International Journal of Geo-Information

MDPI

Article

Cascaded Residual Attention Enhanced Road Extraction from Remote Sensing Images

Shengfu Li [1,2], Cheng Liao [1], Yulin Ding [1,*], Han Hu [1], Yang Jia [2], Min Chen [1], Bo Xu [1], Xuming Ge [1], Tianyang Liu [3] and Di Wu [3]

[1] Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu 611756, China; hustok@sohu.com (S.L.); liaocheng@my.swjtu.edu.cn (C.L.); han.hu@swjtu.edu.cn (H.H.); minchen@home.swjtu.edu.cn (M.C.); xubo@swjtu.edu.cn (B.X.); xuming.ge@swjtu.edu.cn (X.G.)
[2] Sichuan Highway Planning, Survey, Design and Research Institute Ltd., Chengdu 610041, China; yang07031@whu.edu.cn
[3] PLA Key Laboratory of Hydrographic Surveying and Mapping, Dalian Naval Academy, Dalian 116018, China; liutyang@163.com (T.L.); lnwuyashan@163.com (D.W.)
* Correspondence: dingyulin@swjtu.edu.cn

Abstract: Efficient and accurate road extraction from remote sensing imagery is important for applications related to navigation and Geographic Information System updating. Existing data-driven methods based on semantic segmentation recognize roads from images pixel by pixel, which generally uses only local spatial information and causes issues of discontinuous extraction and jagged boundary recognition. To address these problems, we propose a cascaded attention-enhanced architecture to extract boundary-refined roads from remote sensing images. Our proposed architecture uses spatial attention residual blocks on multi-scale features to capture long-distance relations and introduce channel attention layers to optimize the multi-scale features fusion. Furthermore, a lightweight encoder-decoder network is connected to adaptively optimize the boundaries of the extracted roads. Our experiments showed that the proposed method outperformed existing methods and achieved state-of-the-art results on the Massachusetts dataset. In addition, our method achieved competitive results on more recent benchmark datasets, e.g., the DeepGlobe and the Huawei Cloud road extraction challenge.

Keywords: road extraction; remote sensing imagery; semantic segmentation; deep learning; attention mechanism

Citation: Li, S.; Liao, C.; Ding, Y.; Hu, H.; Jia, Y.; Chen, M.; Xu, B.; Ge, X.; Liu, T.; Wu, D. Cascaded Residual Attention Enhanced Road Extraction from Remote Sensing Images. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 9. https://doi.org/10.3390/ijgi11010009

Academic Editor: Wolfgang Kainz

Received: 21 November 2021
Accepted: 26 December 2021
Published: 29 December 2021

1. Introduction

With the rapid development of earth observation technology, large-scale and high-resolution remote sensing imagery has become the most important data source for object extraction. Automatic road recognition and extraction from remote sensing images is an essential step towards many applications, including a high-definition map and the updating of GIS (Geographic Information System) datasets. Despite existing extensive research about automatic road extraction, the accurate and efficient extraction of roads from remote sensing images for GIS applications is still a great challenge. This is partially due to the complexity between roads and backgrounds and partially due to the variation in the width of roads and in the spatial resolution of images [1,2].

Deep learning-based methods can automatically learn and extract representative and distinctive features from a large number of training samples. They have been widely applied in remote sensing because they achieve higher performance than traditional road extraction methods [3–5]. The most popular strategy for road extraction is encoder-decoder-based networks [6–10]. The encoder module extracts multi-scale features from the input images, then the decoder module interprets and upsamples the features end to end for

road extraction. Although recent studies have made great leaps in this regard, there are still some problems that must be solved [3,4,11].

(1) Discontinuity on narrow roads. First, roads are usually linear and continuous. Because there are fewer pixels in the cross-section direction of the road in the image, especially for roads that are narrow, the detailed spatial context is easily lost with repeated downsampling during feature extraction. Existing methods attempt to recover spatial information by fusing the high-resolution features extracted from the shallow convolutional layers using skip connections [9,12,13]. However, the semantic information of features extracted from shallow convolutional layers is insufficient, and it introduces noise that makes the extracted roads discontinuous and jagged, as shown in the yellow dashed ellipse in Figure 1.

(2) Coarse classification at the boundary. Second, existing methods use a specific threshold to classify the upsampled feature maps of roads and backgrounds directly [11,12]. However, threshold-based segmentation usually leads to inaccurate extraction, which is indicated by a zigzag at the boundary of a road. The undesired zigzag effects can be resolved only when the features capture information in a global context rather than locally at the road boundaries, as shown by the red-dashed ellipse in Figure 1.

