remote sensing

MDPI

Article

Road Extraction Convolutional Neural Network with Embedded Attention Mechanism for Remote Sensing Imagery

Shiwei Shao [1,2,*], Lixia Xiao [3,4], Liupeng Lin [5], Chang Ren [6] and Jing Tian [5]

[1] National Research Center of Cultural Industries, Central China Normal University, Wuhan 430056, China
[2] Zhongzhi Software Technology Co., Ltd., Wuhan 430013, China
[3] State Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China; xlx@zrzyhgh.wuhan.gov.cn
[4] Wuhan Natural Resources and Planning Information Center, Wuhan 430014, China
[5] School of Resource and Environment Science, Wuhan University, Wuhan 430079, China; linliupeng@whu.edu.cn (L.L.); tianjing_sres@whu.edu.cn (J.T.)
[6] College of Air Traffic Management, Civil Aviation Flight University of China, Guanghan 618307, China; imrc@whu.edu.cn
* Correspondence: nrcci@ccnu.edu.cn

Abstract: Roads are closely related to people's lives, and road network extraction has become one of the most important remote sensing tasks. This study aimed to propose a road extraction network with an embedded attention mechanism to solve the problem of automatic extraction of road networks from a large number of remote sensing images. Channel attention mechanism and spatial attention mechanism were introduced to enhance the use of spectral information and spatial information based on the U-Net framework. Moreover, residual densely connected blocks were introduced to enhance feature reuse and information flow transfer, and a residual dilated convolution module was introduced to extract road network information at different scales. The experimental results showed that the method proposed in this study outperformed the compared algorithms in overall accuracy. This method had fewer false detections, and the extracted roads were closer to ground truth. Ablation experiments showed that the proposed modules could effectively improve road extraction accuracy.

Keywords: road extraction; U-Net; attention mechanism; residual densely connected blocks; dilated convolution

check for updates

Citation: Shao, S.; Xiao, L.; Lin, L.; Ren, C.; Tian, J. Road Extraction Convolutional Neural Network with Embedded Attention Mechanism for Remote Sensing Imagery. *Remote Sens.* **2022**, *14*, 2061. https://doi.org/10.3390/rs14092061

Academic Editor: Lionel Bombrun

Received: 22 February 2022
Accepted: 20 April 2022
Published: 25 April 2022

1. Introduction

Road network plays an important role in real-world applications, such as intelligent city and intelligent traffic management [1,2]. As an important means of obtaining maps of road networks, road extraction has received extensive attention from scholars in this field. At present, road extraction from images is a basic research task of remote sensing. Compared with the traditional visual interpretation of remote sensing images, the automatic road interpretation algorithm of remote sensing images can obtain road network information quickly and at a low cost. However, their manifestations in remote sensing images are significantly different due to the large differences in shape, width, length, and material of road networks in different regions. Therefore, accurate extraction of road network information from remote sensing images is still a challenging problem.

So far, a series of road extraction algorithms have been developed, which can be mainly divided into two categories: traditional road extraction algorithms and deep learning-based road extraction algorithms.

In the traditional road extraction algorithm, some scholars used a template matching algorithm for semi-automatic extraction, including a rectangular template matching algorithm [3], a circular template matching algorithm [4], and a T-shaped template matching algorithm [5]. This type of semi-automatic template matching algorithm had high requirements for the adaptive ability of the template, and had defects such as inaccurate extraction and unsmooth roads for complex road networks with frequent changes in curvature and width.

Some scholars focused on the positive effect of spectral features of remote sensing images on extraction accuracy. Shi et al. [6] combined the spectral information, spatial information, and shape features of remote sensing images to obtain a road binary map. Coulibaly et al. [7] combined the spectral angle algorithm with Lowe's scale-invariant features transform descriptors to achieve high-quality extraction of road networks. Vector data were also introduced into the road network automatic extraction algorithm to assist road extraction and improve the extraction accuracy. Cao et al. [8] introduced GPS data in the process of road centerline extraction. Manandhar et al. [9] introduced volunteer geographic information to assist road network extraction.

In traditional road extraction methods, satisfactory results have been achieved. However, since traditional methods need to manually design road feature extraction algorithms and adjust the threshold parameters, they are not suitable for road extraction from large-scale and complex remote sensing data.

Some scholars used machine learning algorithms to quickly extract road networks from remote sensing images to solve the aforementioned problems. Mokhtarzade et al. [10] used BP neural networks with a different number of iterations of different hidden layer sizes for road extraction. M. Song et al. [11] used the support vector machine method for road extraction. Maurya et al. [12] first removed non-road areas based on morphological features, and then extracted road areas by K-means clustering. Seppke et al. [13] proposed a parallel super-pixel-based road tracking method combining geometric and topological representations. In the analysis framework of adaptive mean shift, Huang et al. [14] used mean shift to obtain an object-oriented representation of hyperspectral data, and then used a support vector machine for feature interpretation.

Recently, deep learning is widely used in remote sensing and has shown good performance for image classification [15–17] and segmentation tasks [18–22]. Wei et al. [23] proposed to use a patch-based convolutional neural network for road extraction. Alshehhi et al. [24] further developed a neural network that can extract roads and buildings simultaneously. Zhong et al. [25] proposed to use a fully convolutional neural network for road extraction. Based on the fully convolutional neural network, a series of encoder-decoder neural networks are developed. Panboonyuen et al. [26] used the SegNet structure as the benchmark network for road segmentation, and used deconvolution to replace the residual up-sample of the fully convolutional neural network. Zhang et al. [27] combined the advantages of U-Net structure to extract scale information with the advantage of residual structure to connect different features maps. To make the network easier to train, Xu et al. [28] combined the U-Net structure with a densely connected network. In 2017, LinkNet was proposed by Chaurasia et al. [29]. Zhou et al. [30] used LinkNet structure and dilated convolution to segment roads and became the best solution in DeepGlobe-2018. Subsequently, a variety of improved methods based on the U-Net structure are proposed. He et al. [31] added a pyramid pooling structure to the U-Net structure, and used a structural similarity loss function to constrain the network. Wulamu et al. [32] extracted features for roads at different scales through different convolutional layers.

Besides, some scholars focused on the image of the loss function on the road extraction accuracy. He et al. [31] used a structural similarity loss function to constrain the network training process to enhance the detailed information of the road extraction results and avoid over-smoothing the extraction results. Wei et al. [23] used road structure information to adjust the cross-entropy loss function. Mosinska et al. [33] proposed a topological-aware loss function to solve the problem of abnormal road network topology in road extraction results, which effectively reduced road network topological errors. These loss functions only considered strengthening the use of road network structural information and do not fully consider the structural differences of road training samples [34].

In recent years, as a way to recalibrate feature maps, the attention mechanism is widely used in the field of computer vision [35–40]. The attention mechanism can give a higher weight to the discriminative features and reduce the weight of unnecessary information, effectively improving the recognition accuracy.

Therefore, we propose a Road Extraction convolutional neural Network with an embedded Attention mechanism (RENA) for remote sensing images. In the proposed network, we embed the spatial attention mechanism and the channel attention mechanism based on the U-Net. The spatial attention mechanism is used to improve the spatial detail information of road extraction results, and the channel attention mechanism is used to recalibrate the spectral features of remote sensing images. Under this U-Net framework, residual densely connected blocks are also embedded [41] to integrate the information flow transmission advantages of residual densely connected blocks. Besides, we also introduce a residual dilated convolution module to enhance the ability to extract multiscale information.

The main contributions of this study are summarized as follows:

1. This study uses the U-Net structure to achieve end-to-end road network extraction from remote sensing images.
2. This study designs an attention module that combines spatial attention and channel attention to enhance spatial detail information and promote the use of spectral features.
3. Based on the U-Net structure, this study embeds the residual dense connection blocks to achieve information flow transfer and feature reuse, and uses a residual dilated convolution module to achieve multiscale information extraction.

The rest of the paper is organized as follows. Section 2 of the paper presents the details of the proposed method. Section 3 of the paper is the experimental configurations and experimental results including the ablation experiments. Section 4 is the conclusion part of the paper.

2. Methodology

2.1. U-Net Framework with Embedded Attention Mechanism

U-Net framework is a network of encoding and decoding architecture, which is widely used in remote sensing images semantic segmentation [42–45] and classification tasks [46].

The U-Net framework consists of two parts, namely the encoder layer and the decoder layer. The encoder is used for pixel-by-pixel semantic feature extraction of remote sensing images, and the context information of the image is extracted by it. The decoder layer is used to decode the feature maps and locate the region of interest (ROI) in the image to finally obtain semantic segmentation results. In the U-Net framework, in order to effectively transfer the information of low-level feature maps to the high-level feature maps, the feature maps extracted by the encoder layer combine with the feature maps extracted by the decoder layer in the form of a skip connection to achieve effective fusion of low-level feature maps and high-level feature maps.

The attention mechanism is mainly used to recalibrate the weights of the feature maps. At present, there are three main categories of attention mechanisms used in the field of remote sensing, including spatial attention mechanisms, channel attention mechanisms, and temporal attention mechanisms. For the spatial attention mechanism, it is a way to recalibrate the spatial information of remote sensing images. The spatial attention mechanism reduces the dimension of the feature maps to two dimensions by performing feature compression on the feature maps, and obtains the two-dimensional weight through the normalization function to guide the feature maps for spatial recalibration, so that high-frequency information gets higher weight. For the channel attention mechanism, it is to recalibrate the weight of the image channel direction. The channel attention mechanism obtains the global feature of the image channel direction through the global pooling operation, and uses it as the weight of the feature maps to guide the feature maps to recalibrate the channel direction. Make discriminative features get higher weights. In this paper, we extract the global mean weight and global maximum weight of the image through mean pooling and max pooling, respectively. The temporal attention mechanisms are a way of empowering temporal information and are mainly used to process multi-temporal remote sensing data. In this paper, multi-temporal data is not used, therefore, only spatial attention mechanism and channel attention mechanism are used.

