remote sensing

MDPI

Article

A Residual Attention and Local Context-Aware Network for Road Extraction from High-Resolution Remote Sensing Imagery

Ziwei Liu [1], Mingchang Wang [1,2,*], Fengyan Wang [1] and Xue Ji [1]

[1] College of Geo-Exploration Science and Technology, Jilin University, Changchun 130026, China; lzw21@mails.jlu.edu.cn (Z.L.); wangfy@jlu.edu.cn (F.W.); jixuesdqd@jlu.edu.cn (X.J.)
[2] Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources, Shenzhen 518000, China
* Correspondence: wangmc@jlu.edu.cn; Tel.: +86-135-0431-1009

Abstract: Extracting road information from high-resolution remote sensing images (HRI) can provide crucial geographic information for many applications. With the improvement of remote sensing image resolution, the image data contain more abundant feature information. However, this phenomenon also enhances the spatial heterogeneity between different types of roads, making it difficult to accurately discern the road and non-road regions using only spectral characteristics. To remedy the above issues, a novel residual attention and local context-aware network (RALC-Net) is proposed for extracting a complete and continuous road network from HRI. RALC-Net utilizes a dual-encoder structure to improve the feature extraction capability of the network, whose two different branches take different feature information as input data. Specifically, we construct the residual attention module using the residual connection that can integrate spatial context information and the attention mechanism, highlighting local semantics to extract local feature information of roads. The residual attention module combines the characteristics of both the residual connection and the attention mechanism to retain complete road edge information, highlight essential semantics, and enhance the generalization capability of the network model. In addition, the multi-scale dilated convolution module is used to extract multi-scale spatial receptive fields to improve the model's performance further. We perform experiments to verify the performance of each component of RALC-Net through the ablation study. By combining low-level features with high-level semantics, we extract road information and make comparisons with other state-of-the-art models. The experimental results show that the proposed RALC-Net has excellent feature representation ability and robust generalizability, and can extract complete road information from a complex environment.

Keywords: attention mechanism; road extraction; deep learning; remote sensing; multi-scale dilated convolution

check for updates

Citation: Liu, Z.; Wang, M.; Wang, F.; Ji, X. A Residual Attention and Local Context-Aware Network for Road Extraction from High-Resolution Remote Sensing Imagery. *Remote Sens.* **2021**, *13*, 4958. https://doi.org/10.3390/rs13244958

Academic Editors: Zhenwei Shi, Bin Pan and Shuo Yang

Received: 2 November 2021
Accepted: 4 December 2021
Published: 7 December 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

The rapid development of remote sensing technology has led to a sharp increase in high-resolution remote sensing image data and provides an important data source for road extraction. Compared with medium and low-resolution remote sensing images, HRI contains more texture, shape, structure, neighborhood relations, and other information, which can extract road information more accurately [1]. Extracting road information accurately and quickly can provide important geographic information for urban planning and spatial decision-making [2,3]. Accurate extraction of ground objects, such as roads, from remote sensing images has excellent social and economic value, which can play an essential role in disaster investigation [4,5], automatic navigation [6], geographical information system construction [7,8], and other fields.

Generally speaking, road extraction methods are divided into traditional methods and deep learning-based methods [9,10]: (1) Traditional methods. In this method, road types are defined manually based on one particular feature information, and then the corresponding feature extraction model is constructed to recognize and extract the road [11,12]; (2) Deep learning-based methods. This method takes advantage of automatic feature extraction ability, strong generalization ability, high efficiency of fitting ability, and stability robustness of the deep learning model to mine deep characteristic information from HRI based on the prior knowledge to complete the automatic extraction of road information [13–15].

Specifically, the optimization of the network model structure is one of the main strategies to improve performance. Xin et al. [16] optimized U-net by replacing the original encoder with Dense blocks, which reduced the number of parameters and improved the robustness of the network. Xie et al. [17] chose LinkNet as the basic architecture to construct HsgNet by adding a middle block between the encoder and decoder. The function of the middle block is to save global-context semantic information, long-distance spatial information and the relationship, and different feature channels' information and dependencies, which can improve the generalization ability and performance of the network. Liu et al. [18] proposed a multi-task convolutional neural network named Roadnet to extract roads from complex urban scenes, which completed the task of simultaneously predicting road surfaces, edges and center lines for the first time. The other works mainly take advantage of the discriminative probability model as a post-process method to further improve the accuracy of the extracted road results. Liu et al. [19] combined the adaptive sparse model with Markov random field, extracted more abstract and more discriminative high-level features, and then used Gabor filters and nonmaximal suppression combined with ridge transversal method to extract the centerline of the road, which can obtain a more accurate extraction result. Wang et al. [20] introduced the conditional random field into the encoder-decoder structure and enhanced the road extraction structure to improve classification accuracy through the conditional random field.

In the past, road extraction from remote sensing images based on deep learning had the following problems: (1) The spatial heterogeneity between roads caused by issues such as solar elevation angle, occlusion, and lack of semantic information leads to a decrease in the prominence of road geometric texture, discontinuous road extraction results, and fragmented road sections in the road network. In this case, a network structure using only a single encoder cannot effectively extract and abstract road features; (2) Most previous works focused on integrating global context information. They ignored the importance of the local perception information that could not accurately describe the diversity of spatial distribution of complex ground objects in the case of similar spatial-temporal characteristics of different ground objects. However, local information can enhance the feature representation ability of major information and retain spatial topological information such as texture and boundary; (3) The enhancement in the resolution of remote sensing images makes the object information present as highly detailed. At the same time, imaging conditions and the complexity of the composition and distribution of the objects can increase intraclass variability (and interclass similarity), a decrease in the distinguishability of different objects. The category of objects whose spectral curves are very similar cannot be effectively distinguished only according to the image's spectral radiation or grayscale information, resulting in the misclassification of two kinds of objects with similar reflectance characteristics.

To solve the above problems, we propose a residual attention and local context-aware network (RALC-Net) for road extraction from HRI. RALC-Net takes advantage of the encoder–decoder structure to connect context information, a residual attention module to highlight representation ability of local semantics, and a multi-scale dilated convolution module to enhance feature extraction ability. The main contributions of this paper are as follows:

(1) To extract complete road information from HRI, we designed the RALC-Net model. The proposed network consists of a dual-encoder with the residual attention module, a multi-scale dilated convolution module, and a parallel partial decoder. At the same time, the context information is highlighted by using skip connections between the

encoder and decoder. The dual-encoder structure is used to enhance the feature abstraction and fusion capabilities of the network and extract the feature map of the input data. Then the extracted feature map is passed through the multi-scale dilated convolution module to retain the feature information at different spatial scales and obtain high-level semantic information under the multi-scale spatial receptive field, enhancing the network's feature representation ability. The decoders gradually enlarge the spatial size of the obtained feature map until it is consistent with the input data, and the classifier is ultimately used to make classification decisions;

(2) By introducing residual connection into the attention mechanism, the residual attention module is constructed to emphasize the local semantics and improve the generalization ability of the network model. The residual attention module can retain local detailed semantic information and spatial boundary and use residual mapping to extract and abstract deep feature maps to relieve over-fitting;

(3) Multi-feature information is used as the input of the network model to assist the extraction of road information from HRI, which has more abundant spectral and texture information. Color, texture, and shape information extracted from image data can provide the essential decision-making basis for road information extraction. The original image, color feature, texture feature, and shape feature are simultaneously input into the model to extract road information, which improves the generalization ability of the network model and enhances the robustness.

The remainder of this paper is arranged as follows. Section 2 reviews the application of traditional methods, deep learning-based methods, and attention mechanisms in road extraction. Section 3 introduces the methodology of the approach proposed in this article and multi-feature information. Section 4 presents experimental results and evaluations. In Section 5, we discuss the performance of the network and the influence of multi-feature details on road extraction. The conclusion is described in Section 6.

2. Related Work

2.1. Traditional Road Extraction Methods

Traditional road extraction methods can be divided into three categories: pixel-based, object-oriented, and machine learning-based road extraction methods. The pixel-based road extraction method usually targets a single pixel and its surrounding pixels, and considers their spectral characteristics, temporal and spatial relationships, and gradient characteristics to establish the corresponding road extraction model [21]. Liu et al. [22] used the linear features of roads to construct a rural road geometric knowledge database, identified parallel line segments in the image according to the geometric features of the road, grouped and connected the line segments, and finally extracted a complete rural road. Cheng et al. [23] used a combination of semi-supervised classification methods with multi-scale filtering and multi-direction non-maximum suppression (MF&M-NMS) methods to extract road networks based on the relationship between labelled samples and unlabeled samples. Shi et al. [24] divided the road and non-road areas according to the spectral characteristics of the image and then extracted practical road sections by using the shape features of the road and refined and smoothed the extraction results to obtain the complete extraction results. However, this method has certain limitations and is only suitable for extracting urban arterial roads from high-resolution images. Shanmugam et al. [25] proposed a junction-aware water flow model, which automatically recognizes and extracts road networks based on geometric features such as the width, direction, centerline and length of each road. This method can be used to identify complex road intersections with fewer automatically generated anchor points.