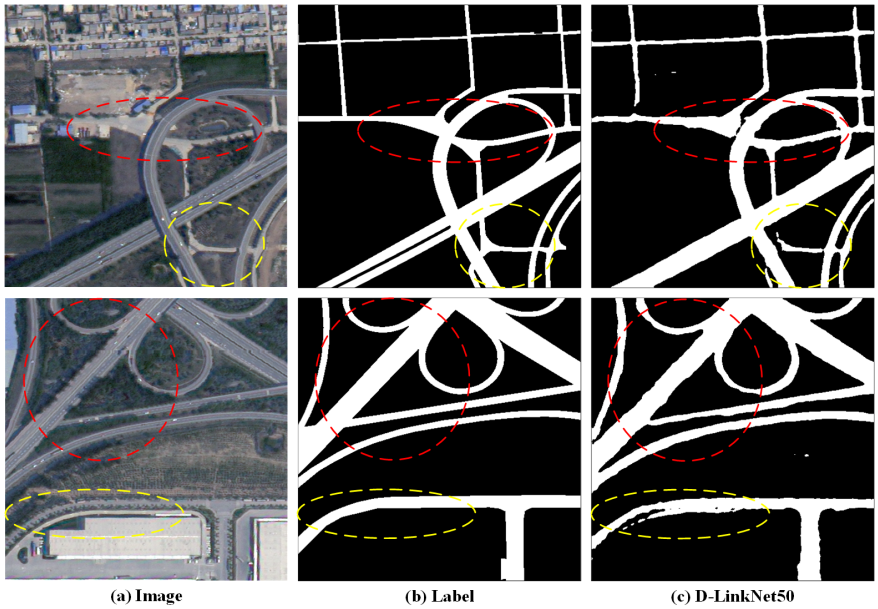


Figure 1. Roads extracted using existing methods are discontinuous and jagged at the boundary, especially when the roads are narrow, as shown in the yellow and red circles, respectively. Columns (**a**–**c**) represent sample images, corresponding labels, and results extracted by D-LinkNet50, respectively.

To address the above problems, we propose a cascaded residual attention-enhanced coarse-to-fine network named CRAE-Net that combines spatial and channel attention modules to optimize and fusion multi-scale features. It preserves the detailed spatial context in high-resolution features and improves threshold-based segments using a lightly refined network to ensure the accuracy of narrow roads and smooth boundary extraction. We designed a parallel multi-path network based on the pretrained ResNet50 to enhance multi-scale features and the combined detailed spatial context and rich semantics. Then, channel and spatial attention were cascaded to optimize and merge multi-scaled features. Additionally, a lightweight network was connected to further refine the boundary of the roads. It sufficiently preserved detailed spatial information and optimized boundary segmentation to achieve accurate road extraction.

In summary, our contributions are as follows: (1) We designed a cascaded attention-based structure to aggregate spatial details and semantic information for continuous road

extraction; (2) we introduced a lightweight coarse-to-fine segment module for smooth road boundary recognition; and (3) we published our benchmark results on the DeepGlobe and the Huawei competition datasets for comparison with related research. The source code is available at: https://github.com/liaochengcsu/Cascade_Residual_Attention_Enhanced_for_Refinement_Road_Extraction, accessed on 26 December 2021.

The rest of this paper is organized as follows. Related works on road extraction are summarized in Section 2. Section 3 introduces the details of the proposed method for refined road extraction. Section 4 describes the experiments and analyzes the results. Section 5 presents the conclusion of this paper.

2. Related Works

Machine learning technology has developed rapidly in recent years, especially with the proposal of the Fully Convolutional Network (FCN) [13], which is a milestone in the field of image processing research and has achieved good results for efficient image segmentation. There have been many deep-learning-based studies related to remote sensing segmentation recently [2,14–26]. For the most relevant problems in road extraction research , we briefly review related works, including refined road boundary extraction and continuous road regional recognition.

Refined road boundary extraction. The U-Net [27] designed encoder-decoder architecture is based on an FCN for medical image segmentation. It introduces skip connections to recover the spatial details lost by the downsampling options at the encoder stage. Many U-Net-derived works [9,28–31] have achieved excellent performance, especially for road boundary extraction, because the potential spatial information represented the details of the road boundary in the shallow layer was focused . In addition, SegNet [6] further refined spatial details and reduced the complexity of computation compared with the FCN.

The ResNet [32] introduced the residual connection to solve the problem of exploding and disappearing gradients in deep convolutional networks. This residual block makes highly complex Convolutional Neural Networks (CNNs) converge faster and more stably. Many studies achieved significant performance by introducing the pre-trained ResNet backbone. ResU-Net [33] improved the accuracy of segmentation significantly by introducing the residual blocks to U-Net. DenseNet [34] designed densely connected blocks using short paths between shallow layers and deep layers for feature reuse, and performed with lower computational complexity than the ResNet. The performance of road extraction was improved using dense blocks in [29,35]. The Coord-Dense-Global (CDG) model [36] introduced an attention mechanism to enhance the edge information and global context of roads based on DenseNet. For road boundary refinement, Conditional Random Fields (CRF) were introduced as post-processing strategies to optimize the extracted result [37,38]. Ref. [39] proposed a coarse-to-fine algorithm, utilizing gray-value distribution to pre-segment the potential roads and using structure context features for final road extraction.

Because the spatial context details are lost in the downsampling operation during encoding, it is difficult to recover the spatial context, especially for narrow road boundaries. Moreover, most existing segmentation-based methods directly classify pixels using a fixed threshold and ignore the structure of the edge, causing severely zigzagged boundaries.

Continuous road regional recognition. Prior knowledge of a road, such as its orientation or topological information, has been used for continuous road extraction [11,35,40–45]. These constraints have been proven effective for road extraction. A series of DeepLab [7,38] methods introduced atrous convolution for segmentation tasks, which enlarges the receptive field without increasing the computational complexity and improves the regional consistency. Ref. [12] introduced an Atrous Spatial Pyramid Pooling (ASPP) module to capture multi-scale global semantic information for efficient road extraction. Moreover, Ref. [8] applied structural similarity loss to improve the continuity of the extracted roads based on multi-scale features.