In this paper, the U-Net framework is used as the main frame. The RGB three-channel remote sensing image is adopted as the input data of the network, and the binary image road network map is obtained through end-to-end network processing. Among them, the network output results are two types of results, road, and non-road. The U-Net framework includes four encoder layers and four decoder layers. The four encoder layers are used to extract image features of different scales, and the four decoder layers are used to restore image details, and realize the information flow transmission and the reuse of different levels of feature maps by means of skip connections. The spatial attention block and the channel attention block are combined in a cascaded form and jointly embedded into the U-Net framework, which enhances the utilization of spectral information while improving spatial details. Residual dense connection blocks are embedded in each encoder layer and decoder layer, which are used to connect feature information at different levels to achieve efficient transmission of information flow. In addition, the residual dilated convolution module is added to the last encoder layer to further expand the receptive field. The flowchart of the proposed method is listed in Figure 1.

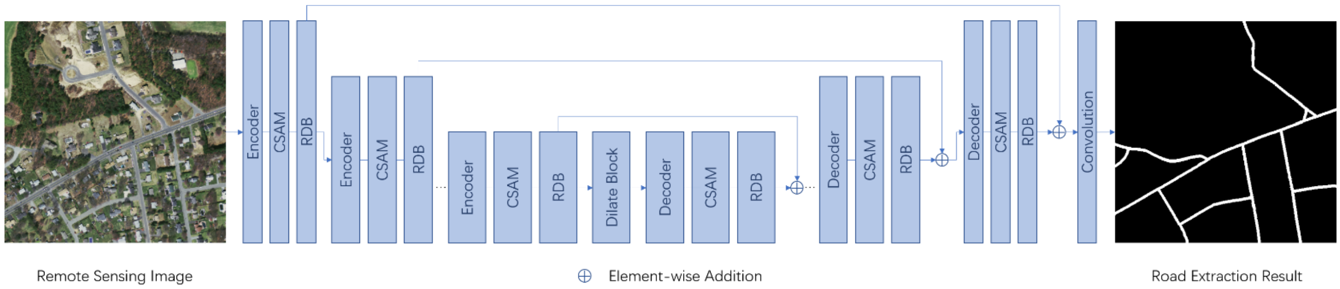


Figure 1. Schematic diagram of the proposed road extraction network.

2.2. U-Net Framework

In the proposed U-Net road extraction framework, we use four encoder layers and four decoder layers, and add the feature maps obtained by the corresponding encoder layers and decoder layers through skip connections.

For the four encoder layers, we adopt the pre-trained model of the ResNet34 network to speed up the model convergence. In each encoder layer, there are multiple residual blocks (ResBlock). For each ResBlock, it consists of two convolutional layers, two batch normalization layers, and a rectified linear unit (ReLU). Among them, the convolution layer is used to extract remote sensing image feature information. The batch normalization layer is used to normalize the data distribution to speed up the convergence of network training, avoid gradient disappearance or gradient explosion, and at the same time, avoid overfitting of the trained model. The ReLU is used to perform nonlinear processing on the data to enhance the model's ability to fit nonlinear relationships. In the encoder layer, the first ResBlock of the second to fourth layers of it also includes a down sampling process, which down-samples the feature maps and increases the number of feature maps channels. In the U-Net framework, there are four decoder layers. In each decoder layer, the structure is roughly the same as that of the encoder layer, including a transposed convolution layer, three batch normalization layers, and three ReLU activation functions. Among them, the transposed convolution is a parameter learnable up sampling operation on the feature maps.

2.3. Channel and Spatial Attention Module (CSAM)

In the U-Net framework, after processing through the encoder layer, the size of the feature maps becomes smaller, and the number of channels increases. Conversely, after being processed by the decoder layer, the number of channels in the feature maps decreases, the size becomes larger, and the spatial information increases. Therefore, in order to maintain the channel information and spatial information of the encoder layer and the decoder layer, we embed the channel and spatial attention modules (Figure 2) after it.

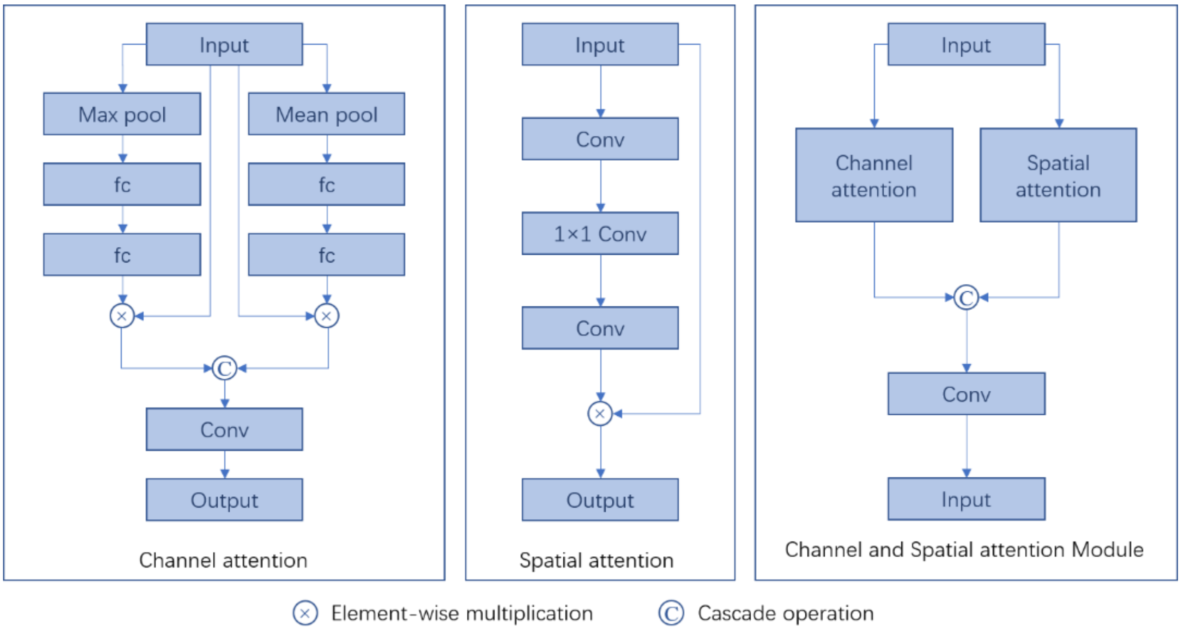


Figure 2. The CSAM schematic.

The channel attention block (CAB) includes pooling layers, fully connected layers, and convolutional layers. Through the maximum value pooling and the mean value pooling, the maximum value and the mean value features are extracted from the feature maps respectively. Then, the maximum value feature and the mean value feature are respectively converted into a one-dimensional vector through the fully connected layer, and the one-dimensional vector is normalized by the sigmoid function to obtain the maximum channel direction weight and the mean channel direction weight. Subsequently, the element-wise multiplication is performed on the two channel weights and the input feature maps to obtain the recalibrated feature maps respectively. Finally, the dimensionality reduction is performed on the result of concatenating two recalibrated feature maps and using convolutional layers.

The channel attention block can be expressed in the following form:

$$fm_{ca} = cat(fm_{input} \otimes \xi(w_{ca} \circ p_{avg}(fm_{input}) + b_{ca}), fm_{input} \otimes \xi(w_{ca} \circ p_{\max}(fm_{input}) + b_{ca})) \tag{1}$$

Among them, fm_{ca} represents the feature maps processed by the channel attention block, fm_{input} represents the input feature maps, w_{ca} represents the convolution kernel of the channel attention block, b_{ca} represents the bias term of the channel attention block, and $\xi(.)$ represents the sigmoid function. $\circ$ represents the convolution operator. and $\otimes$ represents the element-wise multiplication operation. $p_{avg}(.)$ represents the mean value pooling operator, $p_{\max}(.)$ represents the maximum value pooling operator, and $cat(.)$ represents the map cascade operation.

The spatial attention block (SAB) includes convolutional layers and 1 × 1 convolutional layers. The feature maps are extracted through the convolution layer, and then the feature maps are reduced to single-channel features through the 1 × 1 convolution layer, and the single-channel feature map is normalized through the sigmoid function to obtain the spatial weight. Then, element-wise multiplication is performed on the spatial weights and the input feature maps to obtain the recalibrated feature maps. Finally, a convolutional layer is used to perform convolution operation on the recalibrated feature maps.

The spatial attention block can be expressed in the following form:

$$fm_{sa} = fm_{input} \odot \xi\left(w_{sa} \circ fm_{input} + b_{sa}\right) \tag{2}$$

where fm_{sa} represents the feature maps processed by the spatial attention block, fm_{input} represents the input feature maps, w_{sa} represents the convolution kernel of spatial attention block, b_{sa} represents the bias term of spatial attention block

After being processed by the channel attention block and the spatial attention block respectively, the feature maps obtained by the two modules are fused in a cascaded manner, and 1 × 1 convolutional layers are used for dimensionality reduction processing to obtain the final feature map.