Since the pixel-based road extraction method is prone to the salt and pepper phenomenon, Baatz et al. [26] proposed an object-oriented image processing method for image classification and the extraction of various ground features, effectively solving the effect of salt and pepper noise on the results. Li et al. [27] defined structural features based on orientation histograms and morphological profiles, combined the two geometric features of compactness and elongation to merge the roads in the region of interest, and successfully grouped and linked adjacent small sections of roads with high spectral heterogeneity and similar geometric shapes. Huang et al. [28] used an object-oriented method to extract multi-scale information and road structure features to reduce the impact of local spectral changes in the image. Miao et al. [29] proposed a novel object-based automatic road extraction method, which went through five steps in sequence: image segmentation, filtering, feature set extraction, road contour extraction, and post-processing, to overcome the limitations of automatic road extraction to a certain extent. Yin et al. [30] improved the ant colony algorithm to implement global optimization of roads according to the object and edge information, effectively overcoming the shortcoming that the object-oriented method could not make full use of edge information.

Machine learning can automatically extract abstract feature information by mining and abstracting the sample data to fit the correlation and relationship between the samples and then ensure the classification interpretation of the ground objects according to these features [31]. Road extraction methods based on machine learning commonly include support vector machine (SVM), k-nearest neighbor (KNN), random forest (RF), and decision tree (DT) [32–34]. Zhu et al. [35] used K-means to segment the image and then combined SVM and the fuzzy C means method (FCM) to extract the road network from the segmentation map. Xu et al. [36] converted the road network into vector polygons and trained the random forest model by using various geometric driving factors such as compactness, width, circularity, area, perimeter, complexity, parallelism, shape descriptor, and width-to-length ratio to extract the road information of Beijing. Soni et al. [37] combined least square support vector machine (LS-SVM), mathematical morphology, and road shape features to construct a multi-stage road extraction model. Specifically, the image was divided into road and non-road areas. Then, the road centerline was extracted by the method based on Euclidean distance transformation. Finally, the road network result was obtained.

2.2. Road Extraction Methods Based on Deep Learning

In 2006, Hinton formally proposed the concept of deep learning, which is considered the next technological extension to artificial neural networks [38]. A multi-level hierarchical model similar to the human brain's nervous system is constructed to automatically learn and extract the feature information of the input data, from the bottom to the high level, and completes the classification according to the extracted feature information. Under the big data stage, accurately extracting practical information from massive remote sensing data is a hot research direction, and deep learning provides efficient technical approaches [39–41]. Convolutional neural network (CNN) is one of the most common deep learning models in image processing, which shows outstanding performance in road information extraction from remote sensing imagery [42]. However, the CNN-based road extraction method relies too much on the characteristics of the sliding window, resulting in low extraction efficiency and difficulty in meeting the requirements of practical application. The fully convolutional network uses convolutional layers instead of fully-connected layers to achieve end-to-end semantic segmentation and overcome the inefficiency of the CNN method based on a sliding window, which can preserve target spatial information while extracting semantic text information [43–45]. Zhou et al. [46] proposed a boundary and topologically-aware neural network (BT-Roadnet) to solve the problem of boundary accuracy and topological connectivity by using the CMPM block to convert the input optical image into road probability maps and using the FMPM block to do further feature extraction on the road probability map. Cheng et al. [47] used the encoder-decoder structure to construct a cascaded end-to-end convolutional neural network (CasNet), attempting to connect two subtasks using the cascaded network for the first time. Zhang et al. [48] improved the U-net network by adding residual units, which simplified the training of the deep neural network and improved the generalization ability of the network model while reducing the number of parameters by adding multi-layer skip connections. Abdollahi et al. [49] proposed a method to extract road information based on the generative adversarial network (GAN). In this method, the improved U-Net network was used to obtain road network feature maps, which retained edge information, for the most part, reduced the influence of occlusions from trees, and finally got ideal extraction results.

2.3. Attention Mechanisms

In addition, the attention mechanism is an important research area in the field of computer vision. Because of its superior feature representation ability, it has been widely applied to extract remote sensing information [50,51]. The attention mechanism can highlight important information and suppress secondary information by assigning disparate weights to different information [52]. Xu et al. [53] introduced the attention unit into the deep convolutional neural network to extract local and global information of the road in the remote sensing image and improve the accuracy of road network extraction. Wan et al. [54] used a shallow encoder framework to construct a Dual-Attention Network (DA-RoadNet) to explore and analyze the correlation of road features in spatial dimensions and channel dimensions and then extract road information from a complex environment. Ren et al. [55] designed a Dual-Attention Capsule U-Net (DA-CapsUNet) to extract and fuse multi-scale context information of road networks by constructing a feature attention module.

3. Methodology

In this section, we introduce the structure of the proposed RALC-Net in Figure 1. The network uses an encoder–decoder structure to construct the model, encoders with a residual attention module to gradually extract high-level semantic information from the input data, a multi-scale dilated convolution module to extract multi-scale spatial information, decoders to amplify the abstracted feature information layer by layer, which achieves the purpose of pixel-by-pixel semantic segmentation. We introduce the structure of the RA module and the multi-scale dilated convolution and how to extract the spatiotemporal features of the input data. We present the composition of both the residual attention module and the multi-scale dilated convolution and extract the spatiotemporal characteristics of the input data. In addition, we list the types of feature information used in this article and explain how multi-feature information is input into the network model.

3.1. The Structure of Encoder-Decoder

At present, most of the state-of-the-art network models choose to use the encoder–decoder structure such as U-net [56], Deeplab [57] and Hsgnet [17], which can extract richer detailed feature information of the input information. In general, encoders usually choose excellent network structures (such as VGG and ResNet). These network structures generally include convolutional layers, pooling layers, fully connected layers, batch normalization layers (BN layers), and activation layers. The adjusted ResNet-50 is taken as as the baseline of the dual-encoder structure, which is composed of a multi-layer residual mapping block. These residual mapping blocks are divided into two primary components, conv blocks and identity blocks. Each convolution block includes two 1 × 1 convolutional layers and a 3 × 3 convolutional layer. The identity block has one more 1 × 1 convolutional layer than the convolution block on the shortcut, which is used to adjust the dimension of the channel. Compared with the encoder structure, the decoder structure is composed of up-sampling and convolution blocks. Up-sampling enlarges the spatial size of the feature map, and the convolution layer performs local feature extraction on the amplified feature map.

As shown in Figure 1, the RALC-Net network proposed in this paper has two input paths, and the two input information consists of a variety of feature information. Compared with a single encoder, two symmetrical encoders can enhance the feature extraction capability of the model by assigning them different values of weight parameters. The input data are processed through multiple conv blocks and identity blocks, and feature maps in different dimensions are extracted layer by layer. However, the input data lose part of the spatial and spectral information due to multiple dimensionality reduction processing.

Therefore, we combine the underlying feature information with in-depth features by using skip connections between the encoder and the decoder to obtain features with higher dimensions to effectively extract road information in complex environments.

