the training and validation sets into several image patches of 256×256 and obtain 20,000 images for

training and 500 images for validation. Furthermore, we randomly augment the training images to expand the dataset during the training process.

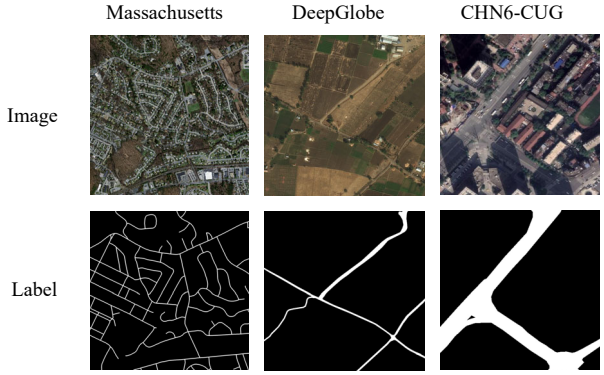

Figure 6. Examples of different road datasets. From left to right, the geographic extents of the images are 1800 m × 1800 m, 512 m × 512 m, and 256 m × 256 m, respectively.

The DeepGlobe Roads Dataset [49] is a satellite remote sensing image dataset containing images collected from Thailand, Indonesia, and India. The dataset includes geographic scenes, such as cities and suburbs with rich road types. The original dataset contains 8570 three-channel satellite remote sensing images, of which only 6226 images contain the corresponding real label data. The size of each image is 1024 × 1024 pixels, and the image spatial resolution is 50 cm. We divide the images containing the ground truth labels according to the ratio of 7:1:2. The training, validation, and test sets contain 5000, 226, and 1000 images, respectively. We randomly crop the images in the training and validation sets into several image patches of 256 × 256 and obtain 25,000 images for training and 1130 images for validation. Furthermore, we randomly augment the training images to expand the dataset during the training process.

The CHN6-CUG Roads Dataset [23] is a large-scale satellite image dataset containing representative cities in China. The remote sensing images within this dataset are acquired from Google Earth. Based on urbanization level, city scale, developmental stage, urban structure, and historical and cultural significance, a careful selection of six Chinese cities is made: Beijing, Shanghai, Wuhan, Shenzhen, Hong Kong, and Macau. The road types in this dataset include railways, highways, urban roads, and rural roads. The dataset contains a total of 4511 remote sensing images with a size of 512 × 512 and their corresponding ground-truth labels. A total of 3608 images are used for model training, and 903 images are utilized for testing. The spatial resolution of the images is 50 cm. We randomly crop the remote sensing images into several image patches of 256 × 256 in the training set and obtain a total of 23,000 images for model training. Moreover, we randomly augment the training images to expand the dataset during the training process.

4.2. Experiment Setting and Evaluation Metrics

In the experimental part, a total of nine mainstream deep convolutional neural networks are selected as comparison models. These models include the FCN-style and encoder–decoder models. All models have better context reasoning ability. For example, the DLinkNet uses a parallel multi-scale atrous convolution model to obtain multi-scale local context information, and the DANet captures the global context information in the spatial and channel dimensions by using a dual attention mechanism. Therefore, these comparative models can effectively test the effectiveness of the proposed method.

All experiments in this chapter are implemented using the PyTorch deep learning framework. We select UNet, deeplabv3+, and other models as comparison models to verify the effect of the proposed road extraction network and train and test these models on three datasets. The specific experimental settings are as follows: an Adam optimizer with a momentum of 0.5 and weight decay of 0.999 is selected as the main optimizer for training,

the parameter weights of all models are randomly initialized, and the learning rates of all models are initialized to 1×10^{-4}. We set the batch size to a dynamic interval of 8 to 16 and the number of iterations to 100 epochs and use binary cross-entropy loss and dice loss to perform supervision on all models, depending on the scale of the model. During the training process, the training learning rate is dynamically adjusted using the poly learning strategy.

To accurately evaluate the performance and accuracy of the proposed model, we use four common and effective metrics to form an evaluation system, which are precision, recall, $F1$, and interaction of union (IoU). The higher the metric value of the above-mentioned four metrics, the better the performance of the road extraction model. The specific calculation formula is as follows:

$$Precision = \frac{TP}{TP + FP}, \tag{8}$$

$$Recall = \frac{TP}{TP + FN}, \tag{9}$$

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}, \tag{10}$$

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

where TP, FN, FP, and TN represent the true positive, false negative, false positive, and true negative, respectively.

4.3. Result Evaluation on Massachusetts Dataset

Table 1 shows the quantitative analysis results of all models on the Massachusetts dataset. The proposed CR-HR-RoadNet can achieve superior performance on the Massachusetts dataset and achieve the highest accuracy on precision, F1, and IoU. The recall of the proposed model is second only to the EMANet, but the value of precision, F1, and IoU is much higher than that of the EMANet, indicating that the comprehensive performance of the proposed model is better. Specifically, the proposed model achieves 78.19% on F1 and 64.19% on IoU. The DLinkNet is the model with the best performance amongst all the comparison models because it achieved 77.17% on F1 and 62.83% on IoU. The proposed model is 1.02% and 1.36% higher on F1 and IoU, respectively, compared with the DLinkNet. This result shows that the context reasoning frame of the proposed model can enhance the feature representation ability and recover the features at the occlusion parts by using the dependencies with the environment and the correlation with the homogeneous geo-objects, thereby greatly improving the extraction accuracy. The results of the quantitative evaluation prove the effectiveness of the multi-scale feature representation module and coordinate attention module on the Massachusetts dataset.

Table 1. Quantitative evaluation results of different methods on the Massachusetts dataset.