The fusion process of channel attention and spatial attention can be expressed as:

$$fm_{fuse} = \delta\left(w_{fuse} \circ cat(fm_{ca}, fm_{sa}) + b_{fuse}\right) \tag{3}$$

where fm_{fuse} represents the fused feature maps, w_{fuse} represents the convolution kernel of spatial attention block, b_{fuse} represents the bias term of spatial attention block, $\delta(.)$ and the ReLU function.

2.4. Residual Densely Connected Blocks (RDCB)

To address the information flow transfer problem, Huang et al., proposed densely connected blocks (DB) [41]. Densely connected blocks have been widely used in the field of computer vision due to their powerful information transfer and information extraction capabilities [41,47]. By using multiple skip connections, the low-level feature maps are connected with the high-level feature maps, while maintaining the characteristics of the feedforward layer, the low-level feature maps information is effectively transferred to the high-level feature maps. In the encoder layer and the channel attention block, the feature maps become smaller after being processed by the encoder layer, and the spatial information of the feature maps is further compressed after being processed by the channel attention block. In order to further utilize the information of the input feature maps, we use residual densely connected blocks to calculate the residual between the input feature map and the DB-processed feature map through the residual structure. The residual densely connected block is shown in Figure 3.

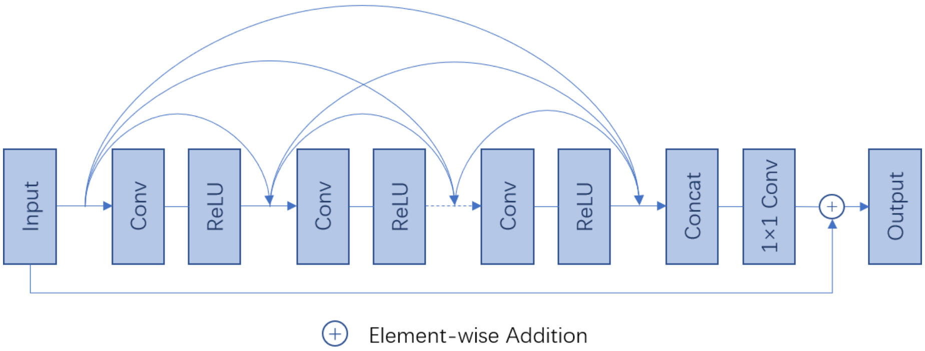


Figure 3. The RDCB schematic.

In the proposed network, in order to preserve the spatial features of the feature map, RDB is used for feature extraction on feature maps. As shown in Figure 1, after using CSAM to recalibrate the features of the encoder layer and the decoder layer, we use RDB to perform feature extraction on the recalibrated feature maps to extract information at different scales.

2.5. Residual Dilated Convolution Module (RDCM)

Dilated convolutional layers are widely used to extract features under different receptive fields to extract features at different scales. A larger receptive field is more favorable for extracting large objects, and a small receptive field can extract the feature information of small objects. The dilated convolution does not reduce the spatial resolution of the image, and at the same time, by setting the dilation rate, it realizes the operation of expanding the receptive field to different degrees.

In this paper, after all of the encoder layer feature extraction, we use a residual dilated convolution module to extract the feature information of different scale targets as shown in Figure 4. The residual dilated convolution module mainly includes four dilated convolution layers with dilation rates of 1, 2, 4, and 8, which are used to extract feature information under four receptive fields and four rectified linear units. In this module, the residual structure is used to perform element addition operations on the extracted four kinds of feature information, so as to fully integrate the feature information at different scales.

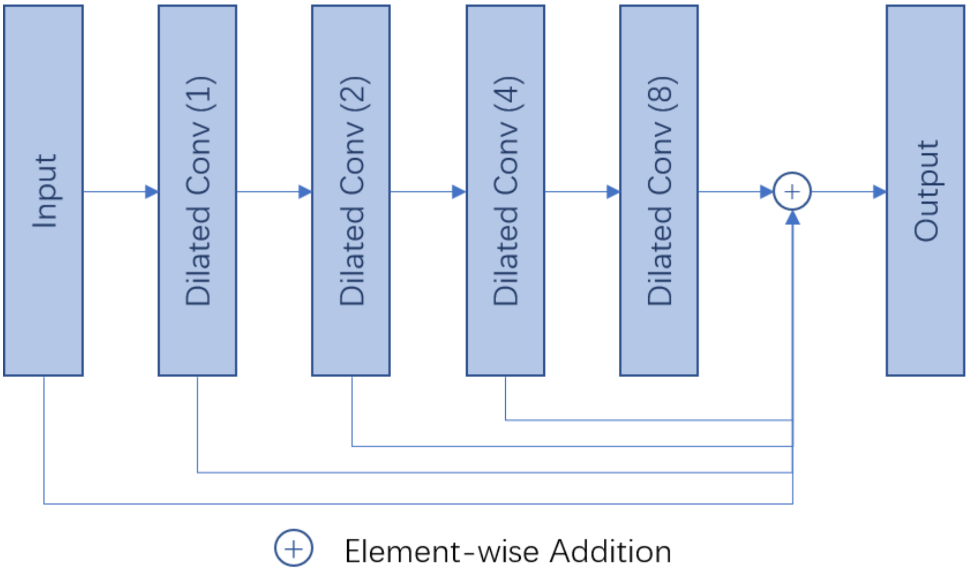


Figure 4. The RDCM schematic.

The residual dilated convolution module can be defined as:

$$fm_{rdcm1} = f_{rdcm}(fm_{input}, dilation = 1) \tag{4}$$

$$fm_{rdcm2} = f_{rdcm}(fm_{rdcm1}, dilation = 2) \tag{5}$$

$$fm_{rdcm3} = f_{rdcm}(fm_{rdcm}, dilation = 4) \tag{6}$$

$$fm_{rdcm4} = f_{rdcm}(fm_{rdcm3}, dilation = 8) \tag{7}$$

$$fm_{rdcm} = fm_{rdcm1} + fm_{rdcm2} + fm_{rdcm3} + fm_{rdcm4} \tag{8}$$

Among them, fm_{input} represents the initial output feature maps, $f_{rdcm}(.)$ represents the dilated convolution layer. $dilation = 1, 2, 4, 8$ represent the dilation rates of 1, 2, 4, and 8, respectively. fm_{rdcm1}, fm_{rdcm2}, fm_{rdcm3}, and fm_{rdcm4} represent the output feature maps at different dilation rates, respectively. fm_{rdcm} represents the final output of the residual dilated convolution module.

As shown in Figure 5, we list the schematic diagrams of dilated convolution layers with dilation rates of 1 and 2, respectively.

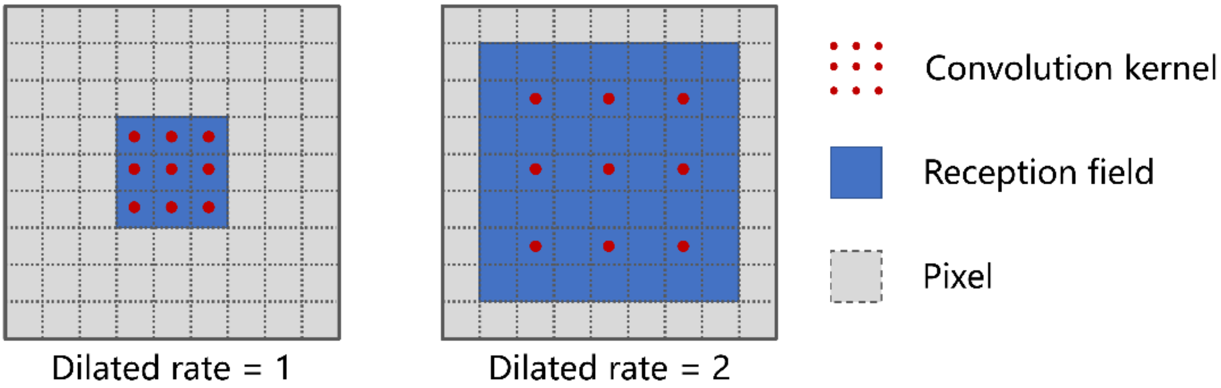


Figure 5. The dilated convolution layers.

3. Experiments and Analysis

3.1. Experimental Dataset Information

To verify the effectiveness of the proposed method, we conduct comparative experiments using the Deep Globe Road Extraction dataset. The Deep Globe Road Extraction dataset is the dataset used in the CVPR Deep Globe 2018 road extraction challenge. The dataset includes 6226 training images, 1243 verification images, and 1101 testing images. The size of each image is 1024 × 1024, and the image resolution is 0.5 m.

3.2. Road Extraction Network Training Configuration Information

In the process of network training in this paper, two loss functions are used to constrain, which are binary cross entropy and dice coefficient loss, respectively.

As shown in Table 1, the parameter configuration information of the main modules of the proposed network is listed.

Table 1. Parameter configuration of the proposed network main module.