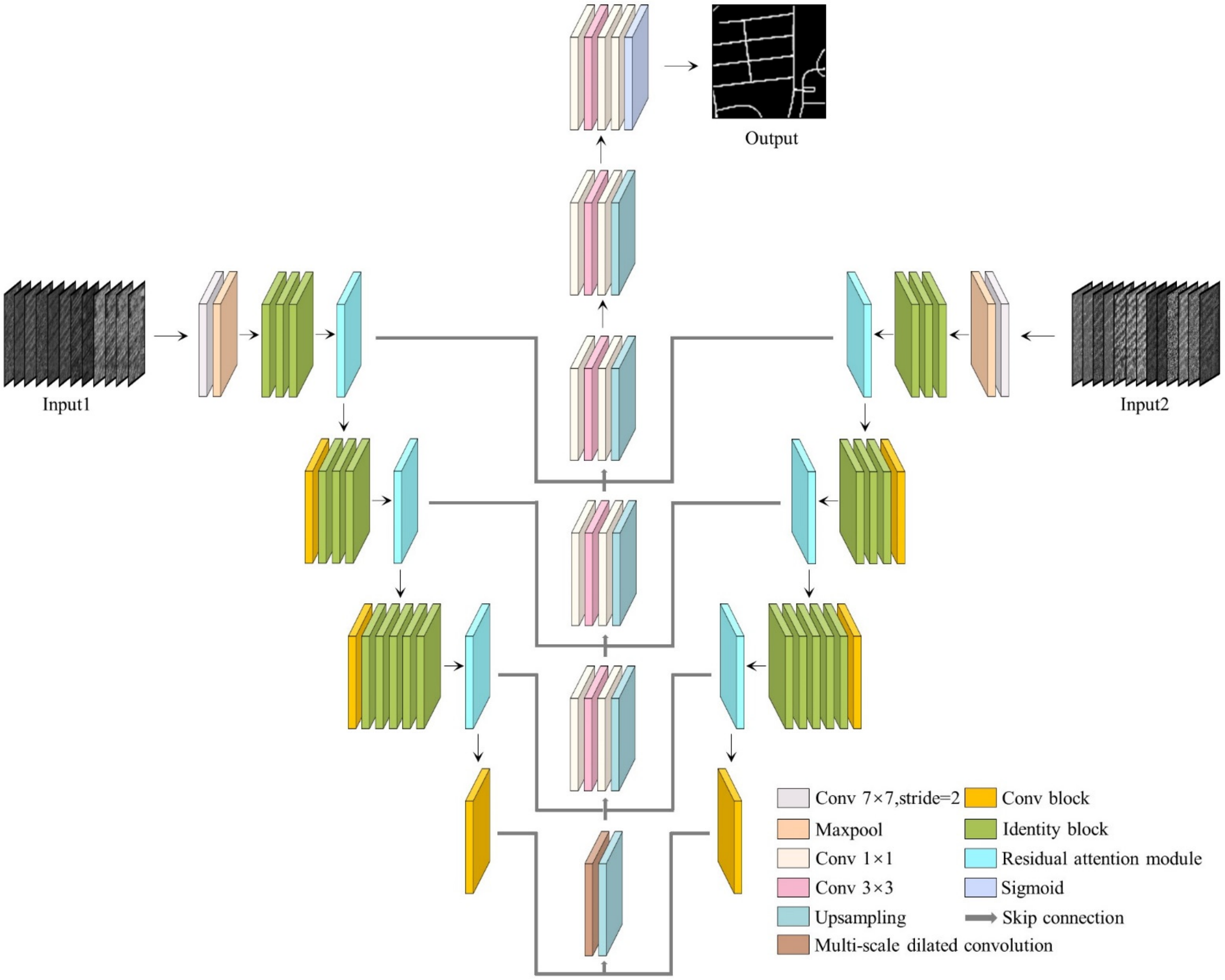

Figure 1. The architecture of the proposed RALC-Net. RALC-Net takes advantage of encoder–decoder structures, residual attention modules, and multi-scale dilated convolution modules to construct the model.

3.2. Residual Attention Module

The attention mechanism is currently the most advanced model for solving multi-tasks, which overcomes some limitations of traditional neural networks and provides certain interpretability for the agnostic nature of deep learning models [58,59]. We construct an attention mechanism model that simulates the human perception system and extracts important feature information from top to bottom. The model uses different weights to highlight important information and suppress irrelevant information. At the same time, we can also take advantage of the attention mechanism to solve a series of problems caused by the correlation between different feature channels, the reduction of computational efficiency resulting from different input sequences, and the lack of extraction and abstraction for important information in the network. The attention mechanism can use high weight to highlight important information and low weight to ignore irrelevant information, enhancing the generalization ability and robustness of the network model to extract meaningful information in different scenarios.

The combination of the attention mechanism and deep learning can successfully improve the performance of deep learning models. As for the encoder–decoder structure, the network model often generates many residuals during the feature mapping process. These residuals will increase as the number of network layers deepens. Correspondingly, while the attention mechanism uses different weights to highlight important information, it also provides a certain degree of interpretability for the "black box" characteristics. Accordingly, we propose a residual attention module (RA module) in this paper, which uses the attention mechanism to highlight crucial local information and residual connections to integrate local context information, and finally achieves the purpose of highlighting local contextual details. The RA module includes two parts: series convolution and shortcut, and the structure of the RA module is shown in Figure 2. In series convolution, a BN layer and a ReLU activation function are followed after each convolution layer to improve the network's generalization ability and speed up the convergence of the model. The convolution kernel of each convolution layer in the series convolution structure is $\left\{2^{(i+5)}, 2^{(i+5)}, 2^{(i+6)}\right\}$ and i means the number of the RA module. Because of the inconsistency of the number of input and output channels in the series convolution structure, the shortcut structure composed of 1 × 1 convolution is introduced to adjust the channel dimension, and the number of convolution kernels is $2^{(i+6)}$.

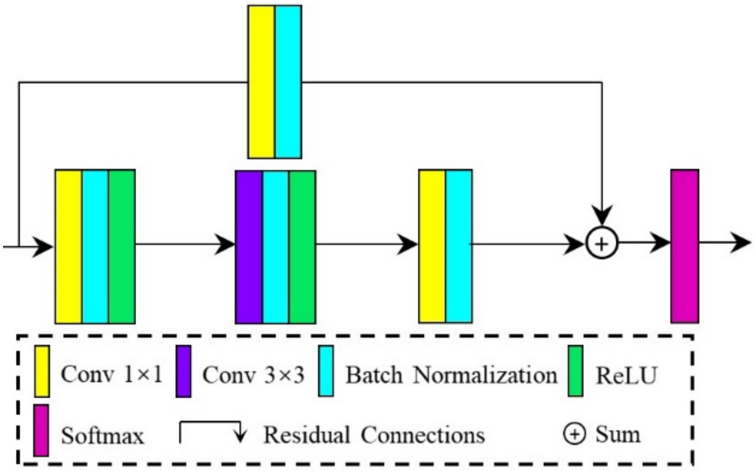

Figure 2. The structure of the residual attention module.

3.3. Multi-Scale Dilated Convolution

We added a multi-scale dilated convolution module between the encoder and the decoder in the model structure, using different dilation rates of dilated convolution and 1 × 1 convolution layer to extract multi-scale feature maps. Figure 3 shows the multi-scale dilated convolution structure with five channels. We only use the 1 × 1 convolutional layer to extract feature information in the first channel and add a 3 × 3 convolutional layer in the second channel. From the third to the fifth channel, we sequentially append dilated convolutions with dilation rates of $\{1, 2, 3\}$ to expand the range of the receptive field without increasing the complexity of the model. The specific calculation process is defined as follows:

$$g(l_0) = \sum_{i=1}^{N} f_i(l_0)\,, \tag{1}$$

where l_0 represents the multi-scale dilated convolution input feature map, and $f_i()$ represents the ith layer multi-scale dilated convolution. Finally, the calculation results of each layer are added, and the fused multi-scale feature information is obtained.

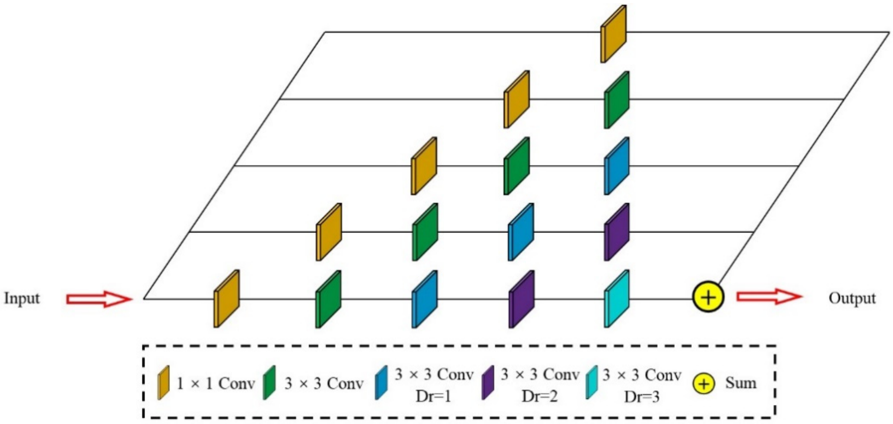

Figure 3. The architecture of the proposed Multi-scale dilated convolution: There are five channels to extract feature maps at different scales simultaneously. The five channels are composed of a 1 × 1 convolutional layer, a 3 × 3 convolutional layer, and dilated convolution layers with dilation rates of $\{1, 2, 3\}$.

3.4. Multi-Feature Information

The improvement of remote sensing image resolution makes the image data have more abundant feature information. This underlying information, including color, texture, spectrum, and neighborhood relations, can be extracted according to the feature information [60]. The underlying information positively affects road extraction and can accurately extract the target road from a complex environment. We use the original image data and the color, texture and shape features extracted from remote sensing images as the two inputs of RALC-Net, making full use of the rich feature information of high-resolution remote sensing images to extract road information accurately. In this article, color moments are selected as color features, Gray Level Co-occurrence Matrix (GLCM) based on statistics is selected as texture features, and edge detection results based on contours are selected as shape features. The specific feature combination is shown in Figure 4. The processing of color features and texture features usually extracts its information pixel by pixel. In this process, the spatial feature information of the original data will be lost, so this article uses the domain method to extract color features and texture features, which uses a 3 × 3 neighborhood window centered on a specific pixel and calculates the feature information in each neighborhood window as the result of the center pixel.

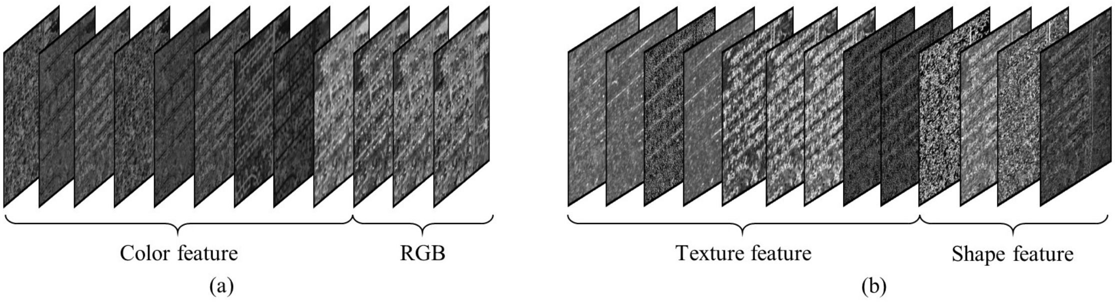

Figure 4. Feature fusion: (**a**) The color feature is superimposed with the original image as input one; (**b**) The texture feature is superimposed with the shape feature as input two.