Methods	Backbone	Precision	Recall	F1	IoU
UNet [37]	None	79.67	74.30	76.89	62.46
DeepLabV3+ [38]	ResNet101	77.33	73.30	75.26	60.34
DenseASPP [46]	DenseNet121	72.15	71.36	71.75	55.95
SENet [50]	ResNet101	74.81	68.08	71.28	55.38
OCNet [51]	ResNet101	71.12	72.25	71.68	55.86
EMANet [52]	ResNet101	73.71	77.72	75.66	60.85
DANet [53]	ResNet101	60.68	74.67	66.95	50.32
ResUNet [16]	UNet	78.31	75.52	76.89	62.45
DLinkNet [21]	ResNet101	78.96	75.46	77.17	62.83
CR-HR-RoadNet (ours)	HRNet	80.34	76.15	78.19	64.19

Figure 7 shows the qualitative analytical results on the Massachusetts dataset. The DANet, DeepLabV3+, and DLinkNet in the comparison model are selected as the main qualitative comparison objects. These three models can comprehensively and objectively compare and evaluate the road extraction effect of the proposed model. The visualization results show that the proposed model can achieve excellent road extraction results. The road boundary in the prediction results is smoother, and the completeness and continuity are better than those of the other three models. Meanwhile, the results also show less misclassification and noise information. Amongst the extraction results of the three comparison models, the results of the DANet are the roughest, and the boundary is not smooth enough, which may be caused by direct upsampling. The extraction results of DeepLabV3+ and DLinkNet have some incompleteness and discontinuity cases. Specifically, the complex road and occlusion areas are marked by red circles in the visualization results. The proposed model can achieve better extraction results more in line with the ground truth and has significantly better performance than the other three models in terms of completeness and continuity. Therefore, the qualitative analytical results can prove that the proposed model has better road extraction effect on the Massachusetts dataset, and the extraction advantage on some complex roads and occlusion areas is more obvious.

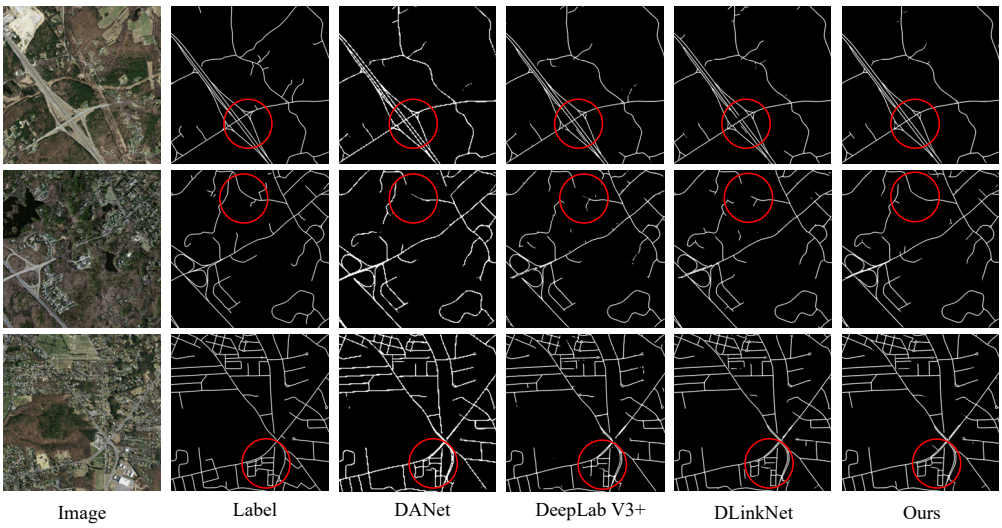

Figure 7. Qualitative evaluation results on the Massachusetts dataset. Each image covers an area of 1800 m × 1800 m.

4.4. Result Evaluation on DeepGlobe Dataset

Table 2 shows the quantitative analysis results of all models on the DeepGlobe dataset. The proposed CR-HR-RoadNet can achieve superior performance on the DeepGlobe dataset and the highest accuracy on recall, F1, and IoU. The precision of the proposed model is lower than that of the DenseASPP, the EMANet, and the DLinkNet, but the recall of the three models is much lower than that of the proposed model. This finding shows that the comprehensive performance of the proposed model is better. Although the EMANet achieves the highest precision of 82.75%, its recall is only 56.65%, resulting in the worst accuracy on F1 and IoU. Specifically, the proposed model can achieve 76.79% on F1 and 62.33% on IoU. Amongst all comparison models, the model with the best performance is the DLinkNet, which achieves 75.74% on F1 and 60.95% on IoU. The proposed model is 1.05% and 1.38% higher on F1 and IoU, respectively, compared with the DLinkNet model. This finding shows that the performance of our model is much better than that of the other comparison models. The quantitative evaluation results prove the effectiveness of the multi-scale feature representation module and coordinate attention module on the DeepGlobe dataset.

Table 2. Quantitative evaluation results of different methods on the DeepGlobe dataset.

Methods	Backbone	Precision	Recall	F1	IoU
UNet [37]	None	73.50	73.45	73.48	58.07
DeepLabV3+ [38]	ResNet101	68.60	76.79	72.46	56.82
DenseASPP [46]	DenseNet121	77.00	65.45	70.76	54.75
SENet [50]	ResNet101	73.55	75.02	74.28	59.08
OCNet [51]	ResNet101	72.82	71.49	72.15	56.43
EMANet [52]	ResNet101	82.75	56.65	67.26	50.66
DANet [53]	ResNet101	69.53	71.43	70.46	54.40
ResUNet [16]	UNet	73.51	72.69	73.10	57.60
DLinkNet [21]	ResNet101	76.68	74.83	75.74	60.95
CR-HR-RoadNet (ours)	HRNet	76.47	77.12	76.79	62.33