Module Name	Module Composition	Parameter Configuration
Channel and Spatial Attention Module	channel attention block	See the configuration information of the channel attention block below for details
	spatial attention block	See the configuration information of the spatial attention block below for details.
	convolutional layer	(input channel = nFeat × 2, output channel = nFeat, kernel size = 3 × 3, stride = 1, pad = 1), ReLU
Channel Attention Block	pooling layer	average pooling layer: Adaptive Average Pooling
	pooling layer	max pooling layer: Adaptive Max Pooling
	fully connected layer	layer 1: (input channel = nFeat, output channel = nFeat/16, kernel size = 1 × 1, stride = 1, pad = 0), ReLU
	fully connected layer	layer 2: (input channel = nFeat/16, output channel = nFeat, kernel size = 1 × 1, stride = 1, pad = 0), Sigmoid
	fully connected layer	layer 3: (input channel = nFeat, output channel = nFeat/16, kernel size = 1 × 1, stride = 1, pad = 0), ReLU
	fully connected layer	layer 4: (input channel = nFeat/16, output channel = nFeat, kernel size = 1 × 1, stride = 1, pad = 0), Sigmoid
	convolutional layer	layer 1: (input channel = nFeat × 2, output channel = nFeat, kernel size = 3 × 3, stride = 1, pad = 1), ReLU
Spatial Attention Block	convolutional layer	layer1: (input channel = nFeat, output channel = nFeat, kernel size = 3 × 3, stride = 1, pad = 1), ReLU
	convolutional layer	layer2: (input channel = nFeat, output channel = 1, kernel size = 1 × 1, stride = 1, pad = 0), Sigmoid
	convolutional layer	layer3: (input channel = nFeat, output channel = nFeat, kernel size = 3 × 3, stride = 1, pad = 1), ReLU

Table 1. *Cont.*

Module Name	Module Composition	Parameter Configuration
ResidualDilated Convolution Module	Dilated convolutional layer	layer 1: (input channel = nFeat, output channel = nFeat, kernel size = 3 × 3, stride = 1, pad = 1, dilation = 1), ReLU
	Dilated convolutional layer	layer 2: (input channel = nFeat, output channel = nFeat, kernel size = 3 × 3, stride = 1, pad = 1, dilation = 2), ReLU
	Dilated convolutional layer	layer 3: (input channel = nFeat, output channel = nFeat, kernel size = 3 × 3, stride = 1, pad = 1, dilation = 4), ReLU
	Dilated convolutional layer	layer 4: (input channel = nFeat, output channel = nFeat, kernel size = 3 × 3, stride = 1, pad = 1, dilation = 8), ReLU
Residual Densely Connected Blocks	convolutional layer	(input channel = nFeat, output channel = nFeat/2, kernel size = 3 × 3, stride = 1, pad = 1), ReLU
	1 × 1 convolutional layer	(input channel = nFeat/2, output channel = nFeat, kernel size = 1 × 1, stride = 1, pad = 0)

nFeat is 64, 128, 256, 512 after processing by different encoder layers, and corresponding to 256, 128, 64 in different decoder layers.

The PyTorch framework is used for network training. The ADAM is used as the network optimizer. The network learning rate is 1×10^{-3}, the network batch size is 4, and the training epoch is 40. In the model training process, PyTorch v1.8 in the Windows environment is used, the CPU used is AMD 5600X@4.5GHz, and the GPU used is NVIDIA RTX3090 with 24 G memory. The training process took 40 h.

3.3. Comparison Algorithms and Quantitative Evaluation Metrics

To better verify the effect of the proposed method, we selected three mainstream remote sensing image road network extraction algorithms for comparison, including the U-Net method [30], LinkNet34 method [29], and D-LinkNet [30], which is the best solution in DeepGlobe-2018 and the state-of-the-art method. Among them, all the comparison algorithms were retrained with the same data and training epochs as in this study.

In quantitative evaluation experiments, five mainstream quantitative metrics are used to evaluate the performance of road extraction algorithms, which include accuracy, precision and recall, F1 score, and IoU (Intersection over Union) [48].

The accuracy is used to calculate the proportion of all correctly predicted samples to the total samples. It is an overall performance evaluation indicator, which is defined as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{9}$$

where TP is true positive, representing the number of samples that predict positive samples as positive samples. FP is false positive, representing the number of samples that predict negative samples as positive samples. TN is true negative, representing the sample that predicts negative samples as negative samples. FN is false negative, representing the number of samples that predict positive samples as negative samples.

The precision is the proportion of correctly predicted positive samples to all predicted positive samples, which is defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

The recall is the proportion of samples that are predicted to be positive among the samples that are actually positive, which is defined as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

The ranges of accuracy, precision, and recall are 0–1, and the higher the value, the better the performance of the two-class model.

F1 score is an indicator used in statistics to measure the performance of the binary classification model. It is a harmonic average indicator of the precision rate and recall rate of the model, which is used to balance the precision rate and the recall rate. Its value range is 0–1, and the higher the value, the better the performance of the binary classification model. The F1 score is defined as follows:

$$\mathrm{F1}-\mathrm{score}=\frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{12}$$

IoU is a standard that measures the accuracy of detecting corresponding objects in a specific dataset. It is used to measure the correlation between the true value and the predicted value. The higher the correlation, the higher the IoU value. The IoU is defined as follows:

$$\mathrm{loU}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{13}$$

3.4. Visual Evaluation Results

The effect of the proposed method was verified in the visual evaluation. According to the density of the road network, visual evaluation from two aspects, dense road network, and sparse road network, was conducted to verify the extraction ability of the proposed method in densely built-up areas and open areas, respectively.

3.4.1. Visual Evaluation Results for Dense Road Network

Three samples were selected for analysis in the densely built-up area, as shown in Figure 6. The results using LinkNet34 and D-LinkNet showed that all roads were effectively identified, but in these two methods, a large number of false detection samples, that is, the nonroad areas, were identified as road areas, and the result roads obtained by these two methods were wider than ground truth. U-Net and the proposed method could obtain similar results as ground truth with relatively few false detections, but could not identify areas occluded by trees. In the lower-left corner of Figure 6, U-Net and the proposed method had broken roads. From the remote sensing images, it was seen that trees were present in this area. In addition, the U-Net framework extraction results showed multiple fine false detection roads.

As shown in Figure 7, the three comparison algorithms had different degrees of false detection in the upper left corner of Figure 7, and the proposed method had fewer false detections. In terms of missed detection, LinkNet34 and D-LinkNet missed relatively few cases, while U-Net and the proposed method were relatively more.

In Figure 8, D-LinkNet had many false detections in the large-scale road network above the figure, and some buildings were identified as roads. The LinkNet34 method could identify roads more completely, but many false detections still existed. Moreover, the identified road width was wider than ground truth. The proposed method had few false detections, but some missed detections existed. The proposed method was closer to ground truth in terms of road shape and road width.

3.4.2. Visual Evaluation Results for Sparse Road Network

Another three samples in the open area were selected for analysis. As shown in Figure 9, U-Net, LinkNet34, and RENA methods could extract roads effectively, while D-LinkNet could not extract complete roads. In addition, the roads extracted using U-Net, and RENA methods were closer to ground truth in morphology, while LinkNet34 had false detections. As shown in Figure 10, compared with ground truth, U-Net, LinkNet34, and D-LinkNet had different degrees of false detection, while RENA had no false detection. As shown in Figure 11, the remote sensing image showed that there were two parallel roads. The U-Net and RENA methods could effectively extract the road and divide it into two roads, while LinkNet34 and D-LinkNet recognized the two parallel roads as a single road. In addition, LinkNet34 and D-LinkNet had false detections in the upper left corner.

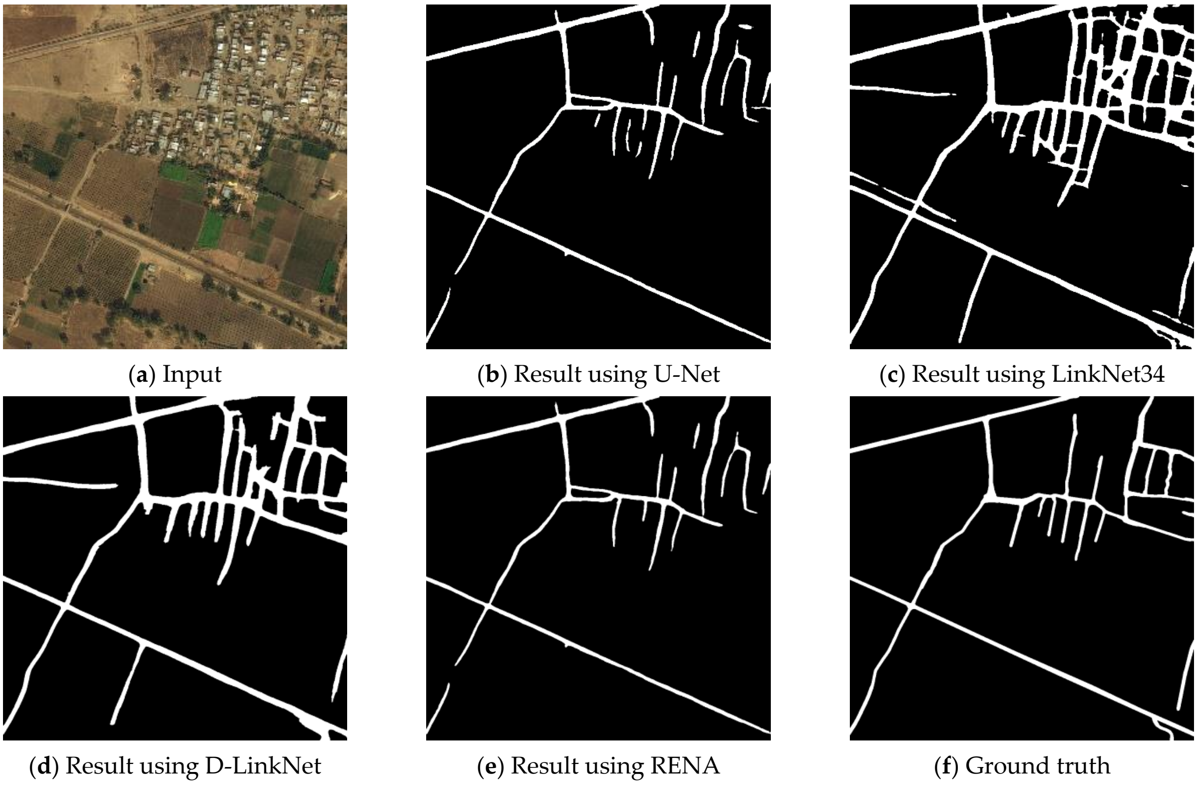

(**a**) Input (**b**) Result using U-Net (**c**) Result using LinkNet34
(**d**) Result using D-LinkNet (**e**) Result using RENA (**f**) Ground truth

Figure 6. Visual evaluation results of the first group of dense road networks.