3.4.1. Color Feature

The color feature is one of the most commonly used information features in tasks such as target detection and image classification. It can intuitively reflect the surface properties of objects and is less affected by image size and shooting angle. This paper uses color moments as color features, and the information in color moments is often reflected in low-order moments. Therefore, the low-order moments including mean, variance, and skewness, are selected as the color driving factors. The calculation formula is as follows:

$$M_i = \frac{1}{N}\sum\nolimits_j^N p_{ij} \tag{2}$$

$$V_i = \sqrt{\frac{1}{N}\sum\nolimits_j^N (p_{ij} - M_i)^2} \tag{3}$$

$$S_i = \sqrt[3]{\frac{1}{N}\sum\nolimits_j^N (p_{ij} - M_i)^3} \tag{4}$$

where M_i, V_i and S_i represent the mean, variance, and skewness of the ith pixel respectively, and p_{ij} represents the pixel value of the ith row and the jth column.

3.4.2. Texture Feature

The texture feature is a quantitative description of the texture properties of an image. It is also a visual feature that does not depend on color or brightness information to reflect the homogeneity of the image and contains important arrangement information of the image surface structure and its relationship with the surrounding environment [61]. The Gray Level Co-occurrence Matrix (GLCM) is a statistically-based texture feature that can conduct a statistical analysis on all pixels of the image according to the distance and angle between each pixel in the imagery, and extract texture features in a conditional probability manner to describe the spatial correlation of image pixels in grayscale. As the result of the statistical analysis, GLCM cannot be directly used to describe texture information, but the feature value calculated based on GLCM is taken as the description factor of texture information. This article uses homogeneity, contrast, and energy as description factors of texture features.

$$Hom = \sum_i \sum_j \frac{g(i,j)}{1 + (i-j)^2} \tag{5}$$

$$Con = \sum_i \sum_j (i-j)^2 g(i,j) \tag{6}$$

$$Ene = \sum_i \sum_j g(i,j)^2 \tag{7}$$

where $g(i,j)$ is the value of the ith row and the jth column in GLCM.

3.4.3. Shape Feature

The shape feature is a high-level visual feature in the image. Under certain extreme conditions, where other features do not exist, the category of the object can still be identified through the shape and contour of the object. The edge of an image is a collection of positions where the grayscale value of the pixel changes sharply, and it is one of the main description methods of shape features. Image edge detection usually uses a binarization method to distinguish the edge area from other areas, highlight the region of change in the object's shape, and show the object's contour. The edge detection methods used in this paper include Robert, Laplace, Prewitt and Canny.

4. Experiments

In this section, the Massachusetts Roads Dataset and DeepGlobe Roads Dataset are taken as the experimental data to validate the performance of the proposed RALC-Net. For experiment 1, different encoders with RA modules are used to extract road information in the two datasets, and results of qualitative and quantitative comparisons are presented in Sections 4.3.1 and 4.3.2. For experiment 2, we compared the impact of multi-feature information on the performance of other state-of-the-art models on the Massachusetts Roads Dataset in Section 4.3.3. In addition, Adam is chosen as the optimizer with the batch size of 1 and Binary Cross-Entropy loss is taken as the loss function. The initial learning rate is 0.0002, and when the loss value does not decrease for three epochs, the learning rate is multiplied by 0.5. Tensorflow is taken as the backend, and Keras is used to implement all network models trained on NVIDIA GTX2080 GPU.

4.1. Dataset

4.1.1. Massachusetts Roads Dataset

The Massachusetts Roads Dataset [62] is composed of 1171 aerial images, with a total coverage area of 2.25 square kilometres, and the size of each image is 1500 × 1500. The original dataset includes 1108 training images, 14 validating images, and 49 test images. To validate the performance of the model proposed in this article, we screened the original data set and selected 718 images with excellent quality for training and 49 images for testing by discarding some incomplete images. Furthermore, to prevent computer memory from overflowing, we cropped the image data to a size of 512 × 512 and obtained 10,052 training images and 441 test images after data enhancement.

4.1.2. DeepGlobe Roads Dataset

The DeepGlobe Roads Dataset [63] consists of 6226 training images, 1243 validating images, and 1101 test images, including cities and rural areas in many countries such as Thailand, Indonesia, and India. Each image has three bands with a spatial resolution of 0.5 m. In the original dataset, only 6226 training images have label data. Therefore, we randomly divide 6226 images into 4359 training images and 1867 test images. Similarly, we cropped and enhanced the original images to obtain 26,154 training images and 7468 test images with a spatial size of 512 × 512.

4.2. Evaluation Metrics

We use the mean union score ($mIOU$) as an evaluation metric in this paper. For the $mIOU$, we need to calculate the union score (IOU) of each original dataset image and then take the average of all the calculation results.

$$IOU_i = \frac{TP_i}{TP_i + FP_i + FN_i} \tag{8}$$

$$mIOU = \frac{1}{n}\sum_{i=1}^{n} IOU_i \tag{9}$$

where TP_i, FP_i and FN_i is the true positive, false positive and false negative of the ith image.

In addition, we also take the $F1$-score as an evaluation metric. Firstly, we need to obtain precision (P) and recall (R). Then the $F1$-score is calculated based on their correlation.

$$P = \frac{TP}{TP + FP} \tag{10}$$

$$R = \frac{TP}{TP + FN} \tag{11}$$

$$F1 = 2 * \frac{P * R}{P + R} \tag{12}$$

However, the calculation results of precision, recall, and $F1$-score excessively depend on the number of samples and categories, and there will be uncertainties in the results. Therefore, we can combine the *Kappa* coefficient to more objectively evaluate the accuracy of the results. The *Kappa* coefficient uses a discrete multivariate analysis method to calculate the degree of similarity between the actual spatial distribution of the ground features and the classification results predicted by the classifier, which can more objectively and fairly express the pros and cons of the results.

$$Pe = \frac{\sum_{i=1}^{n} t_i p_i}{M * M} \tag{13}$$

$$Kappa = \frac{Po - Pe}{1 - Pe}, \tag{14}$$

where t_i is the number of actual results of ith category, p_i is the number of prediction results of the ith category, M is the total number of samples, and Po is the overall accuracy.

4.3. Experimental Results

We mainly validate the performance of the proposed RALC-Net through two experiments. In the first experiment, we chose the three state-of-the-art network structures of VGG16, Res34, and Res50 as the encoder to analyze the performance of the different encoders and the RA module on the two datasets. In the second experiment, the proposed RALC-Net was compared with four state-of-the-art road extraction methods based on deep learning models, including SegNet [64], U-Net [56], DeeplabV3+ [57], and D-LinkNet [65], which compare the extraction results of two input ways using only image data and using multi-feature information.

4.3.1. Ablation Study on the Encoders and the RA Module

Figure 5 shows the visual comparison of different encoders on the Massachusetts Roads Dataset. In the first three rows of Figure 5, the spectrum and texture of the rural roads and non-road areas are very similar, and the continuous road network cannot be extracted entirely. However, after the RA module is added to encoders, the local context perception performance is used to extract crucial local information and improve the accuracy of the results. The three original models show dissimilar performances in the fourth and fifth rows. The Res50 model extracts relatively complete road network information, with better visual outcomes than the Res34 and VGG16 models. After adding the RA module, the three models effectively improve the fragmented and fractured conditions and obtain complete road networks. In general, the performance of the Res50 model performance is significantly better than the other two models through the comprehensive analysis of the extraction results of the three models. At the same time, the experimental results verified the challenging performance of the RA module.

As shown in Table 1, the quantitative comparisons of the three encoders after adding the RA module include *mIOU*, *F*1-score, and *Kappa* coefficient. Before adding the RA module, among the three encoder models, the *mIOU*, *F*1-score, and *Kappa* coefficient of the VGG16 model were 0.5325, 0.6949, and 0.6824, respectively. However, the accuracy of the Res34 model outperforms that of the VGG16 model, indicating that the residual connection can effectively improve the performance of the model. Besides, the *mIOU*, *F*1-score, and *Kappa* coefficient of the Res50 model are higher than the other two models, reaching 0.5563, 0.7149, and 0.7023. After adding the RA module, the accuracy of the three encoders has been increased, and the use of the RA module improves the performance by 0.0383, 0.0108, and 0.0271 in *mIOU*, respectively.