Figure 8 shows the qualitative analytical results on the DeepGlobe dataset. ResUNet, OCNet and DLinkNet are selected as the main qualitative analytical objects. The visualization results show that the CR-HR-RoadNet can achieve the best road extraction accuracy and can obtain more complete and continuous extraction results with smoother boundaries and less noise. Multiple areas are marked by red circles in the visualization results. The proposed model can obtain better road extraction results in these areas. Specifically, some narrow roads are occluded by vegetation in the remote sensing images in the first row. Neither the ResUNet nor DLinkNet models can completely extract the narrow roads. Although the OCNet can completely extract the narrow roads, the boundaries are rough. The proposed model can completely extract narrow roads and ensure that the road boundaries are smooth enough, which is mainly due to the high-resolution feature encoder that can effectively capture detailed information. In the visualization results of other rows, a large number of road occlusions can be observed in the remote sensing images. Neither ResUNet, OCNet, nor DLinkNet can effectively recover road information at occlusions, resulting in severe incompleteness and discontinuity in the prediction maps. Given the existence of effective local and global context reasoning modules in the proposed model, the proposed model can use the dependence with background and the correlation with homogeneous geo-objects to enhance the feature representation and effectively restore the road information at the occlusion areas. Hence, the proposed model can obtain better completeness and continuity results.

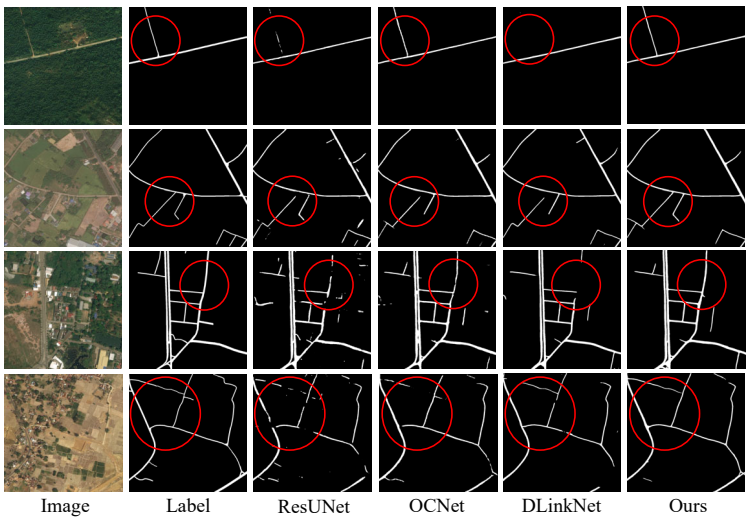

Figure 8. Qualitative evaluation results on the DeepGlobe dataset. Each image covers an area of 512 m × 512 m.

4.5. Result Evaluation on CHN6-CUG Dataset

Table 3 shows the quantitative analysis results of all models on the CHN6-CUG dataset. The proposed CR-HR-RoadNet can achieve superior performance on the CHN6-CUG dataset and achieve the highest accuracy on recall, F1, and IoU. The precision of the proposed model is second only to the UNet and the SENet. However, the recall of these two models is much lower than that of the proposed model, indicating that the comprehensive performance of the proposed model is better. The proposed model achieves 77.92% on F1 and 63.83% on IoU. Amongst all the comparison models, most models achieve good extraction accuracy, and the model with the best performance is EMANet, which achieves 77.16% on F1 and 62.82% on IoU. The proposed model is 0.76% and 1.01% higher on F1 and IoU, respectively, compared with the EMANet, indicating that the proposed model has better road extraction performance. It is worth noting that the extraction accuracy of the UNet and ResUNet is lower, which may be due to the complex background information on the CHN6-CUG dataset and a large amount of noisy information. The skip connection operation in the encoder–decoder structure will introduce some irrelevant information into the decoder, resulting in a decrease in extraction accuracy, which also proves the advantages of the high-resolution network used in this paper. The results of quantitative evaluation prove the effectiveness of multi-scale feature representation module and coordinate attention module on CHN6-CUG dataset.

Table 3. Quantitative evaluation results of different methods on the CHN6-CUG dataset.

Methods	Backbone	Precision	Recall	F1	IoU
UNet [37]	None	78.69	67.38	72.60	56.98
DeepLabV3+ [38]	ResNet101	74.85	76.80	75.81	61.04
DenseASPP [46]	DenseNet121	76.84	74.86	75.83	61.08
SENet [50]	ResNet101	78.59	74.16	76.31	61.70
OCNet [51]	ResNet101	79.52	72.57	75.89	61.14
EMANet [52]	ResNet101	77.77	76.57	77.16	62.82
DANet [53]	ResNet101	77.74	72.88	75.23	60.30
ResUNet [16]	UNet	77.52	66.18	71.40	55.52
DLinkNet [21]	ResNet101	77.67	73.29	75.41	60.53
CR-HR-RoadNet (ours)	HRNet	78.40	77.44	77.92	63.83

Figure 9 shows the visual qualitative analysis results on the CHN6-CUG dataset. ResUNet, DLinkNet, and EMANet are selected as the main qualitative analytical objects to evaluate the road extraction effect of the proposed model comprehensively and objectively. The visualization results demonstrate that the proposed model can obtain the best road extraction results, regardless of whether it is in terms of road completeness or road continuity. Moreover, the noise information is less, and the road boundary is smoother. Multiple areas are marked by red circles in the visualization results. Specifically, the visualization results in the first row show that the proposed model can obtain better extraction results in complex and dense road areas. The roads on the label maps of the second and third rows are not smooth enough and are different from the real situation. However, the proposed model can obtain smoother and more complete prediction results. The proposed model can extract the road areas that are not in the label (bottom right of the image in the second row and top of the image in the third row). The results of the fourth row show that the proposed model has the advantage of maintaining the road completeness. Road occlusions can be observed in the remote sensing images in the fifth row. The proposed model can also use local and global context information to obtain extraction results with better continuity. This finding shows that the proposed model has good anti-occlusion ability. In summary, the proposed model can achieve far better extraction results than the other models on the CHN6-CUG dataset, which fully proves the effectiveness of the proposed method.