The attention mechanism is another popular method for improving regional continuity, Refs. [31,46–50] introduced spatial and channel attention layers, which effectively improved

the segmentation performance, especially for continuity in the road area. Benefiting from the Generative Adversarial Networks (GAN), Refs. [40,51–53] obtained impressive results from images through adversarial machine learning between generative and discriminative models. D-LinkNet [54] introduced dilated convolutional layers in the center part of a pre-trained encoder-decoder structure to enlarge the receptive field with efficient computation and memory without reducing the resolution of the features, and achieved first place in the CVPR DeepGlobe Road Extraction Challenge [55].

However, for road extraction research, inaccurate boundary recognition and discontinuous segmentation results, especially for narrow roads, are still unavoidable.

3. Materials and Methods

3.1. Overview

In this work, we propose a cascaded attention-aggregated refinement network to extract roads from remote sensing images. It aims to solve the problems of discontinuous road extraction and jagged boundary identification in existing methods. The main structure of the proposed method is illustrated in Figure 2.

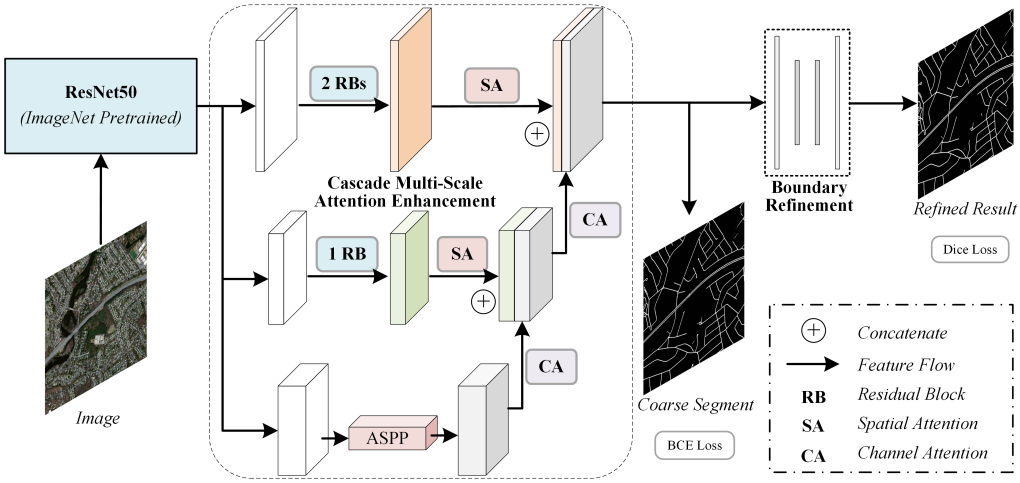


Figure 2. Structure of the proposed method.

Specifically, we base our study on the ResNet50 backbone pre-trained on ImageNet. (1) For the discontinuity of extracted roads, we introduce a cascaded attention-based residual module to enhance extracted multi-scale features, especially for maintaining detailed spatial information. Because the high-resolution features extracted from shallow convolutional layers contain spatial details that are vital for small object recognition, such as narrow roads, it may introduce noise for the insufficiently represent features extracted through a shallow convolutional layer. The designed module not only combines exact spatial details and rich semantic information, but also improves the ability to capture the long-distance similarity of roads. (2) For jagged road boundary recognition, we designed a lightweight U-Net-like network to refine the boundaries of roads in the original scale, and we achieved, without introducing much computational complexity, smoother road boundary extraction than the existing methods that directly filter using a threshold.

3.2. Cascaded Attention Feature Enhancement

The attention mechanism was first proposed to address the bottleneck problem that arises with the use of a fixed-length encoding vector [56], where the decoder would have limited access to the information provided by the input. It uses a weighted sum of all of the encoder hidden states to flexibly focus the attention of the decoder on the most relevant parts of the input sequence, which greatly improved the performance of the sequence model, especially for machine translation. It has been successfully transferred to image processing applications, especially for semantic segmentation in recent years, and

significantly improved the performance in remote sensing image processing. The attention mechanism is mainly divided into spatial attention and channel attention in image processing applications. The spatial attention aims to capture long-distance correlation using a space pixel-by-pixel similarity calculation, while the channel attention is mainly used to assign weights to each feature channel by calculating the correlation in channel levels.

For completeness, we briefly introduce the attention mechanisms used in our work, including the spatial attention module and channel attention module [46–48], as shown in Figure 3.

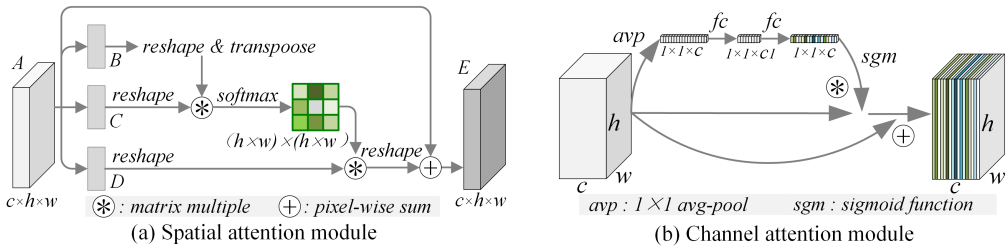


Figure 3. The detail of spatial attention module and channel attention module.