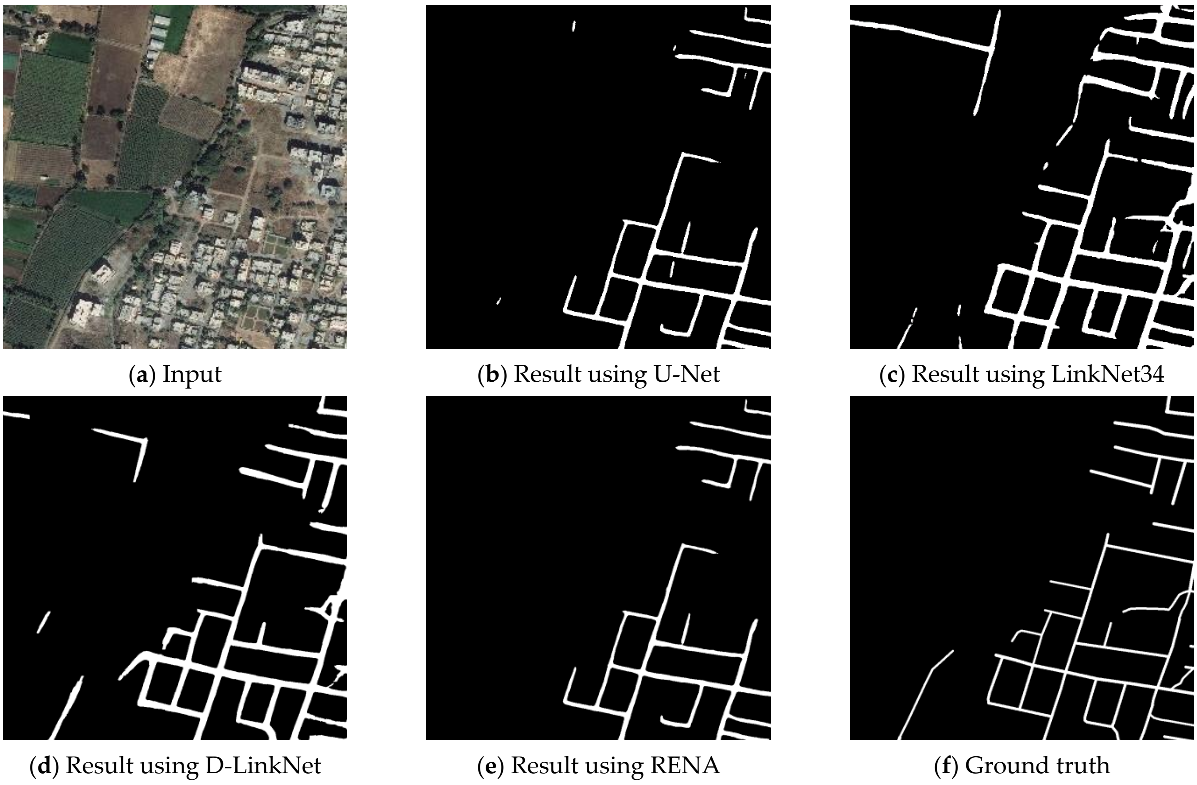

(**a**) Input (**b**) Result using U-Net (**c**) Result using LinkNet34
(**d**) Result using D-LinkNet (**e**) Result using RENA (**f**) Ground truth

Figure 7. Visual evaluation results of the second group of dense road networks.

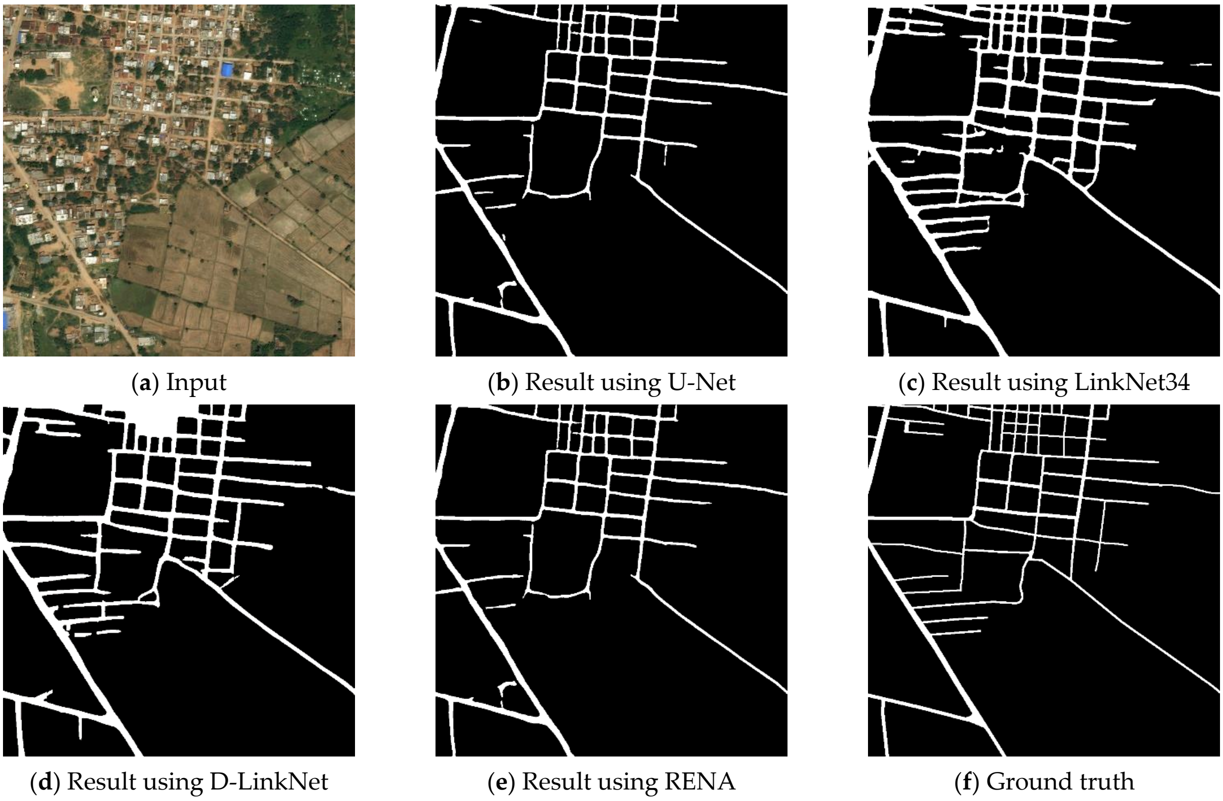

(**a**) Input
(**b**) Result using U-Net
(**c**) Result using LinkNet34
(**d**) Result using D-LinkNet
(**e**) Result using RENA
(**f**) Ground truth

Figure 8. Visual evaluation results of the third group of dense road networks.