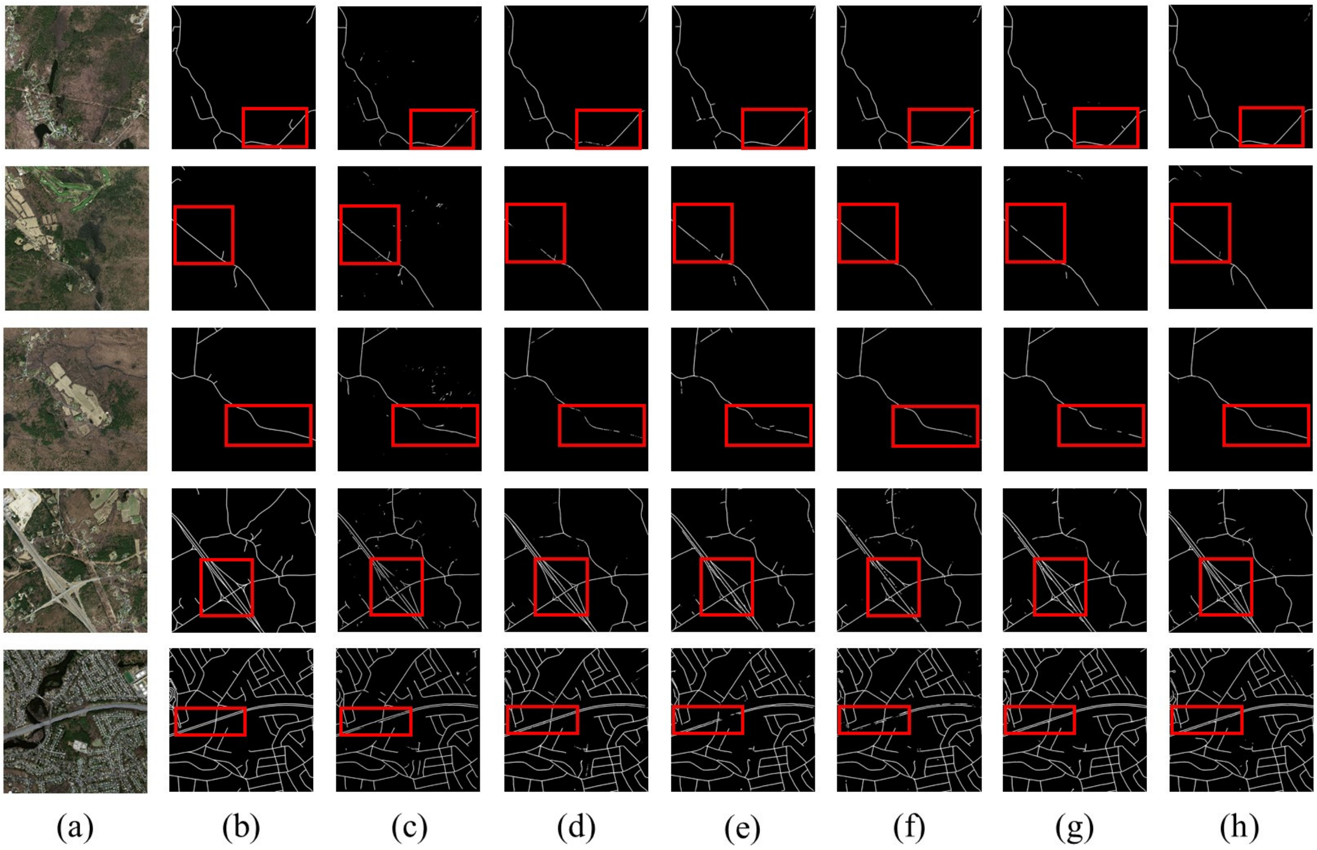

Figure 5. Road extraction results using the Massachusetts Roads Dataset. (**a**) Aerial image; (**b**) Ground truth; (**c**) VGG16; (**d**) VGG16 + RA module; (**e**) Res34; (**f**) Res34 + RA module; (**g**) Res50; (**h**) Res50 + RA module.

Table 1. Quantitative comparison of different encoders adding RA module on the Massachusetts Roads Dataset. The best results for each evaluation metric are highlighted in bold.

Encoder model	*mIOU*	*F1*	*Kappa*
VGG16	0.5325	0.6949	0.6824
Res34	0.5497	0.7094	0.6967
Res50	0.5563	0.7149	0.7023
VGG16 + RA module	0.5708	0.7268	0.7153
Res34 + RA module	0.5605	0.7184	0.7066
Res50 + RA module	**0.5834**	**0.7369**	**0.7255**

To further verify the feature extraction capabilities of different encoders and the performance of the RA module, we compared the extraction effects of the three encoder models on the DeepGlobe Roads Dataset. As shown in Figure 6, the results of the VGG16 model have a large number of fragmented patterns, and it is difficult to extract a complete road network before the RA module is added. There are also fractured road fragments in the results of Res34 and Res50 models. However, the extraction results of the Res50 model are better than the results of Res34 and VGG16 from visual comparison. After adding the RA module, the VGG16 model effectively alleviated the "salt and pepper" phenomenon, indicating that the RA module can highlight the local information of the feature, maintain the boundary information of the road, and effectively distinguish the road from the complex surrounding environment. In addition, the Res34 and Res50 models also extract more complete road networks after adding the RA module.

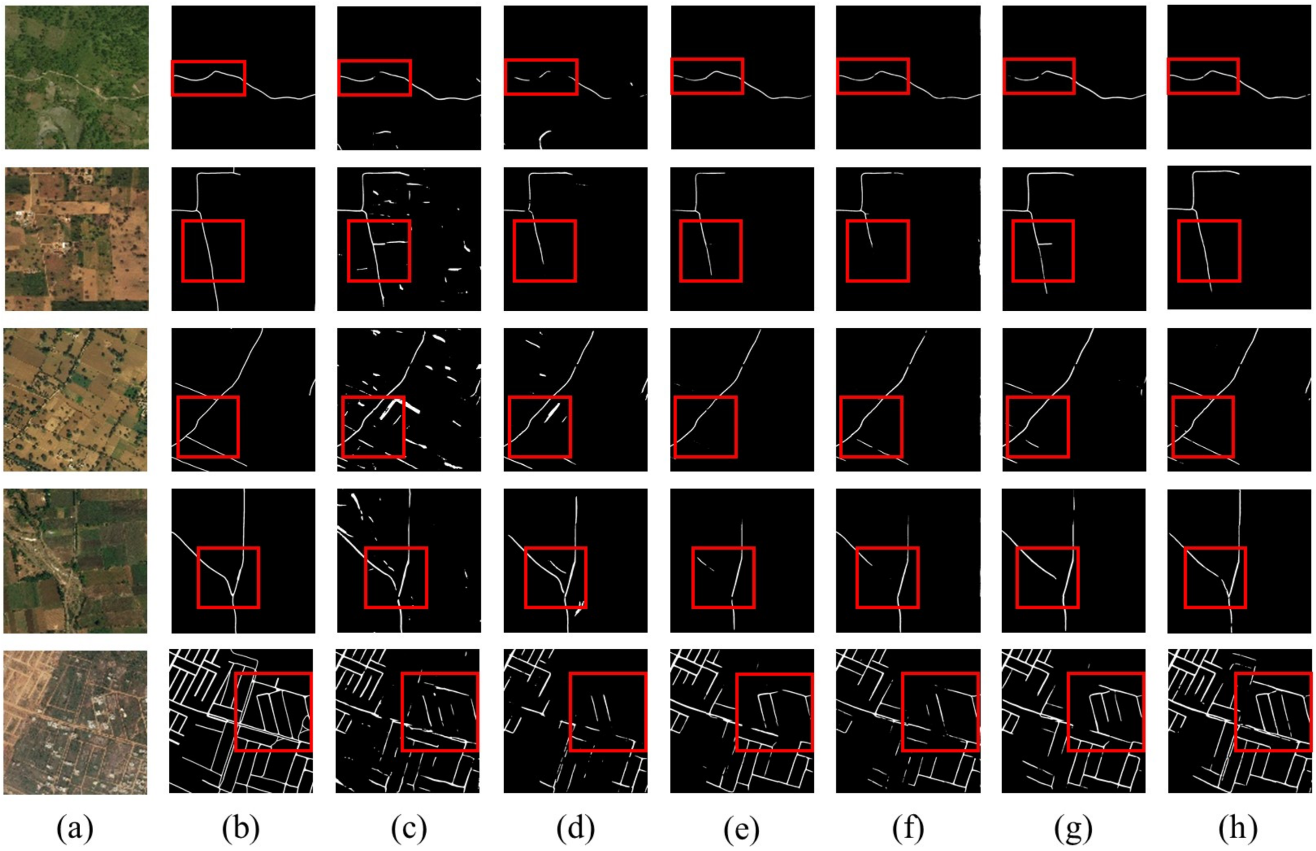

Figure 6. Road extraction results using the DeepGlobe Roads Dataset. (**a**) Satellite image; (**b**) Ground truth; (**c**) VGG16; (**d**) VGG16 + RA module; (**e**) Res34; (**f**) Res34 + RA module; (**g**) Res50; (**h**) Res50 + RA module.

Table 2 shows the quantitative comparison of the three encoder models on the DeepGlobe Roads Dataset. On the whole, the accuracy of the Res50 model is significantly higher than the other two models, which not only illustrates the positive effect of residual connections on the road extraction results but also verifies that the depth of the model directly affects the accuracy of the results. After adding the RA module, the Res50 model reaches the optimal performance, and the values of *mIOU*, *F*1-score, and *Kappa* coefficient reached 0.5643, 0.7214, 0.7107. The experimental results show that the Res50 model has outstanding feature extraction capabilities. Highlighting the characteristics of crucial local information through the RA module can accurately distinguish the difference between the road and the surrounding area and improve the road extraction results.