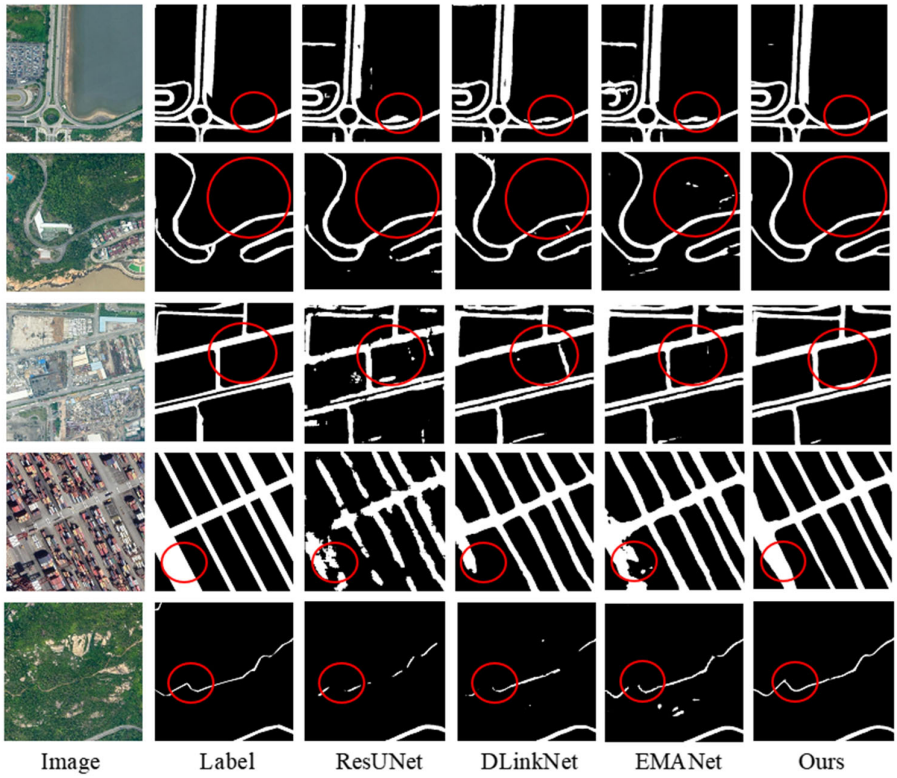

Figure 9. Qualitative evaluation results on the CHN6-CUG dataset. Each image covers an area of 256 m × 256 m.

4.6. Performance Analysis

In addition, this study also conducted performance analysis regarding the parameter size and computational complexity of the CR-HR-RoadNet model. Table 4 presents the efficiency analysis results of several convolutional neural network models. The Params and FLOPs of our proposed model are only 15.28 Mb and 248.90 Gbps, respectively, demonstrating a precision advantage without significantly increasing the number of parameters and computational load.

Table 4. Efficiency analysis results of selected models.

	UNet	**DeepLabV3+**	**DANet**	**ResUNet**	**DLinkNet**	**Ours**
Params (Mb)	13.40	59.44	54.36	13.04	198.89	15.28
FLOPs (Gbps)	124.36	90.35	313.70	323.73	129.96	248.90

Comparing our proposed model (Ours) with other popular models, it is evident that our model achieves competitive results in terms of parameter size and computational complexity. With only 15.28 Mb of parameters and 248.90 Gbps of FLOPs, the model strikes a balance between computational efficiency and accuracy.

The smaller number of parameters is beneficial for reducing model size, making it more lightweight and easier to deploy in resource-constrained environments. Moreover, the lower computational complexity (FLOPs) implies faster inference times and reduced energy consumption during model execution, which is essential for real-time applications and scenarios with limited computational resources.

4.7. Ablation Study

To further verify the role of the multi-scale feature representation module and coordinate attention module, we design the corresponding ablation experiments to analyze and verify the role of each module.

Table 5 shows the quantitative ablation experimental results of the proposed model on the DeepGlobe dataset. The quantitative comparison results show that the precision of the complete model is lower than that of the model that does not include the two modules, and the recall of the complete model is lower than that of the model that only contains the coordinate attention module. The proposed complete model can achieve the highest accuracy on the comprehensive indicators of F1 and IoU and has better comprehensive performance. Specifically, the proposed model only obtains the worst road extraction accuracy when it does not include the two modules. Meanwhile, the F1 is improved by 1.33%, and the IoU is improved by 1.74% when both modules are included, which shows that the proposed modules can play a better positive role and proves their necessity and effectiveness.

Table 5. Ablation experimental results of the CR-HR-ROADNET on the DeepGlobe dataset.

Multi-Scale Feature Representation Module	Coordinate Attention Module	Precision	Recall	F1	IoU
× [1]	×	**77.32** [3]	73.69	75.46	60.59
√ [2]	×	75.49	76.79	76.13	61.46
×	√	74.52	**77.46**	75.96	61.24
√	√	76.47	77.12	**76.79**	**62.33**

[1] The symbol "×" represents the non-use of the corresponding module. [2] The symbol "√" represents the use of the corresponding module. [3] The optimal accuracy values are in bold font.

Specifically, the proposed model can achieve an accuracy improvement of 0.67% on F1 and 0.87% on IoU when only the multi-scale feature representation module is included, thereby proving the importance of multi-scale local context for road extraction tasks. When only the coordinate attention module is included, the proposed model can achieve an accuracy improvement of 0.5% on F1 and 0.65% on IoU, which proves the importance of global context information for road extraction tasks. When the two modules are included, the magnitude of the accuracy improvement on F1 and IoU is higher than the sum of the individual improvements of the two modules. This condition may be due to the tight coupling of the two modules in the model; the multi-scale feature representation module is in each feature stage, and the coordinate attention module is between different feature stages. Hence, these two modules can influence and promote each other. The more effective features the multi-scale feature representation module extracts, the more effective global context reasoning the coordinate attention module performs. By contrast, the better optimization effect the coordinate attention module obtains, the more effective local context reasoning the multi-scale feature representation module performs.