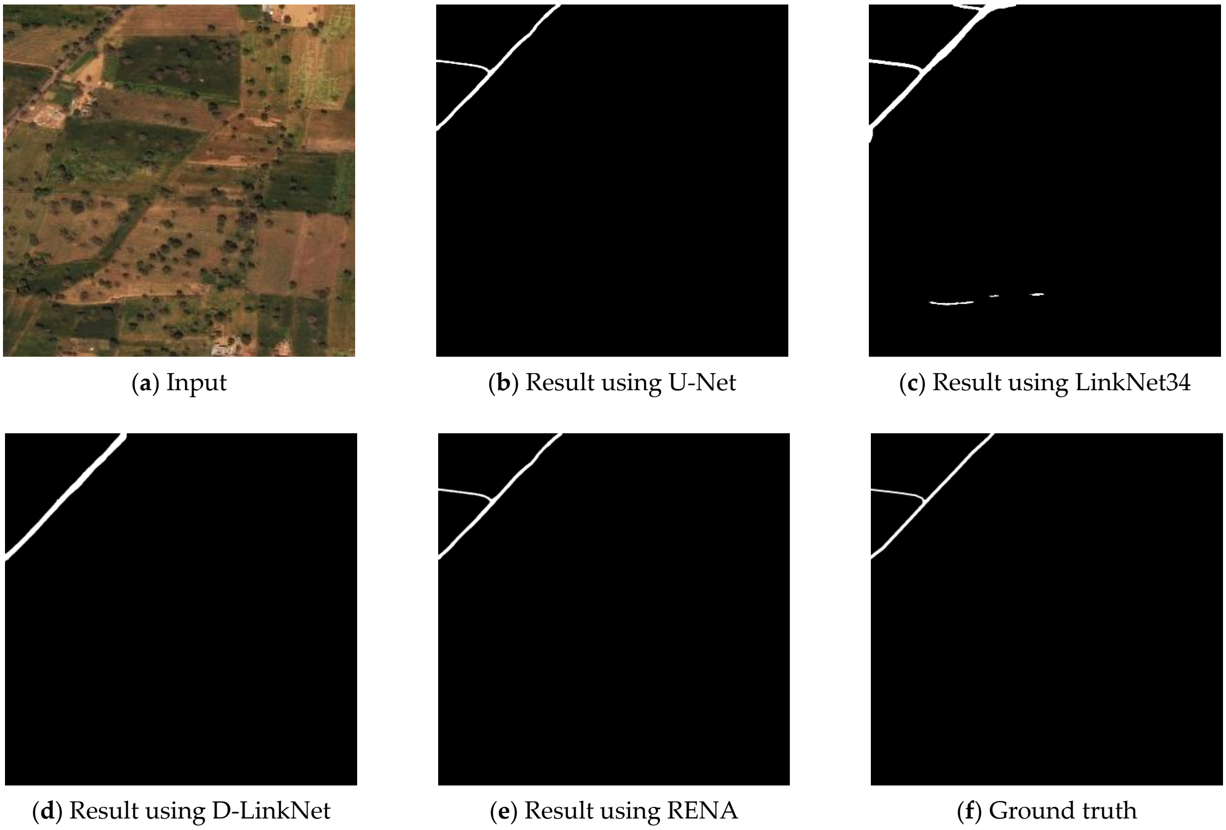

(**a**) Input
(**b**) Result using U-Net
(**c**) Result using LinkNet34
(**d**) Result using D-LinkNet
(**e**) Result using RENA
(**f**) Ground truth

Figure 9. Visual evaluation results of the first group of sparse road networks.

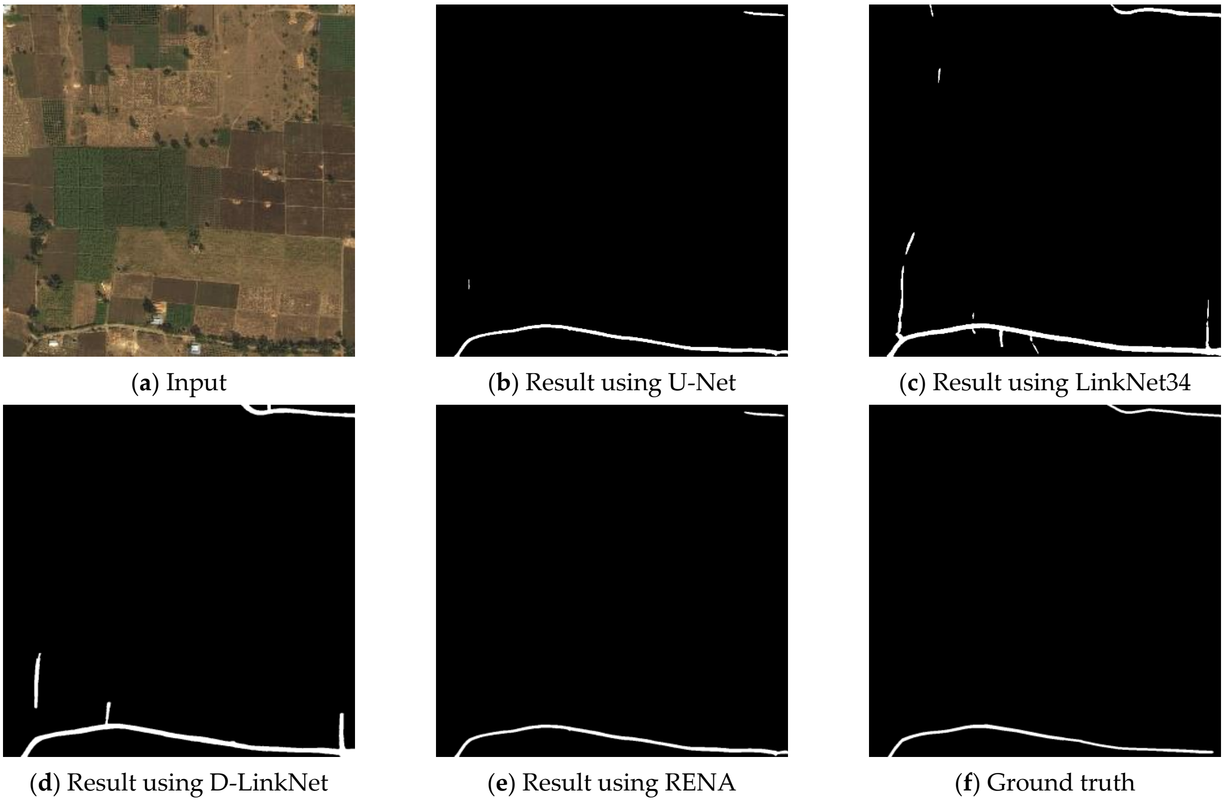

(**a**) Input
(**b**) Result using U-Net
(**c**) Result using LinkNet34
(**d**) Result using D-LinkNet
(**e**) Result using RENA
(**f**) Ground truth

Figure 10. Visual evaluation results of the second group of sparse road networks.

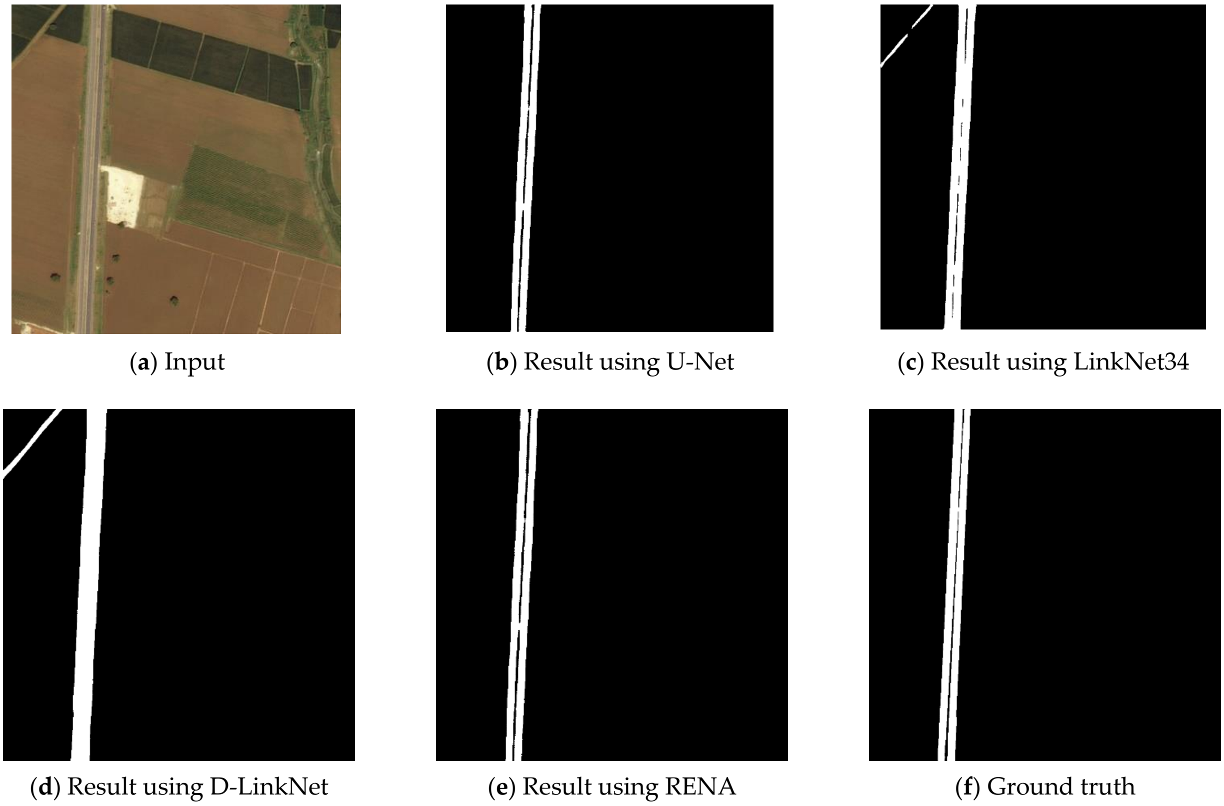

(**a**) Input
(**b**) Result using U-Net
(**c**) Result using LinkNet34
(**d**) Result using D-LinkNet
(**e**) Result using RENA
(**f**) Ground truth

Figure 11. Visual evaluation results of the third group of sparse road networks.

From the six sets of visual experiments in densely built-up areas and open areas, it was seen that LinkNet34 and D-LinkNet could completely extract the road network, but many false detections were present, and the shape of the extracted road was different from that of ground truth. U-Net and RENA had fewer false detections, but the cases of missed detection were present, and U-Net had some fine false detection roads. In addition, the roads extracted by U-Net and RENA were closer to ground truth in morphology, and the road width was also closer to ground truth.

3.5. Quantitative Evaluation Results

This study used 101 images with a size of 1024 × 1024 for road extraction experiments to validate the proposed method in quantitative evaluation, and calculated quantitative evaluation indicators with the real label results. The results are shown in Table 2.

Table 2. Quantitative evaluation results.

Method	Accuracy	Precision	Recall	F1_Score	IoU
U-Net	0.988	0.769	0.782	0.763	0.630
LinkNet34	0.978	0.539	0.966	0.680	0.528
DlinkNet	0.982	0.584	**0.975**	0.724	0.576
RENA	**0.989**	**0.784**	0.770	**0.764**	**0.631**

Note: Bold font indicates column maximum.