In this work, there are four scales of features extracted by the ResNet50 pre-trained backbone. Features having high resolution that are extracted from shallow layers maintain great spatial detail; features having low resolution that are extracted from deeper convolution layers containing rich semantic features with spatial information loss. Roads usually appear as long and narrow linear structures in remote sensing images. There are few pixels in the cross-section of the road due to the limited road width. Therefore, we only utilize the three scales of features having a high resolution, which better preserves the necessary road spatial information. In order to enlarge the receptive field and extract a wide range of continuous roads, we introduced the ASPP module in the path with the highest feature resolution to obtain more global features.

Since the features with high resolution flow through fewer network layers, which retain more spatial information while introducing noise information and make the road edge appear jagged. We added the cascaded residual blocks represented as RB to extract rich semantic features and preserve the spatial details by maintaining the spatial resolution of features. Furthermore, the Spatial Attention (SA) layer is introduced to capture the long-distance similarity of roads as well as to enhance the consistency of the characteristics of the road, especially for the narrow roads. At the decoding stage, the enhanced multi-scale features are fused through skip connections, and at the same time, the Channel Attention (CA) layer is utilized to perform channel-level filtering on the multi-scale features at the upsampling stage and to obtain features with aggregated rich semantic and accurate spatial details. The details of these blocks were shown in Figure 4.

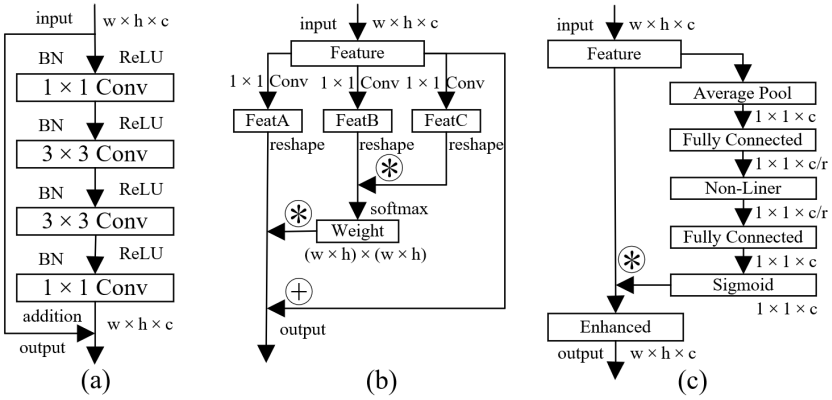


Figure 4. Details of the related blocks. (**a**) The residual block (RB). (**b**) The spatial attention block (SA). (**c**) The channel attention block (CA).

Considering the roads are narrow and continuous in remote sensing imagery, the tiny road is susceptible to interference from nearby background pixels, causing the problem of extracting discontinuity results. We enhance the semantic features of the road area by introducing a spatial attention mechanism to capture the long-distance correlation of roads, which improves the continuity of narrow roads significantly. Besides, the decoder of semantic segmentation networks fuses the features through skip connection by spatial addition or channel concatenation. It ignores the capability of extracting spatial details and semantic features by features with different scales. Based on this investigation, we introduced the channel attention mechanism to realize the adaptive fusion of features at different scales, and optimizes the spatial detail and semantic information of features, thereby enhancing the feature for road representation. The comparison of different feature fusion methods is shown in Figure 5.

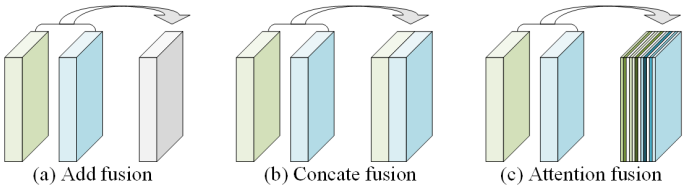


Figure 5. Comparison of different feature fusion methods. (**a**) Addition fusion. (**b**) Concatenation fusion. (**c**) Channel attention fusion.

3.3. Coarse-to-Fine Boundary Optimization

Almost all of the decoder strategies used in existing segmentation-based methods upsample the feature maps to the same scale as the input and classify the pixels of a road or background according to a specific threshold. As roads are always obscured by another object such as a tree or building shadows, especially at the edges of roads so that the context characteristics of the road boundary are unsmooth, making the boundary of the roads difficult to be recognized as smoothly as realistically. The yellow dash circle in Figure 6, from (a) to (d), represent the original imagery, visualizing a feature map of the last layer in DLinkNet50, the threshold-based segmentation on the feature map, and the labels, respectively. Obviously, the roads extracted based on the threshold are coarse at the boundary of occluded roads.