Meanwhile, the accuracy improvement of the multi-scale feature representation module is better than that of the coordinate attention module, which may be because the multi-scale feature representation module is closely integrated with the feature learning process of the model. This module can extract effective multi-scale local context information and enhance the feature representation ability of the model by capturing the dependencies between the road and the background environment.

Figure 10 shows the ablation experimental results of the proposed model on the DeepGlobe dataset. The visualization results indicate that the best road extraction results can be achieved when the CR-HR-RoadNet model includes the multi-scale feature representation module and the coordinate attention module (namely, Model C). The extraction results of Model C have better completeness and continuity, smoother road boundaries, and less noise information compared with those of Models A and B. Specifically, some areas are marked by the red circle in Figure 10. Model C can achieve the best road extraction effect in these areas. The results in the first row show that Model C can effectively distinguish geo-objects similar to roads, thereby avoiding the problem of road misclassification. The results in the second row show that Model C can remove irrelevant noise information caused by the complex background environment. The results in the third row show that Model C can effectively handle the incomplete and discontinuous roads caused by complex backgrounds.

The results in the fourth row show that Model C can effectively handle the problem of road discontinuity that is due to occlusion. Overall, the qualitative experimental results fully demonstrate the effectiveness and necessity of the multi-scale feature representation module and coordinate attention module.

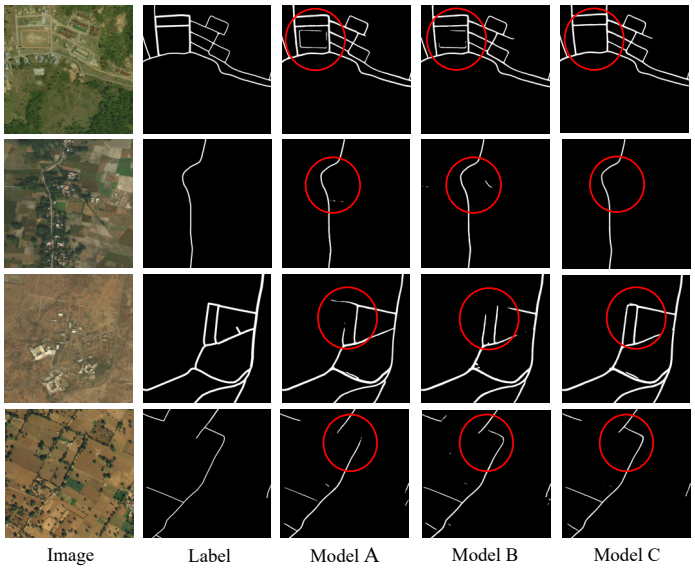

Figure 10. Ablation experimental results of the CR-HR-RoadNet on the DeepGlobe dataset. Model A represents the CR-HR-RoadNet containing only the multi-scale feature representation module; Model B represents the CR-HR-RoadNet containing only the coordinate attention module; and Model C represents the CR-HR-RoadNet containing both modules. Each image covers an area of 512 m $\times$ 512 m.

4.8. Comprehensive Analysis and Evaluation of Algorithmic Performance

In the analysis of the Massachusetts dataset, we undertook a systematic approach to achieve a more profound understanding of the distribution of algorithmic accuracy across the dataset. Additionally, we sought to explore the potential impact of stochastic factors present in the data on algorithmic outcomes. This endeavor involved conducting a comprehensive evaluation of the algorithm's stability and consistency when confronted with diverse data samples. Our aim was to achieve a more precise assessment of the algorithm's overall performance, thereby enhancing its reliability in practical applications. To accomplish this, we initiated the process by independently executing the algorithm for each individual sample within the dataset and subsequently calculating the accuracy of the algorithm's outcomes. This enabled us to obtain a baseline understanding of its performance. Building upon this, we conducted statistical analysis using techniques like box plots to analyze the distribution of accuracy values for each sample. We also conducted a thorough investigation to identify the potential sources of any outliers that might have influenced the results. Furthermore, we selected four distinct representative scenes from the Massachusetts dataset: urban arterial roads, urban residential area roads, forest pathways, and village roads. The results of road extraction in these four scenes are showcased to elucidate the algorithm's road extraction capabilities in various scenes.

Based on the statistical results presented in Figure 11, several significant observations have been drawn. Through an examination of quartiles, the relatively narrow interquartile range underscores the limited variability in performance among distinct images, showcasing the algorithm's stability and its capacity to maintain consistent outcomes when presented with diverse data samples. Concurrently, the proximity of the quartiles, medians, and means for all four metrics suggests a consistent trend within the dataset. This signifies that the algorithm generally attains reliable results across various scenarios.

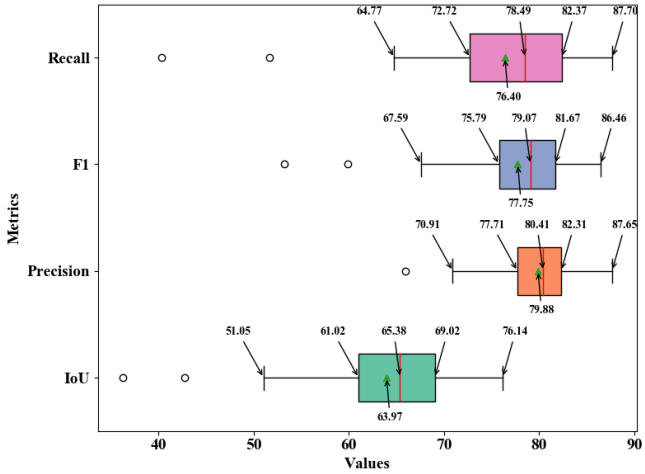

Figure 11. Statistical analysis of samples in Massachusetts dataset. The box segment represents the range between the first quartile (Q1) and the third quartile (Q3) of the data. The upper limit is calculated by adding 1.5 times the interquartile range (IQR) to the third quartile (Q3), where IQR is the difference between Q3 and Q1. The lower limit is calculated by subtracting 1.5 times the IQR from the first quartile (Q1). Values exceeding these limits are considered outliers. The black hollow points represent outliers. The red vertical line denotes the median, which is the middle value when the data is sorted. The green triangular points represent the mean.