As shown in Table 2, the RENA method achieved the highest accuracy in four out of five metrics, indicating that the proposed method outperformed the compared algorithms in road extraction accuracy. In the recall indicator, the D-LinkNet obtained the highest accuracy, indicating that it had fewer missed detections, which was consistent with the conclusion of visual evaluation.

In the precision indicator, U-Net and RENA scored higher than LinkNet34 and D-LinkNet, indicating that U-Net and RENA had less false detection. In the recall indicator, LinkNet34 and D-LinkNet scored higher than U-Net and RENA, indicating that LinkNet34 and D-LinkNet had fewer missed detections. It was difficult to achieve high precision and high recall at the same time due to their mutual influence. Therefore, the methods were evaluated in terms of accuracy, F1-score, and IoU. The proposed method was higher than the compared algorithms in terms of three overall indicators for evaluation, indicating that the proposed method was better than the compared algorithms.

3.6. Discussion

This study conducted ablation experiments on channel attention block, spatial attention block, residual dense connection block, and residual dilated convolution module to verify the actual effect of each proposed module in road extraction experiments.

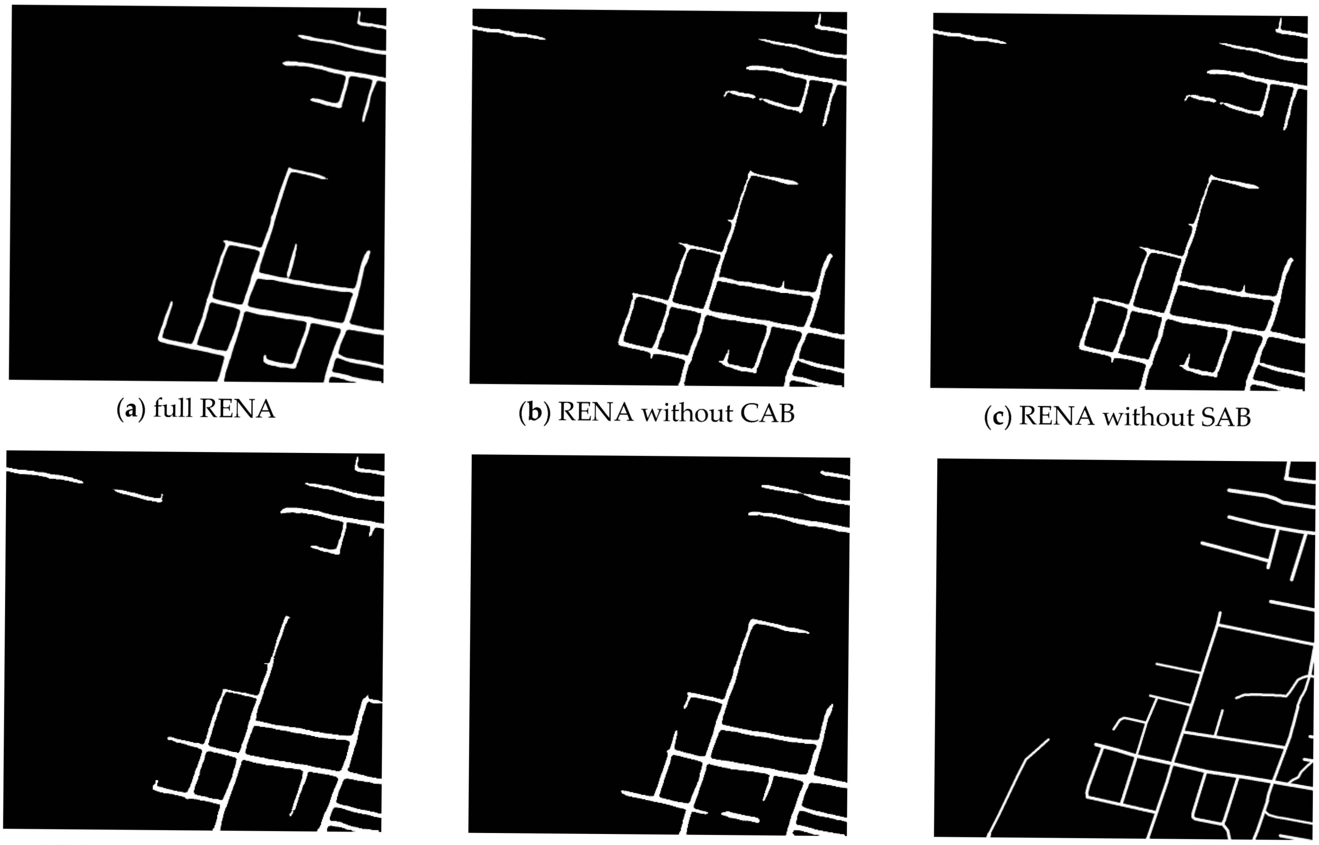
As shown in Table 3, RENA without removing any module achieved the highest accuracy among the four metrics and the second-best accuracy among one metric. In addition to a certain increase in the precision of removing the channel attention block, the results after removing each module had a certain degree of reduction in accuracy, indicating that the modules proposed in this study had a positive effect on accuracy improvement.

Table 3. Quantitative evaluation results of ablation experiments.

Method	Accuracy	Precision	Recall	F1_Score	IoU
RENA	**0.989**	0.784	**0.770**	**0.764**	**0.631**
CAB removed	0.988	**0.807**	0.722	0.747	0.611
SAB removed	0.988	0.784	0.745	0.750	0.611
RDCM removed	0.987	0.758	0.748	0.734	0.595
RDCB removed	0.987	0.739	0.763	0.732	0.596

Note: Bold font indicates column maximum.

As shown in Table 3, in terms of recall, the result with a module removed was lower than the result without the module removal to a certain extent, indicating that the result with the module removal had more wrong selections than RENA. Combinedwith Figure 12, it was seen that in addition to the original RENA results, the results without the removed modules had varying degrees of false selection, which was in line with the conclusion indicated by the quantitative indicators. For residual dilated convolution module and residual densely connected blocks, the model dropped more in accuracy, F1-score, and IoU after removing these two modules. It showed that residual dilated convolution module and residual densely connected blocks contributed more to the model in feature extraction. The spatial attention block and channel attention block were used as feature recalibration modules; the model accuracy decreased to a lesser extent after removal. The results showed that all the proposed modules could positively affect the experimental results; the residual dilated convolution module and residual densely connected blocks contributed more to the network, and the attention module contributed relatively less.


(**a**) full RENA (**b**) RENA without CAB (**c**) RENA without SAB

(**d**) RENA without RDCM (**e**) RENA without RDCB (**f**) Ground truth

Figure 12. Visual evaluation results of the second group of sparse road networks.

In the overall indicators Accuracy, F1-score and IoU, the original RENA results were higher than the results with the removal of modules, indicating that the proposed module could effectively enhance the performance of road extraction.

4. Conclusions

In this study, we used U-Net as the main framework to construct an end-to-end road extraction network with the embedded attention mechanism. From remote sensing images, we obtained road and non-road binary images, and extracted the road network from the input. In this network framework, the use of spectral information of remote sensing images was strengthened through the channel attention mechanism, and the detailed information in the road extraction results was enhanced through the spatial attention mechanism. Residual densely connected blocks were introduced in the network to enhance the information flow transfer and feature reuse of feature maps at different levels. At the same time, we introduced a residual dilated convolution module to enhance the extraction ability of road networks of different scales. Visual and quantitative experiments showed that the proposed method had higher overall accuracy and lower false detection rate than the comparison algorithms. Ablation experiments showed that the proposed modules could effectively improve the accuracy of road extraction.

However, in the case of missed detection, the proposed method still had certain shortcomings, and when faced with information occlusion such as vegetation, the proposed method still had certain defects. Future studies should employ hyperspectral imagery to better use spectral information to address the aforementioned issues. In addition, future studies should introduce more input information, such as Open Street Map road information, to assist the road extraction work and improve the road extraction accuracy.

Author Contributions: Conceptualization, S.S.; methodology, L.X.; software, L.L.; validation, S.S.; formal analysis, L.X.; investigation, L.L., C.R. and J.T.; resources, S.S. and J.T.; data curation, S.S.; writing—original draft preparation, S.S. and L.L.; writing—review and editing, L.X., C.R. and J.T.; visualization, S.S. and L.L.; supervision, S.S.; project administration, S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Dataset used in this study is obtained from DeepGlobe, which is publicly available at https://www.kaggle.com/balraj98/deepglobe-road-extraction-dataset (accessed on 21 February 2022; created by Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, Ramesh Raskar; maintained by Balraj Ashwath) under their license.

Conflicts of Interest: Shao has been working at Zhongzhi Software Technology Company Limited as a technical consultant. There is no potential conflict of interest from this company with regard to this paper.