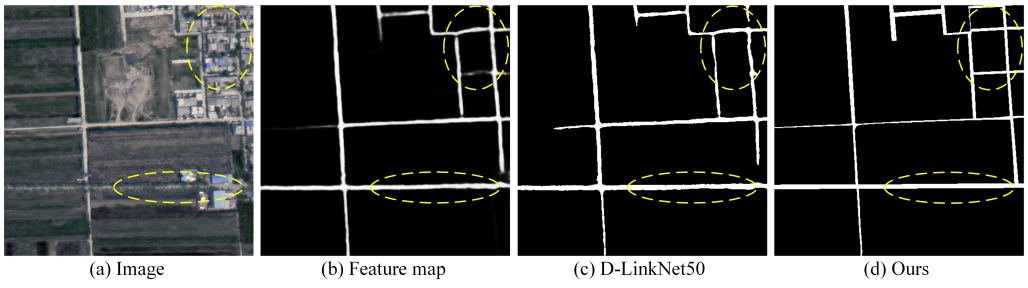


Figure 6. Comparison of results of the existing threshed-based method and ours.

The binary cross-entropy loss function L_bce represented as Formula (1) is usually used to evaluate the distance between predicted results and true labels in binary segmentation applications. Since, the L_bce is a pixel-based metric that does not consider the overall consistency of the prediction results and the labels. The detailed structures of roads that characterize boundaries are valuable in the field of remote sensing. Therefore, we designed a U-Net-like network optimized using the Dice [57] loss to refine the upsampled feature maps. The Dice loss function was introduced to optimize output features, and is represented as L_dice in Formula (2). The Dice loss function measures the similarity between the prediction results and the labels; thus, the boundaries of the road could be further optimized.

The added U-Net is lightweight, considering the complexity of the model. It is connected at the end of a segmentation branch and includes two encoding units with a low number of feature channels to refine the enhanced features efficiently, as shown in Figure 7. The cross-entropy loss function optimizes feature extraction, and the Dice loss function achieves road boundary refinement, which forms a coarse-to-fine road extraction structure. The total loss is the sum of the two loss functions as shown in Formula (3). Because the L_bce is much smaller than L_dice, we set the weights α and β to 4 and 1, respectively, in our experiment to keep a balanced weight between the segmentation and optimization branch.

$$L_{bce} = -\frac{1}{N}\Sigma_i[y_i * \log(p_i) + (1 - y_i) * \log(1 - p_i)] \tag{1}$$

$$L_{dice} = 1 - \frac{2\Sigma_i^N y_i * p_i + 1}{\Sigma_i^N y_i^2 + \Sigma_i^N p_i^2 + 1} \tag{2}$$

$$Loss = \alpha * L_{bce} + \beta * L_{dice} \tag{3}$$

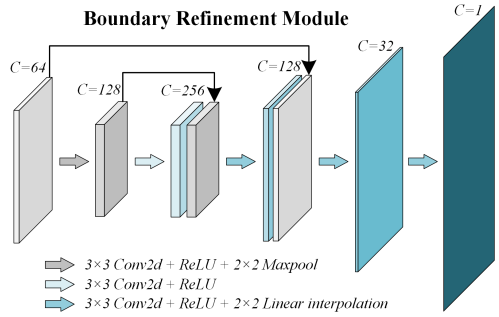


Figure 7. Detail of lightweight road boundary refinement block.

3.4. Evaluation Metrics

Semantic segmentation-based road extraction from remote sensing images is a typical binary segmentation task that seeks to classify every pixel as the road or background. Thus, we evaluated the performance of the proposed method through general semantic segmentation-based evaluation metrics.

The prediction results comprise four cases, including correctly classified positive samples, incorrectly classified positive samples, correctly classified negative samples, and incorrectly classified negative samples. They are represented as True-Positive (TP), False-Positive (FP), True-Negative (TN), and False-Negative (FN), respectively. Precision, Recall, F1-Score, and Intersection over Union (IoU) are four common comprehensive evaluation metrics based on the above indicators. Their calculations are shown in Formulas (4)–(7).

Because the test set includes multiple images, we calculate the accuracy of each image separately, and we finally average the evaluation results of all the images.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{F1} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

$$\text{IoU} = \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall} - \text{Precision} * \text{Recall}} \tag{7}$$

4. Experiments and Results

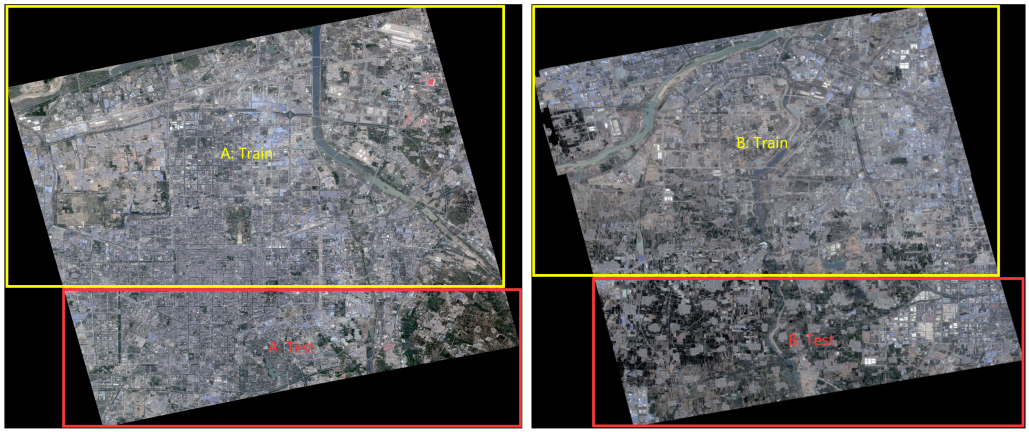
4.1. Datasets and Strategies

In this section, we describe the experimental datasets, including the Massachusetts Road Dataset [58], the DeepGlobe Road Extraction Challenge dataset [55], and the Huawei Cloud competition dataset [59].