Specifically, the complex road and occlusion areas are marked by red circles in the visualization results. While the algorithm demonstrates robustness across quartiles and similar statistical measures, the presence of outliers also indicates its sensitivity to data uncertainty and randomness. The samples represented by the accuracy outliers in Figure 12 reveal that although the algorithm excels in accurately locating road positions, it faces challenges in distinguishing lane quantities because of factors such as image clarity and algorithmic structure, as demonstrated by the area highlighted by the red circle. This difficulty results in the appearance of accuracy outliers.

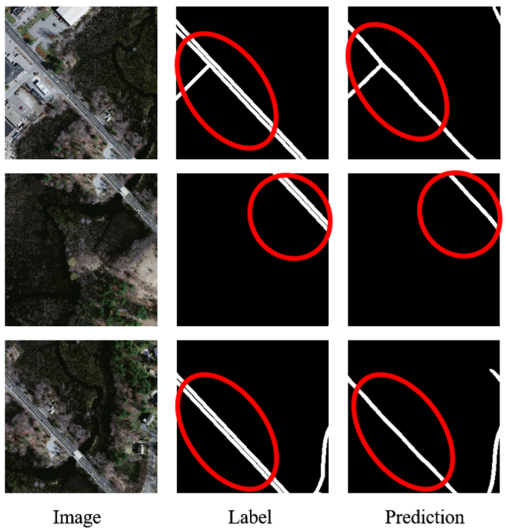

Figure 12. Samples corresponding to accuracy outliers. Each image covers an area of 307.2 m × 307.2 m.

Figure 13 illustrates the extraction results of the algorithm in four representative scenes. The algorithm's capability to ensure a comprehensive and uninterrupted depiction of roads remains robust across dissimilar scenarios. Urban arterial roads are typically wide and bustling with traffic, presenting complex and variable occlusions from vehicles and structures. In contrast, roadways within urban residential areas exhibit a relatively

dense architectural layout and diverse path trajectories. Conversely, within dense forested roadways, the occlusion from foliage often renders road recognition highly challenging. In the context of village road scenes, in comparison to urban areas, there is a heightened prevalence of obstructive elements such as vegetation and trees, leading to increased occlusion. Additionally, the road pathways tend to be narrower. These diverse contextual scenarios pose formidable challenges to our algorithm. The presented results underscore the algorithm's prowess in delivering accurate and coherent road extraction outcomes, underscoring its adaptability to a spectrum of scenarios.

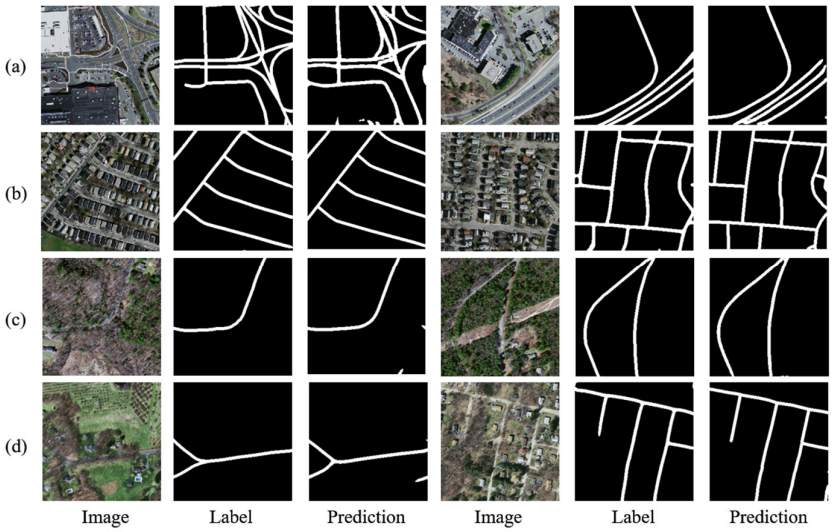

Figure 13. Algorithm extraction performance in different typical scenes: (**a**) represents urban arterial roads, (**b**) represents urban residential area roads, (**c**) represents forest pathways, and (**d**) represents village roads.

5. Conclusions

CR-HR-RoadNet employs a road-adapted high-resolution network as the core feature encoder and comprises two essential modules. The multi-scale feature representation module enhances the feature representation capacity of the neural network model by combining multi-scale information with feature learning, effectively capturing local context at various scales. Meanwhile, the coordinate attention module captures long-range dependencies and extracts vital global context information, significantly improving road feature representation in both spatial and channel dimensions.

Through comprehensive experiments on three diverse datasets, our proposed model has demonstrated remarkable extraction accuracy and strong anti-occlusion capabilities. The predicted results exhibit enhanced road completeness and continuity, validating the effectiveness and generalization of CR-HR-RoadNet. Ablation experiments further confirm the importance and necessity of the multi-scale feature representation module and the coordinate attention module.

Despite the overall success of CR-HR-RoadNet, we acknowledge that challenges may arise in handling certain complex occlusions, such as cases where roads are entirely obscured by dense and tall vegetation.