Table 2. Quantitative comparison of different encoders adding RA module on the DeepGlobe Roads Dataset. The best results for each evaluation metric are highlighted in bold.

Encoder Model	*mIOU*	*F*1	*Kappa*
VGG16	0.5042	0.6703	0.6554
Res34	0.5094	0.6750	0.6638
Res50	0.5280	0.6911	0.6793
VGG16 + RA module	0.5558	0.7145	0.7031
Res34 + RA module	0.5351	0.6972	0.6857
Res50 + RA module	**0.5643**	**0.7214**	**0.7107**

4.3.2. Ablation Study on the Components of the Network

In this section, we use the Massachusetts Roads Dataset to compare the performance of different components of the RALC-Net model. We take the Res50 network as the baseline and then increase the RA module, multi-scale dilated convolution (MD), and multi-feature information (MF), respectively. From Table 3, the addition of the RA module (the second row of Table 3) improves the baseline from 0.5563 to 0.5834 in terms of *mIoU*, indicating that the RA module has a good feature representation ability, can highlight the importance of local perception context, and has a positive effect in the task of road extraction. The *mIOU* increased from 0.5834 to 0.5872 and 0.5917, respectively, after adding multi-scale dilated convolution and multi-feature information in the baseline (the third and fourth row of Table 3). This consequence shows that the multi-scale spatial receptive field extracted by the multi-scale dilated convolution module can abstract and fuse feature information in different spatial dimensions to improve the accuracy of the result, and multi-feature information can provide more helpful semantic information in the task of road extraction. Finally, we combine all the components to obtain the best performance of the RALC-Net model (the fifth row of Table 3), and the *mIOU* reaches 0.5961. In a word, every component in the model is indispensable for obtaining a complete and accurate road extraction result.

Table 3. Ablation study on the Massachusetts Roads Dataset. The best results for each evaluation metric are highlighted in bold.

	Baseline	**RA**	**MD**	**MF**	***mIOU***	***F*1**	***Kappa***
1	√				0.5563	0.7149	0.7023
2	√	√			0.5834	0.7369	0.7255
3	√	√		√	0.5872	0.7399	0.7285
4	√	√	√		0.5917	0.7435	0.7322
5	√	√	√	√	**0.5961**	**0.7470**	**0.7358**

4.3.3. Road Extraction Using Multi-Feature Information

Due to the heterogeneity of the complex environment, the road has strong similarities to the surrounding environment in texture and spectral features, which also makes it difficult to extract a complete and continuous road extraction result. Therefore, we propose to use feature information such as texture, color, and shape as the model's inputs to improve the ability of the network model to extract and abstract deep feature information. In Section 3.4, we introduced the types and extraction methods of feature information used in this article. To further demonstrate the impact of multiple feature information on the road extraction, the proposed RALC-Net is compared with other state-of-the-art models such as SegNet, U-Net, DeeplabV3+, and D-LinkNet in this experiment.

The visual comparisons of different models are shown in Figure 7, and this experiment consists of two parts. One part of the experiment is to use only the original image as input data to extract road information (Figure 7c,e,g,i,k), and the other part is to combine various feature information with the original image as input data (Figure 7d,f,h,j,l). For SegNet and U-Net, there is a salt and pepper phenomenon in the extraction map (Figure 7c,e) using only image data. However, after adding multi-feature information, the result (Figure 7d,f) effectively improves the salt and pepper phenomenon. As for DeeplabV3+, the outcome using only image data (Figure 7g) is more easily misidentified in some areas where the spectral and texture features are similar to the road, which leads to difficultly extracting a complete and continuous road network. Compared with the result using only image data, multi-feature information (Figure 7h) can obtain better prediction outcomes. As shown in Figure 7, the proposed RALC-Net outperforms D-LinkNet having multi-scale dilated convolution after adding multi-feature information, indicating that the advantages of RALC-Net with dual-inputs make fuller use of multi-feature information and improve the generalization ability of the model.

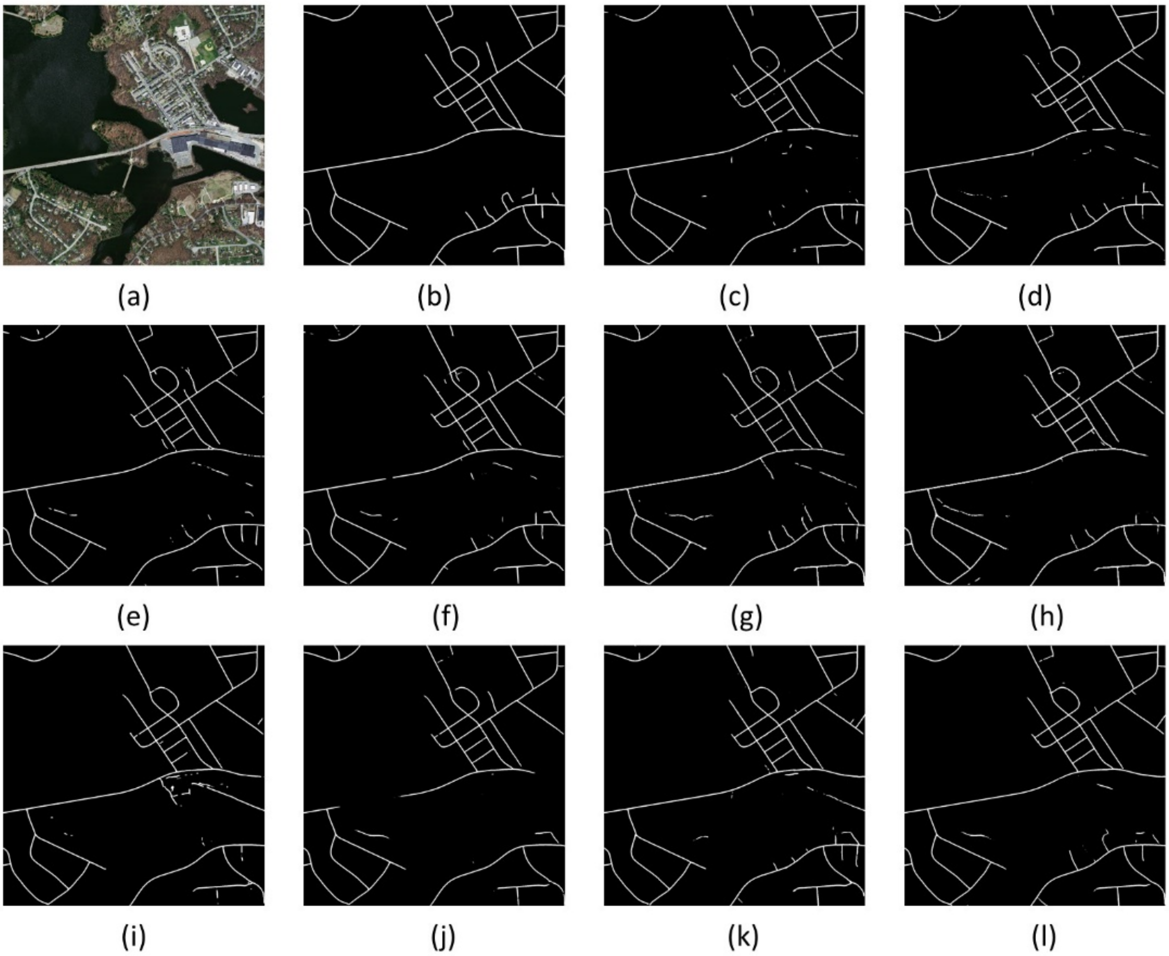

Figure 7. Visual comparisons using only image data and using multi-feature information. (**a**) Aerial image; (**b**) Ground truth; (**c**) SegNet; (**d**) SegNet using Multi-feature; (**e**) U-Net; (**f**) U-Net using Multi-feature; (**g**) DeeplabV3+; (**h**) DeeplabV3+ using Multi-feature; (**i**) D-LinkNet; (**j**) D-LinkNet using Multi-feature; (**k**) RALC-Net; (**l**) RALC-Net using Multi-feature.

Table 4 shows the quantitative comparisons of different models. In the first part of the experiment (the first five rows of Table 4), we can see that the accuracy of D-LinkNet is higher than the other models except for RALC-Net. This is because D-LinkNet has an outstanding multi-scale dilated convolution module, which can extract feature maps at different spatial scales and improve the model's generalization ability. The proposed RALC-Net achieves the best performance, with *mIOU*, *F*1-score, and *Kappa* coefficient reaching 0.5917, 0.7435, and 0.7322, respectively. The evaluation metrics are better than other models because the RALC-Net proposed in this paper includes a multi-scale dilated convolution module and uses the RA module to highlight local information and contact the spatial context, making the model more robust. In another part of the experiment (the last five rows of Table 4), we take multi-feature information as input data. As shown in Table 4, the results of SegNet, U-Net, DeeplabV3+, D-LinkNet, and RALC-Net are all enhanced after adding multi-feature information. The experimental results show that we can accurately distinguish the target road from other areas by using feature information such as texture, color, and shape in a complex environment.