The Massachusetts dataset contains 1171 tiled images, including 1108 for training, 14 for validation, and 49 for testing. The size of all images is 1500 × 1500 pixels with a resolution of 1 m.

The DeepGlobe dataset contains 8570 images, including 6226 for training, 1243 for validation, and 1101 for testing, that were captured over Thailand, Indonesia, and India. We only utilized the training subset in our experiment because the corresponding labels of the other two subsets were not available. The size of all the images is 1024 × 1024 pixels with a resolution of 0.5 m. We randomly split these images into 4316 for training, 617 for validation, and 1293 for testing according to the ratio 7:1:2.

The road extraction dataset for the Huawei Cloud competition contains training and testing subsets. We only utilized the training set in our experiments because the labels for the testing set are private. There are two tiles of images from the Beijing2 satellite with a resolution of 0.8 m in the training set. The size of the images is 40,391 × 33,106 and 34,612 × 29,810 pixels, respectively. We divided the images into training and testing areas according to a 7:3 ratio, as illustrated in Figure 8. We clipped the test areas to patches of 1500 × 1500 pixels with an overlap of 18 pixels for testing, considering the limitations of a video memory size. Although it is widely acknowledged that a larger overlapping region and fusing results using voting strategy will significantly improve performance, we want to directly examine the performances without extensive engineering.



(a) Area A (b) Area B

Figure 8. Road extraction dataset of Huawei Cloud competition.

We trained our model on only the training set and tested the performance on the testing set. The validation set was used only to validate the method on the above three datasets. We clipped the training set to patches of size 512 × 512 during the training stage, with an adapted overlap due to the limitations of video memory size. The test and validation sets were predicted and cropped to the original size to evaluate the accuracy.

Our experiments were conducted on a server with single 2080Ti GPU. The Adam was used to optimize the models with the recommended hyper-parameters, and the initial learning rate was 0.01. We adopted a data augmentation strategy with a random flip in the vertical and horizontal directions, color jitter, and random rotation during the training stage. We utilized the Test Time Augmentation (TTA) strategy in the testing stage for all comparison methods. The predict result was the arithmetic mean of the predictions, including the original as well as the horizontally and vertically flipped images.

4.2. Performance on Massachusetts

To verify the proposed method, we conducted comparative experiments on the Massachusetts dataset to compare the performance of our method with other classical semantic segmentation methods, including the U-Net [27], SegNet [6], Res-U-Net [28], DeepLabv3+ [7], and D-LinkNet [54] series methods. Res-U-Net introduces the residual block based on the U-Net. D-LinkNet34 is based on ResNet34 pre-trained on ImageNet, and the backbone of DeepLabv3+ and D-LinkNet50 is ResNet50 pre-trained on ImageNet.

The experimental results are shown in Table 1. Our proposed method achieved significant improvement compared with the classical semantic segmentation methods and achieved SOTA (State-of-the-Art) results on the Massachusetts test dataset. Furthermore, the proposed method gained a large performance margin over D-LinkNet50 and achieved a 4.92% improvement in IoU metrics compared with D-LinkNet34, which attained first place in the CVPR DeepGlobe 2018 Road Extraction Challenge. We made the best result in bold under each metrics.

Table 1. The performance comparison of related methods on the Massachusetts test dataset.

Methods	Precision (%)	Recall (%)	F1-Score (%)	IoU (%)
U-Net	77.71	66.09	71.19	55.66
SegNet	78.89	67.73	72.25	57.02
Res-U-Net	**80.76**	71.49	75.69	61.21
DeepLabv3+	75.47	77.97	76.43	62.25
D-LinkNet34	72.77	79.53	75.75	61.35
D-LinkNet50	73.38	**82.66**	77.51	63.63
Ours	80.04	79.35	**79.52**	**66.27**

For the convenience of comparing the extraction results of related methods, we illustrate the extraction sample and local details in Figure 9 and Figure 10, respectively. Subparts (a) to (h) show the results of related methods in Figure 9. In Figure 10, subparts (c) to (i) display the local details of the proposed method and the results for U-Net, SegNet, Res-U-Net, DeepLabv3+, D-LinkNet34, and D-LinkNet50 corresponding to the yellow circle marked in (a). From the visualized local detail results, it can be inferred that our method extracted a smoother boundary than the comparison methods.