References

1. Wang, J.; Qin, Q.; Gao, Z.; Zhao, J.; Ye, X. A New Approach to Urban Road Extraction Using High-Resolution Aerial Image. *ISPRS Int. Geo-Inf.* **2016**, *5*, 114. [CrossRef]
2. Hinz, S.; Baumgartner, A.; Ebner, H. Modeling Contextual Knowledge for Controlling Road Extraction in Urban Areas. In Proceedings of the IEEE/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas (Cat. No.01EX482), Rome, Italy, 8–9 November 2001; pp. 40–44.
3. Lin, X.; Shen, J.; Liang, Y. Semi-Automatic Road Tracking Using Parallel Angular Texture Signature. *Intell. Autom. Soft Comput.* **2012**, *18*, 1009–1021. [CrossRef]
4. Fu, G.; Zhao, H.; Li, C.; Shi, L. Road Detection from Optical Remote Sensing Imagery Using Circular Projection Matching and Tracking Strategy. *J. Indian Soc. Remote Sens.* **2013**, *41*, 819–831. [CrossRef]
5. Lin, X.; Zhang, J.; Liu, Z.; Shen, J. Semi-Automatic Extraction of Ribbon Roads from High Resolution Remotely Sensed Imagery by T-Shaped Template Matching. In Proceedings of the Geoinformatics 2008 and Joint Conference on GIS and Built Environment: Geo-Simulation and Virtual GIS Environments, Guangzhou, China, 28–29 June 2008; p. 71470J.
6. Shi, W.; Miao, Z.; Debayle, J. An integrated method for urban main-road centerline extraction from optical remotely sensed imagery. *IEEE Trans. Geosci. Remote. Sens.* **2013**, *52*, 3359–3372. [CrossRef]
7. Coulibaly, I.; Spiric, N.; Lepage, R.; St-Jacques, M. Semiautomatic road extraction from VHR images based on multiscale and spectral angle in case of earthquake. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2017**, *11*, 238–248. [CrossRef]
8. Cao, C.; Sun, Y. Automatic Road Centerline Extraction from Imagery Using Road GPS Data. *Remote Sens.* **2014**, *6*, 9014–9033. [CrossRef]
9. Manandhar, P.; Marpu, P.R.; Aung, Z. Segmentation Based Traversing-Agent Approach for Road Width Extraction from Satellite Images Using Volunteered Geographic Information. *Appl. Comput. Inf.* **2021**, *17*, 131–152. [CrossRef]
10. Mokhtarzade, M.; Zoej, M.J.V. Road Detection from High-Resolution Satellite Images Using Artificial Neural Networks. *Int. J. Appl. Earth Obs. Geoinf.* **2007**, *9*, 32–40. [CrossRef]
11. Song, M.J.; Civco, D. Road Extraction Using SVM and Image Segmentation. *Photogramm. Eng. Remote Sens.* **2004**, *70*, 1365–1371. [CrossRef]
12. Maurya, R.; Gupta, P.R.; Shukla, A.S. Road Extraction Using K-Means Clustering and Morphological Operations. In Proceedings of the 2011 International Conference on Image Information Processing, Shimla, India, 3–5 November 2011; pp. 1–6.

13. Seppke, B.; Dreschler-Fischer, L.; Wilms, C. A Robust Concurrent Approach for Road Extraction and Urbanization Monitoring Based on Superpixels Acquired from Spectral Remote Sensing Images. In Proceedings of the ESA-SP, Prague, Czech, 9 May 2016; Volume 740, p. 113.
14. Huang, X.; Zhang, L. An Adaptive Mean-Shift Analysis Approach for Object Extraction and Classification From Urban Hyperspectral Imagery. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 4173–4185. [CrossRef]
15. Cheng, G.; Yang, C.; Yao, X.; Guo, L.; Han, J. When Deep Learning Meets Metric Learning: Remote Sensing Image Scene Classification via Learning Discriminative CNNs. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2811–2821. [CrossRef]
16. Cheng, G.; Xie, X.; Han, J.; Guo, L.; Xia, G.-S. Remote Sensing Image Scene Classification Meets Deep Learning: Challenges, Methods, Benchmarks, and Opportunities. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2020**, *13*, 3735–3756. [CrossRef]
17. Kussul, N.; Lavreniuk, M.; Skakun, S.; Shelestov, A. Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 778–782. [CrossRef]
18. Yuan, X.; Shi, J.; Gu, L. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Syst. Appl.* **2021**, *169*, 114417. [CrossRef]
19. Wang, S.; Chen, W.; Xie, S.M.; Azzari, G.; Lobell, D.B. Weakly Supervised Deep Learning for Segmentation of Remote Sensing Imagery. *Remote Sens.* **2020**, *12*, 207. [CrossRef]
20. Kemker, R.; Salvaggio, C.; Kanan, C. Algorithms for Semantic Segmentation of Multispectral Remote Sensing Imagery Using Deep Learning. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 60–77. [CrossRef]
21. Xu, Y.; Wu, L.; Xie, Z.; Chen, Z. Building Extraction in Very High Resolution Remote Sensing Imagery Using Deep Learning and Guided Filters. *Remote Sens.* **2018**, *10*, 144. [CrossRef]
22. Diakogiannis, F.I.; Waldner, F.; Caccetta, P.; Wu, C. ResUNet-a: A Deep Learning Framework for Semantic Segmentation of Remotely Sensed Data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 94–114. [CrossRef]
23. Wei, Y.; Wang, Z.; Xu, M. Road Structure Refined CNN for Road Extraction in Aerial Image. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 709–713. [CrossRef]
24. Alshehhi, R.; Marpu, P.R.; Woon, W.L.; Mura, M.D. Simultaneous Extraction of Roads and Buildings in Remote Sensing Imagery with Convolutional Neural Networks. *ISPRS J. Photogramm. Remote Sens.* **2017**, *130*, 139–149. [CrossRef]
25. Zhong, Z.; Li, J.; Cui, W.; Jiang, H. Fully Convolutional Networks for Building and Road Extraction: Preliminary Results. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 1591–1594.
26. Panboonyuen, T.; Vateekul, P.; Jitkajornwanich, K.; Lawawirojwong, S. An Enhanced Deep Convolutional Encoder-Decoder Network for Road Segmentation on Aerial Imagery. Advances in Intelligent Systems and Computing. In *Recent Advances in Information and Communication Technology 2017*; Meesad, P., Sodsee, S., Unger, H., Eds.; Springer International Publishing: Cham, Germany, 2018; Volume 566, pp. 191–201. ISBN 978-3-319-60662-0.
27. Zhang, Z.; Liu, Q.; Wang, Y. Road Extraction by Deep Residual U-Net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
28. 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]
29. Chaurasia, A.; Culurciello, E. Linknet: Exploiting encoder representations for efficient semantic segmentation. In Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP), Petersburg, FL, USA, 10–13 December 2013; pp. 1–4.
30. 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. 192–1924.
31. He, H.; Yang, D.; Wang, S.; Wang, S.; Li, Y. Road Extraction by Using Atrous Spatial Pyramid Pooling Integrated Encoder-Decoder Network and Structural Similarity Loss. *Remote Sens.* **2019**, *11*, 1015. [CrossRef]
32. Wulamu, A.; Shi, Z.; Zhang, D.; He, Z. Multiscale Road Extraction in Remote Sensing Images. *Comput. Intell. Neurosci.* **2019**, *2019*, 2373798. [CrossRef] [PubMed]
33. Mosinska, A.; Marquez-Neila, P.; Koziński, M.; Fua, P. Beyond the pixel-wise loss for topology-aware delineation. In Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3136–3145.
34. Wan, J.; Xie, Z.; Xu, Y.; Chen, S.; Qiu, Q. DA-RoadNet: A dual-attention network for road extraction from high resolution satellite imagery. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2021**, *14*, 6302–6315. [CrossRef]
35. 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, 15–21 June 2019; pp. 3146–3154.
36. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
37. Kim, J.-H.; Choi, J.-H.; Cheon, M.; Lee, J.-S. RAM: Residual Attention Module for Single Image Super-Resolution. *arXiv* **2018**, arXiv:1811.12043 [cs.CV].
38. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual Attention Network for Image Classification. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6450–6458.

39. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. Lecture Notes in Computer Science. In *Computer Vision–ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Germany, 2018; Volume 11211, pp. 294–310. ISBN 978-3-030-01233-5.
40. Guo, H.; Shi, Q.; Du, B.; Zhang, L.; Wang, D.; Ding, H. Scene-Driven Multitask Parallel Attention Network for Building Extraction in High-Resolution Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 4287–4306. [CrossRef]
41. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
42. Jiao, L.; Huo, L.; Hu, C.; Tang, P. Refined UNet: UNet-Based Refinement Network for Cloud and Shadow Precise Segmentation. *Remote Sens.* **2020**, *12*, 2001. [CrossRef]
43. McGlinchy, J.; Johnson, B.; Muller, B.; Joseph, M.; Diaz, J. Application of UNet Fully Convolutional Neural Network to Impervious Surface Segmentation in Urban Environment from High Resolution Satellite Imagery. In Proceedings of the IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 3915–3918.
44. Shamsolmoali, P.; Zareapoor, M.; Wang, R.; Zhou, H.; Yang, J. A Novel Deep Structure U-Net for Sea-Land Segmentation in Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2019**, *12*, 3219–3232. [CrossRef]
45. Li, R.; Liu, W.; Yang, L.; Sun, S.; Hu, W.; Zhang, F.; Li, W. DeepUNet: A Deep Fully Convolutional Network for Pixel-Level Sea-Land Segmentation. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2018**, *11*, 3954–3962. [CrossRef]
46. Cao, K.; Zhang, X. An improved res-unet model for tree species classification using airborne high-resolution images. *Remote Sens.* **2020**, *12*, 1128. [CrossRef]
47. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y. Residual Dense Network for Image Super-Resolution. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2472–2481.
48. Wei, Y.; Ji, S. Scribble-based weakly supervised deep learning for road surface extraction from remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–12. [CrossRef]