Table 4. Quantitative comparison of different methods using only image data and using multi-feature information. The best results for each evaluation metric are highlighted in bold.

	Method	*mIOU*	*F1*	*Kappa*
part one	SegNet	0.5477	0.7077	0.6957
	U-Net	0.5694	0.7256	0.7140
	DeeplabV3+	0.5439	0.7045	0.6917
	D-LinkNet	0.5765	0.7313	0.7196
	RALC-Net	0.5917	0.7435	0.7322
part two	SegNet using Multi-feature	0.5755	0.7305	0.7188
	U-Net using Multi-feature	0.5800	0.7342	0.7228
	DeeplabV3+ using Multi-feature	0.5632	0.7205	0.7080
	D-LinkNet using Multi-feature	0.5813	0.7352	0.7239
	RALC-Net using Multi-feature	**0.5961**	**0.7470**	**0.7358**

5. Discussion

In Section 4.3.1, we compare the feature extraction capabilities of different encoders and analyze the feature representation capabilities of the RA module in the Massachusetts Roads Dataset and the DeepGlobe Roads Dataset. The experimental results show that taking the Res50 model as the basic structure of the network can reflect outstanding feature extraction capabilities of the encoders and excellent feature representation capabilities of the RA module. As shown in Figures 5 and 6, the three models after adding the RA module are better than the three original models, which reduces the salt and pepper phenomenon and further improves the continuity of road results. The RA module constructed using the residual connection to contact spatial context and the attention mechanism to highlight crucial local information can effectively retain road boundaries and improve the model's ability to extract and abstract feature maps. The RALC-Net model has the characteristics of highlighting the local context by adding the RA module, which can further solve the problem of discontinuous road extraction caused by spatial heterogeneity. Tables 1 and 2 show that the Res50 model combined with the RA module outperforms the other two models, with *mIOU* reaching 0.5834 and 0.5643, respectively. Therefore, comprehensively considering both quantitative and qualitative results, the RALC-Net proposed in this paper selects the Res50 model with excellent feature extraction capabilities as the basic encoder of the network.

In Section 4.3.2, we conduct an ablation study on RALC-Net and comprehensively analyze the necessity of each component of the network model. As shown in Table 3, we discussed the performance of different components through five quantitative comparisons. The RA module can highlight local semantic information and has the best effect on enhancing network performance. The result illustrates that local information plays a vital role in road extraction, and the road network can be extracted more accurately by preserving spatial topological information such as road texture and boundary, and multi-feature information can provide more helpful semantic information for road extraction in complex environments.

High-resolution remote sensing images include a wealth of feature information, and feature information such as color, texture, and shape extracted from remote sensing images helps accurately extract target features from a complex environment. The deep learning model extracts high-level semantic features layer by layer through a multi-layer hierarchical model. As the depth of the network model increases, the underlying feature information is often easily overlooked. However, taking the feature information such as color, texture, and shape as the input of the deep learning model effectively solves this problem, which successfully combines the low-level feature information with the high-level feature information, enhancing the generalization ability and robustness of the network model. As is shown in Figure 7, the visual results after adding multi-feature information have been improved, indicating using multi-feature information can obtain more continuous road network and reduce the salt and pepper phenomenon. Moreover, the accuracy of the extraction result of adding multi-feature information is also significantly better than using only the original image to extract the feature information in Table 4. In Section 4.3.3, we compare the results of RALC-Net with other state-of-the-art models in the second part of the experiment. The experimental results show that the performance of RALC-Net proposed in this paper is better than other models. This outcome also illustrates that the dual-encoder model constructed in this paper has more robust feature extraction capabilities. Compared with other single-encoder models, RALC-Net can obtain more accurate road networks.

6. Conclusions

This study proposes a novel end-to-end residual attention local perception network called RALC-Net, which extracts road information from high-resolution remote sensing images. Specifically, the encoder–decoder structure is used to construct the network model, and the Res50 model is selected as the encoding baseline. A residual attention module is proposed by combining the residual connection and the attention mechanism. We have verified the feature representation ability of the RA module on two public datasets and analyzed the outstanding performance of the RA module from both qualitative and quantitative perspectives. In addition, by integrating the multi-scale dilated convolution module, the RALC-Net can extract comprehensively multi-scale feature information and strengthen spatial context semantic information. It is also combined with the RA module to enhance the performance of highlighting local perceptual information. At the same time, the experimental results show that the dual-encoder structure can enhance the feature extraction capability of the network. In addition, we use the low-level feature information such as color, texture, and shape as the model input and combine the low-level feature information with the high-level abstract semantic features to extract complete road information from the complex environment. Compared with other state-of-the-art models on the Massachusetts Roads Dataset, the experimental results indicate that multi-feature information positively affects the road extraction result and further verify that the RALC-Net proposed in this paper outperforms other models.

Although deep learning technology has made excellent achievements in image processing, its "black box" nature, whose internal information cannot be interpreted, is still challenging to explain. This article uses the characteristics of the attention mechanism to have a specific positive effect on the interpretability of deep learning models, but there are still certain limitations. Future research will address the interpretability problem of deep learning by using other advanced technologies.

Author Contributions: The conceptualization was proposed by Z.L.; F.W. supervised the research and administrated this project; X.J. conducted the investigation and offered supporting algorithms; the methodology and experiments were conducted by Z.L.; the article was written by Z.L.; M.W. professionally optimized the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (Grant No. 42171407, 42077242); Natural Science Foundation of Jilin province (Grant No. 20210101098JC); Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources (Grant No. KF-2020-05-024); Scientific research project of the 13th five-year plan of Jilin province education department (Grant No. JJKH20200999KJ).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Yang:, Y.; Newsam, S. Geographic Image Retrieval Using Local Invariant Features. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 818–832. [CrossRef]
2. Weng, Q.H. Remote sensing of impervious surfaces in the urban areas: Requirements, methods, and trends. *Remote Sens. Environ.* **2012**, *117*, 34–49. [CrossRef]
3. Wang, M.C.; Zhang, X.Y.; Niu, X.F.; Wang, F.Y.; Zhang, X.Q. Scene Classification of High-Resolution Remotely Sensed Image Based on ResNet. *J. Geovisualization Spat. Anal.* **2019**, *3*, 16. [CrossRef]
4. Youssef, A.M.; Sefry, S.A.; Pradhan, B.; Alfadail, E.A. Analysis on causes of flash flood in Jeddah city (Kingdom of Saudi Arabia) of 2009 and 2011 using multi-sensor remote sensing data and GIS. *Geomat. Nat. Hazards Risk* **2015**, *7*, 1018–1042. [CrossRef]
5. Ghaffarian, S.; Kerle, N.; Filatova, T. Remote Sensing-Based Proxies for Urban Disaster Risk Management and Resilience: A Review. *Remote Sens.* **2018**, *10*, 1760. [CrossRef]
6. Heiselberg, H.; Stateczny, A. Remote Sensing in Vessel Detection and Navigation. *Sensors* **2020**, *20*, 5841. [CrossRef]
7. Bi, Q.; Qin, K.; Zhang, H.; Zhang, Y.; Li, Z.; Xu, K. A Multi-Scale Filtering Building Index for Building Extraction in Very High-Resolution Satellite Imagery. *Remote Sens.* **2019**, *11*, 482. [CrossRef]
8. Cardim, G.; Silva, E.; Dias, M.; Bravo, I.; Gardel, A. Statistical Evaluation and Analysis of Road Extraction Methodologies Using a Unique Dataset from Remote Sensing. *Remote Sens.* **2018**, *10*, 620. [CrossRef]
9. Lu, X.; Zhong, Y.; Zheng, Z.; Liu, Y.; Zhao, J.; Ma, A.; Yang, J. Multi-Scale and Multi-Task Deep Learning Framework for Automatic Road Extraction. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 9362–9377. [CrossRef]
10. Jia, J.; Sun, H.; Jiang, C.; Karila, K.; Karjalainen, M.; Ahokas, E.; Khoramshahi, E.; Hu, P.; Chen, C.; Xue, T.; et al. Review on Active and Passive Remote Sensing Techniques for Road Extraction. *Remote Sens.* **2021**, *13*, 4235. [CrossRef]
11. Wang, J.; Song, J.; Chen, M.; Yang, Z. Road network extraction: A neural-dynamic framework based on deep learning and a finite state machine. *Int. J. Remote Sens.* **2015**, *36*, 3144–3169. [CrossRef]
12. Zhang, X.; Han, X.; Li, C.; Tang, X.; Zhou, H.; Jiao, L. Aerial Image Road Extraction Based on an Improved Generative Adversarial Network. *Remote Sens.* **2019**, *11*, 930. [CrossRef]
13. Vo, D.M.; Lee, S.-W. Semantic image segmentation using fully convolutional neural networks with multi-scale images and multi-scale dilated convolutions. *Multimed. Tools Appl.* **2018**, *77*, 18689–18707. [CrossRef]
14. Wang, M.; Zhang, H.; Sun, W.; Li, S.; Wang, F.; Yang, G. A Coarse-to-Fine Deep Learning Based Land Use Change Detection Method for High-Resolution Remote Sensing Images. *Remote Sens.* **2020**, *12*, 1933. [CrossRef]
15. Xu, Z.; Shen, Z.; Li, Y.; Xia, L.; Wang, H.; Li, S.; Jiao, S.; Lei, Y. Road Extraction in Mountainous Regions from High-Resolution Images Based on DSDNet and Terrain Optimization. *Remote Sens.* **2020**, *13*, 90. [CrossRef]
16. Xin, J.; Zhang, X.; Zhang, Z.; Fang, W. Road Extraction of High-Resolution Remote Sensing Images Derived from DenseUNet. *Remote Sens.* **2019**, *11*, 2499. [CrossRef]
17. Xie, Y.; Miao, F.; Zhou, K.; Peng, J. HsgNet: A Road Extraction Network Based on Global Perception of High-Order Spatial Information. *ISPRS Int. J. Geoinf.* **2019**, *8*, 571. [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. Liu, R.; Miao, Q.; Zhang, Y.; Gong, M.; Xu, P. A Semi-Supervised High-Level Feature Selection Framework for Road Centerline Extraction. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 894–898. [CrossRef]
20. 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]
21. 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]
22. Liu, J.; Qin, Q.M.; Li, J.; Li, Y.P. Rural Road Extraction from High-Resolution Remote Sensing Images Based on Geometric Feature Inference. *ISPRS Int. J. Geoinf.* **2017**, *6*, 314. [CrossRef]
23. Cheng, G.; Zhu, F.; Xiang, S.; Pan, C. Road Centerline Extraction via Semisupervised Segmentation and Multidirection Nonmaximum Suppression. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 545–549. [CrossRef]
24. 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.* **2014**, *52*, 3359–3372. [CrossRef]
25. Shanmugam, L.; Kaliaperumal, V. Junction-aware water flow approach for urban road network extraction. *IET Image Process* **2016**, *10*, 227–234. [CrossRef]
26. Baatz, M.; Schape, A. Multiresolution Segmentation: An Optimization Approach for High Quality Multi-Scale Image Segmentation. *Angew. Geogr. Inf. Sev.-Beitung* **2000**, *12*, 12–23.
27. Li, M.; Stein, A.; Bijker, W.; Zhan, Q. Region-based urban road extraction from VHR satellite images using Binary Partition Tree. *Int. J. Appl. Earth Obs. Geoinf.* **2016**, *44*, 217–225. [CrossRef]
28. Huang, X.; Zhang, L. Road centreline extraction from high-resolution imagery based on multiscale structural features and support vector machines. *Int. J. Remote Sens.* **2009**, *30*, 1977–1987. [CrossRef]
29. Miao, Z.; Shi, W.; Gamba, P.; Li, Z. An Object-Based Method for Road Network Extraction in VHR Satellite Images. *IEEE J.-STARS* **2015**, *8*, 4853–4862. [CrossRef]