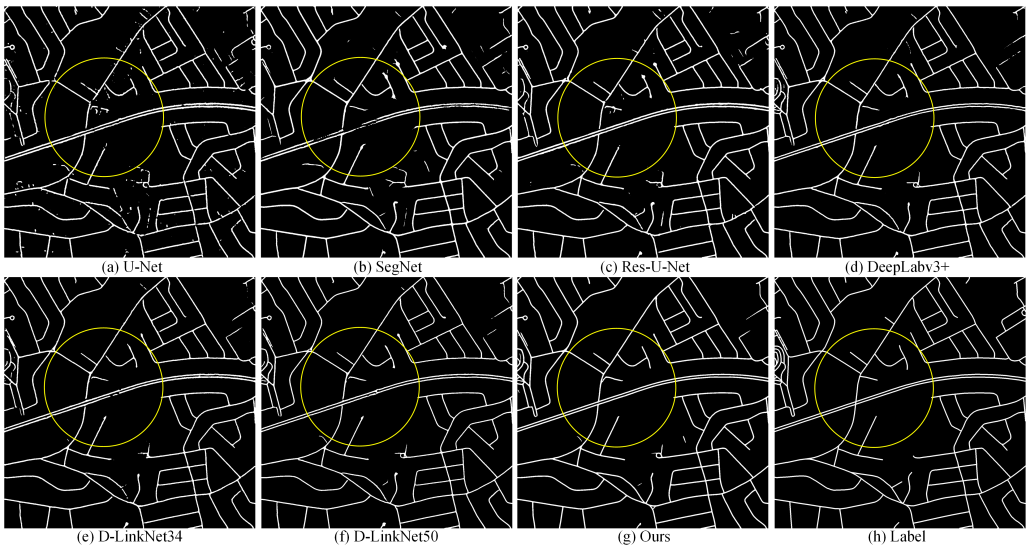


Figure 9. Extracted sample of related methods on the Massachusetts dataset.

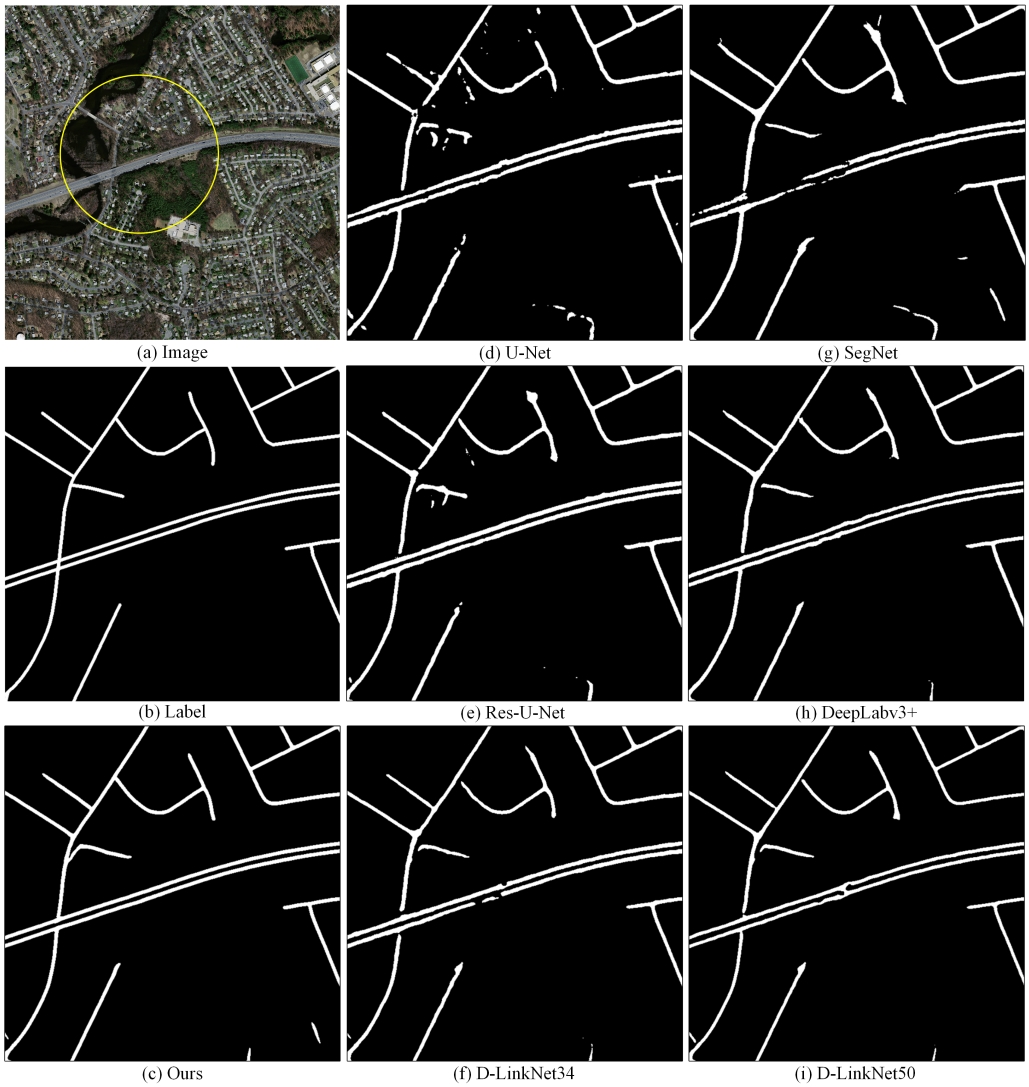


Figure 10. Local detail comparison on the Massachusetts dataset. Subparts (**d**–**i**) show the results of related methods corresponding to the area in the original image (**a**) marked with yellow. Subparts (**b**,**c**) are the ground truth and our method.