Furthermore, due to limitations within the dataset, our method's analysis has been primarily focused on regions concentrated in mid to low latitudes, with no exploration or discussion regarding higher latitude areas characterized by prolonged snow cover and more pronounced vegetation seasonality. Additionally, given the diverse developmental trajectories of individual cities, variations in road network construction and structure exist. However, this study has not specifically investigated the generalizability and applicability of the algorithm under various geographical factors affecting different road network configurations. As part of our future research, we will focus on exploring post-processing

optimization methods to recover occluded road information effectively. Simultaneously, we will establish datasets for higher latitude regions and areas with distinct road network structures to analyze and enhance the generalizability and applicability of our approach.

In conclusion, our work presents a promising approach to address the road occlusion problem in high-resolution remote sensing images using deep learning techniques. The proposed CR-HR-RoadNet shows considerable potential for advancing road extraction tasks in challenging environmental conditions, paving the way for further advancements in geospatial image analysis and understanding.

Author Contributions: Conceptualization, J.C. and H.W.; methodology, L.Y. and H.W.; software, L.Y. and H.W.; validation, J.Z. and G.S.; formal analysis, J.Z. and G.S.; investigation, J.Z., G.S. and X.D.; data curation, G.S. and X.D.; visualization, G.S. and X.D.; writing—original draft preparation, J.C., L.Y. and H.W.; writing—review and editing, J.C., M.D. and Y.S.; supervision, M.D.; project administration, Y.S.; funding acquisition, Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the National Key Research and Development Program of China Grant 2020YFA0713503 and the National Natural Science Foundation of China Grant 42071427.

Data Availability Statement: The Massachusetts dataset can be downloaded from https://www.cs.toronto.edu/~vmnih/data/ (accessed on 19 April 2023), the DeepGlobe dataset can be downloaded from http://deepglobe.org/challenge.html (accessed on 19 April 2023), and the CHN6-CUG dataset can be downloaded from http://cugurs5477.mikecrm.com/ZtMn5tR (accessed on 19 April 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chen, Z.; Wang, C.; Li, J.; Fan, W.; Du, J.; Zhong, B. Adaboost-like End-to-End multiple lightweight U-nets for road extraction from optical remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *100*, 102341. [CrossRef]
2. Shan, B.; Fang, Y. A Cross Entropy Based Deep Neural Network Model for Road Extraction from Satellite Images. *Entropy* **2020**, *22*, 535. [CrossRef] [PubMed]
3. Chen, S.-B.; Ji, Y.-X.; Tang, J.; Luo, B.; Wang, W.-Q.; Lv, K. DBRANet: Road Extraction by Dual-Branch Encoder and Regional Attention Decoder. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 3002905. [CrossRef]
4. Yang, K.; Yi, J.; Chen, A.; Liu, J.; Chen, W. ConDinet++: Full-Scale Fusion Network Based on Conditional Dilated Convolution to Extract Roads from Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 8015105. [CrossRef]
5. Lian, R.; Wang, W.; Mustafa, N.; Huang, L. Road Extraction Methods in High-Resolution Remote Sensing Images: A Comprehensive Review. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 5489–5507. [CrossRef]
6. Abdollahi, A.; Pradhan, B.; Shukla, N.; Chakraborty, S.; Alamri, A. Deep Learning Approaches Applied to Remote Sensing Datasets for Road Extraction: A State-of-the-Art Review. *Remote Sens.* **2020**, *12*, 1444. [CrossRef]
7. Ge, Z.; Zhao, Y.; Wang, J.; Wang, D.; Si, Q. Deep Feature-Review Transmit Network of Contour-Enhanced Road Extraction from Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 3001805. [CrossRef]
8. Shao, Z.; Zhou, Z.; Huang, X.; Zhang, Y. MRENet: Simultaneous Extraction of Road Surface and Road Centerline in Complex Urban Scenes from Very High-Resolution Images. *Remote Sens.* **2021**, *13*, 239. [CrossRef]
9. Shamsolmoali, P.; Zareapoor, M.; Zhou, H.; Wang, R.; Yang, J. Road Segmentation for Remote Sensing Images Using Adversarial Spatial Pyramid Networks. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 4673–4688. [CrossRef]
10. Xu, Y.; Chen, H.; Du, C.; Li, J. MSACon: Mining Spatial Attention-Based Contextual Information for Road Extraction. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5604317. [CrossRef]
11. Wang, S.; Mu, X.; Yang, D.; He, H.; Zhao, P. Road Extraction from Remote Sensing Images Using the Inner Convolution Integrated Encoder-Decoder Network and Directional Conditional Random Fields. *Remote Sens.* **2021**, *13*, 465. [CrossRef]
12. Lu, X.; Zhong, Y.; Zheng, Z.; Zhang, L. GAMSNet: Globally aware road detection network with multi-scale residual learning. *ISPRS J. Photogramm. Remote Sens.* **2021**, *175*, 340–352. [CrossRef]
13. Cira, C.-I.; Alcarria, R.; Manso-Callejo, M.-Á.; Serradilla, F. A Deep Learning-Based Solution for Large-Scale Extraction of the Secondary Road Network from High-Resolution Aerial Orthoimagery. *Appl. Sci.* **2020**, *10*, 7272. [CrossRef]
14. Wei, Y.; Zhang, K.; Ji, S. Simultaneous Road Surface and Centerline Extraction from Large-Scale Remote Sensing Images Using CNN-Based Segmentation and Tracing. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8919–8931. [CrossRef]
15. Wang, Y.; Seo, J.; Jeon, T. NL-LinkNet: Toward Lighter but More Accurate Road Extraction with Nonlocal Operations. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 3000105. [CrossRef]
16. Zhang, Z.; Liu, Q.; Wang, Y. Road Extraction by Deep Residual U-Net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]