30. Yin, D.; Du, S.; Wang, S.; Guo, Z. A Direction-Guided Ant Colony Optimization Method for Extraction of Urban Road Information From Very-High-Resolution Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 4785–4794. [CrossRef]
31. Basavaraju, A.; Du, J.; Zhou, F.J.; Ji, J. A Machine Learning Approach to Road Surface Anomaly Assessment Using Smartphone Sensors. *IEEE Sens. J.* **2020**, *20*, 2635–2647. [CrossRef]
32. Zhang, Q.; Kong, Q.; Zhang, C.; You, S.; Wei, H.; Sun, R.; Li, L. A new road extraction method using Sentinel-1 SAR images based on the deep fully convolutional neural network. *Eur. J. Remote Sens.* **2019**, *52*, 572–582. [CrossRef]
33. Lv, Y.; Wang, G.F.; Hu, X.Y. Machine Learning Based Road Detection from High Resolution Imagery. In Proceedings of the 23rd Congress of the International-Society-for-Photogrammetry-and-Remote-Sensing (ISPRS), Prague, Czech Republic, 12–19 July 2016; pp. 891–898.
34. Guo, Q.; Wang, Z. A Self-Supervised Learning Framework for Road Centerline Extraction From High-Resolution Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 4451–4461. [CrossRef]
35. Zhu, D.M.; Wen, X.; Xia, R. Road Extraction Based on the Algorithms of K-Means Clustering and Hybrid Model of SVM and FCM. *Adv. Mat. Res.* **2012**, *518–523*, 5738–5743. [CrossRef]
36. Xu, Y.; Xie, Z.; Wu, L.; Chen, Z. Multilane roads extracted from the OpenStreetMap urban road network using random forests. *Trans. GIS.* **2018**, *23*, 224–240. [CrossRef]
37. Soni, P.K.; Rajpal, N.; Mehta, R. Semiautomatic Road Extraction Framework Based on Shape Features and LS-SVM from High-Resolution Images. *J. Indian Soc. Remote Sens.* **2020**, *48*, 513–524. [CrossRef]
38. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507. [CrossRef] [PubMed]
39. Ma, L.; Liu, Y.; Zhang, X.L.; Ye, Y.X.; Yin, G.F.; Johnson, B.A. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *152*, 166–177. [CrossRef]
40. Paoletti, M.E.; Haut, J.M.; Plaza, J.; Plaza, A. Deep learning classifiers for hyperspectral imaging: A review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *158*, 279–317. [CrossRef]
41. Senthilnath, J.; Varia, N.; Dokania, A.; Anand, G.; Benediktsson, J.A. Deep TEC: Deep Transfer Learning with Ensemble Classifier for Road Extraction from UAV Imagery. *Remote Sens.* **2020**, *12*, 245. [CrossRef]
42. Cheng, G.; Zhu, F.; Xiang, S.; Wang, Y.; Pan, C. Accurate urban road centerline extraction from VHR imagery via multiscale segmentation and tensor voting. *Neurocomputing* **2016**, *205*, 407–420. [CrossRef]
43. Buslaev, A.; Seferbekov, S.; Iglovikov, V.; Shvets, A. Fully Convolutional Network for Automatic Road Extraction from Satellite Imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 197–1973.
44. Li, P.; Zang, Y.; Wang, C.; Li, J.; Cheng, M.; Luo, L. Road network extraction via deep learning and line integral convolution. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 1599–1602.
45. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1337–1342.
46. 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]
47. 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]
48. Zhang, Z.; Liu, Q.; Wang, Y. Road Extraction by Deep Residual U-Net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
49. Abdollahi, A.; Pradhan, B.; Sharma, G.; Maulud, K.N.A.; Alamri, A. Improving Road Semantic Segmentation Using Generative Adversarial Network. *IEEE Access* **2021**, *9*, 64381–64392. [CrossRef]
50. Wei, X.; Fu, X.; Yun, Y.; Lv, X. Multiscale and Multitemporal Road Detection from High Resolution SAR Images Using Attention Mechanism. *Remote Sens.* **2021**, *13*, 3149. [CrossRef]
51. Wang, Z.; Gao, X.; Zhang, Y. HA-Net: A Lake Water Body Extraction Network Based on Hybrid-Scale Attention and Transfer Learning. *Remote Sens.* **2021**, *13*, 4121. [CrossRef]
52. Zhang, H.M.; Wang, M.C.; Wang, F.Y.; Yang, G.D.; Zhang, Y.; Jia, J.Q.; Wang, S.Q. A Novel Squeeze-and-Excitation W-Net for 2D and 3D Building Change Detection with Multi-Source and Multi-Feature Remote Sensing Data. *Remote Sens.* **2021**, *13*, 440. [CrossRef]
53. Xu, Y.Y.; Xie, Z.; Feng, Y.X.; Chen, Z.L. Road Extraction from High-Resolution Remote Sensing Imagery Using Deep Learning. *Remote Sens.* **2018**, *10*, 1461. [CrossRef]
54. 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 Obs. Remote Sens.* **2021**, *14*, 6302–6315. [CrossRef]
55. 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]
56. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; pp. 234–241.

57. 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]
58. Li, X.; Zhang, W.; Ding, Q. Understanding and improving deep learning-based rolling bearing fault diagnosis with attention mechanism. *Signal. Process.* **2019**, *161*, 136–154. [CrossRef]
59. Tan, Z.; Su, J.; Wang, B.; Chen, Y.; Shi, X. Lattice-to-sequence attentional Neural Machine Translation models. *Neurocomputing* **2018**, *284*, 138–147. [CrossRef]
60. Zhang, X.M.; He, G.J.; Zhang, Z.M.; Peng, Y.; Long, T.F. Spectral-spatial multi-feature classification of remote sensing big data based on a random forest classifier for land cover mapping. *Clust. Comput.* **2017**, *20*, 2311–2321. [CrossRef]
61. Mishra, V.N.; Prasad, R.; Rai, P.K.; Vishwakarma, A.K.; Arora, A. Performance evaluation of textural features in improving land use/land cover classification accuracy of heterogeneous landscape using multi-sensor remote sensing data. *Earth Sci. Inform.* **2018**, *12*, 71–86. [CrossRef]
62. Mnih, V. *Machine Learning for Aerial Image Labeling*; University of Toronto: Toronto, TO, Canada, 2013.
63. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 172–181.
64. 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]
65. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 192–196.