17. Cheng, G.; Wang, Y.; Xu, S.; Wang, H.; Xiang, S.; Pan, C. Automatic Road Detection and Centerline Extraction via Cascaded End-to-End Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3322–3337. [CrossRef]
18. Liu, Y.; Yao, J.; Lu, X.; Xia, M.; Wang, X.; Liu, Y. RoadNet: Learning to Comprehensively Analyze Road Networks in Complex Urban Scenes from High-Resolution Remotely Sensed Images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 2043–2056. [CrossRef]
19. Zhou, M.; Sui, H.; Chen, S.; Wang, J.; Chen, X. BT-RoadNet: A boundary and topologically-aware neural network for road extraction from high-resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2020**, *168*, 288–306. [CrossRef]
20. Gao, X.; Sun, X.; Zhang, Y.; Yan, M.; Xu, G.; Sun, H.; Jiao, J.; Fu, K. An End-to-End Neural Network for Road Extraction from Remote Sensing Imagery by Multiple Feature Pyramid Network. *IEEE Access* **2018**, *6*, 39401–39414. [CrossRef]
21. Zhou, L.; Zhang, C.; Wu, M. D-LinkNet: LinkNet with Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 182–186.
22. Wu, Q.; Luo, F.; Wu, P.; Wang, B.; Yang, H.; Wu, Y. Automatic Road Extraction from High-Resolution Remote Sensing Images Using a Method Based on Densely Connected Spatial Feature-Enhanced Pyramid. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 3–17. [CrossRef]
23. Zhu, Q.; Zhang, Y.; Wang, L.; Zhong, Y.; Guan, Q.; Lu, X.; Zhang, L.; Li, D. A Global Context-aware and Batch-independent Network for road extraction from VHR satellite imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *175*, 353–365. [CrossRef]
24. Tan, X.; Xiao, Z.; Wan, Q.; Shao, W. Scale Sensitive Neural Network for Road Segmentation in High-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 533–537. [CrossRef]
25. Tao, C.; Qi, J.; Li, Y.; Wang, H.; Li, H. Spatial information inference net: Road extraction using road-specific contextual information. *ISPRS J. Photogramm. Remote Sens.* **2019**, *158*, 155–166. [CrossRef]
26. Lin, Y.; Xu, D.; Wang, N.; Shi, Z.; Chen, Q. Road Extraction from Very-High-Resolution Remote Sensing Images via a Nested SE-Deeplab Model. *Remote Sens.* **2020**, *12*, 2985. [CrossRef]
27. Xu, Y.; Xie, Z.; Feng, Y.; Chen, Z. Road Extraction from High-Resolution Remote Sensing Imagery Using Deep Learning. *Remote Sens.* **2018**, *10*, 1461. [CrossRef]
28. Li, J.; Liu, Y.; Zhang, Y.; Zhang, Y. Cascaded Attention DenseUNet (CADUNet) for Road Extraction from Very-High-Resolution Images. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 329. [CrossRef]
29. Ren, Y.; Yu, Y.; Guan, H. DA-CapsUNet: A Dual-Attention Capsule U-Net for Road Extraction from Remote Sensing Imagery. *Remote Sens.* **2020**, *12*, 2866. [CrossRef]
30. Ding, C.; Weng, L.; Xia, M.; Lin, H. Non-Local Feature Search Network for Building and Road Segmentation of Remote Sensing Image. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 245. [CrossRef]
31. Song, M.; Civco, D. Road Extraction Using SVM and Image Segmentation. *Photogramm. Eng. Remote Sens.* **2004**, *70*, 1365–1371. [CrossRef]
32. Jing, R.; Gong, Z.; Zhu, W.; Guan, H.; Zhao, W. Island Road Centerline Extraction Based on a Multiscale United Feature. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 3940–3953. [CrossRef]
33. Maboudi, M.; Amini, J.; Malihi, S.; Hahn, M. Integrating fuzzy object based image analysis and ant colony optimization for road extraction from remotely sensed images. *ISPRS J. Photogramm. Remote Sens.* **2018**, *138*, 151–163. [CrossRef]
34. Chen, L.; Zhu, Q.; Xie, X.; Hu, H.; Zeng, H. Road Extraction from VHR Remote-Sensing Imagery via Object Segmentation Constrained by Gabor Features. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 362. [CrossRef]
35. Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef] [PubMed]
36. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
37. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Springer: Cham, Switzerland, 2015; pp. 234–241.
38. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Computer Vision—ECCV 2018*; Springer: Cham, Switzerland, 2018; pp. 833–851. [CrossRef]
39. Yang, X.; Li, X.; Ye, Y.; Lau, R.Y.K.; Zhang, X.; Huang, X. Road Detection and Centerline Extraction Via Deep Recurrent Convolutional Neural Network U-Net. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 7209–7220. [CrossRef]
40. Ding, L.; Bruzzone, L. DiResNet: Direction-Aware Residual Network for Road Extraction in VHR Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 10243–10254. [CrossRef]
41. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 3349–3364. [CrossRef]
42. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
43. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *arXiv* **2014**, arXiv:14127062v4.
44. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef]

45. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:170605587v3.
46. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. DenseASPP for Semantic Segmentation in Street Scenes. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 3684–3692.
47. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 13708–13717.
48. Mnih, V. *Machine Learning for Aerial Image Labeling*; University of Toronto: Toronto, ON, Canada, 2013; Available online: https://www.zhangqiaokeyan.com/academic-degree-foreign_mphd_thesis/02061189498.html (accessed on 3 March 2022).
49. Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; Raskar, R. DeepGlobe 2018: A Challenge to Parse the Earth through Satellite Images. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 172–181.
50. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef] [PubMed]
51. Yuan, Y.; Huang, L.; Guo, J.; Zhang, C.; Chen, X.; Wang, J. OCNet: Object Context Network for Scene Parsing. *arXiv* **2018**, arXiv:180900916v4.
52. Li, X.; Zhong, Z.; Wu, J.; Yang, Y.; Lin, Z.; Liu, H. Expectation-Maximization Attention Networks for Semantic Segmentation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9166–9175.
53. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3141–3149.