To further verify the performance of our proposed method, we evaluated it with the latest related road extraction methods on the Massachusetts dataset. Numerous studies have referenced this dataset, but the applied evaluation metrics are not uniform. Therefore, we select methods include [12,30,36,60,61] for a fair comparison. All of these methods were proposed within the last two years. The TSKD-Road [60] proposed a lightweight knowledge distillation-based topological space network to produce more continuous roads and maintain low computational complexity and network parameters. The CDG [36] introduced an attention mechanism based on DenseNet to enhance the edge information in the global context of the road. DGRN [12] designed a global dense residual network to aggregate abundant multi-scale features based on the ASPP for remedying the loss of spatial features. JointNet [61] combined dense connections with dilated convolution to enlarge the receptive field, which enabled the extraction of multi-scale roads effectively. RB-UNET [30] proposed a reconstruction biased U-Net for road extraction to capture rich semantic information from multiple upsampling operations.

Because the source code of the related works were not available, we referenced their reported results. As shown in Table 2, the proposed method achieved significant improvement both in the F1 Score and IoU metrics. Ours outperformed the latest related studies and achieved SOTA results on the Massachusetts test dataset. We made the best result in bold under each metrics.

Table 2. The performance comparison of latest related methods on the Massachusetts test dataset.

Methods	Precision (%)	Recall (%)	F1-Score (%)	IoU (%)
TSKD-Road [60]	69.85	**79.38**	74.15	59.16
CDG [36]	81.41	71.80	76.10	61.90
DGRN [12]	**81.84**	71.97	76.59	62.48
JointNet [61]	85.36	71.90	78.05	64.00
RB-UNET [12]	79.14	78.53	78.83	65.06
Ours	80.04	79.35	**79.52**	**66.27**

4.3. Performance on DeepGlobe

We designed an experiment on the DeepGlobe dataset, which contains high-resolution images, to evaluate the performance of the proposed method. Because there is no uniform data division performed by the related road extraction method, we compared our method with only the classical semantic segmentation methods introduced above.

Although many studies have used this dataset for experiments, they do not make public the details of the data split, which makes it impossible to compare performance between related studies. Therefore, we provide the details of data division and a new benchmark, which is convenient for comparisons in follow-up research. The experiment result is shown in Table 3. We made the best result in bold under each metrics.

Table 3. The performance comparison of related methods on the DeepGlobe test dataset.

Methods	Precision (%)	Recall (%)	F1-Score (%)	IoU (%)
U-Net	75.13	55.53	62.00	46.30
SegNet	54.45	76.38	60.00	45.45
Res-U-Net	**84.97**	64.13	71.95	57.36
DeepLabv3+	78.20	76.24	75.69	62.33
D-LinkNet34	77.74	78.65	77.04	63.98
D-LinkNet50	80.18	77.46	77.72	64.85
Ours	79.12	**80.39**	**78.75**	**66.38**

The results show that our method outperforms the comparison methods significantly both on the F1 Score and IoU metrics. In particular, our method achieved a 1.53% improvement over the IoU metric achieved by D-LinkNet50. For the convenience of comparing the extracted results of related methods, we illustrate the extracted sample and the exact local details in Figure 11 and Figure 12, respectively. In Figure 12, subparts (c) to (i) represent the local detail of the proposed method and the results for U-Net, SegNet, Res-U-Net, DeepLabv3+, D-LinkNet34, and D-LinkNet50 that correspond to the yellow circle marked in (a). According to the visualized results, the extracted results of the proposed method are more continuous than those of the comparison methods. Moreover, the road boundaries of our method are smoother than those of others.

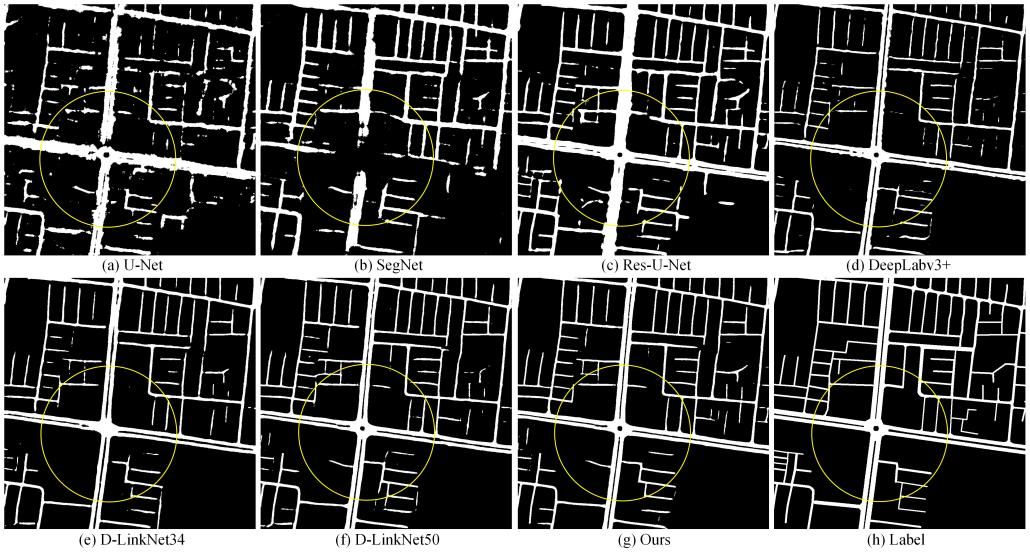


Figure 11. Extracted samples of related methods on the DeepGlobe dataset.

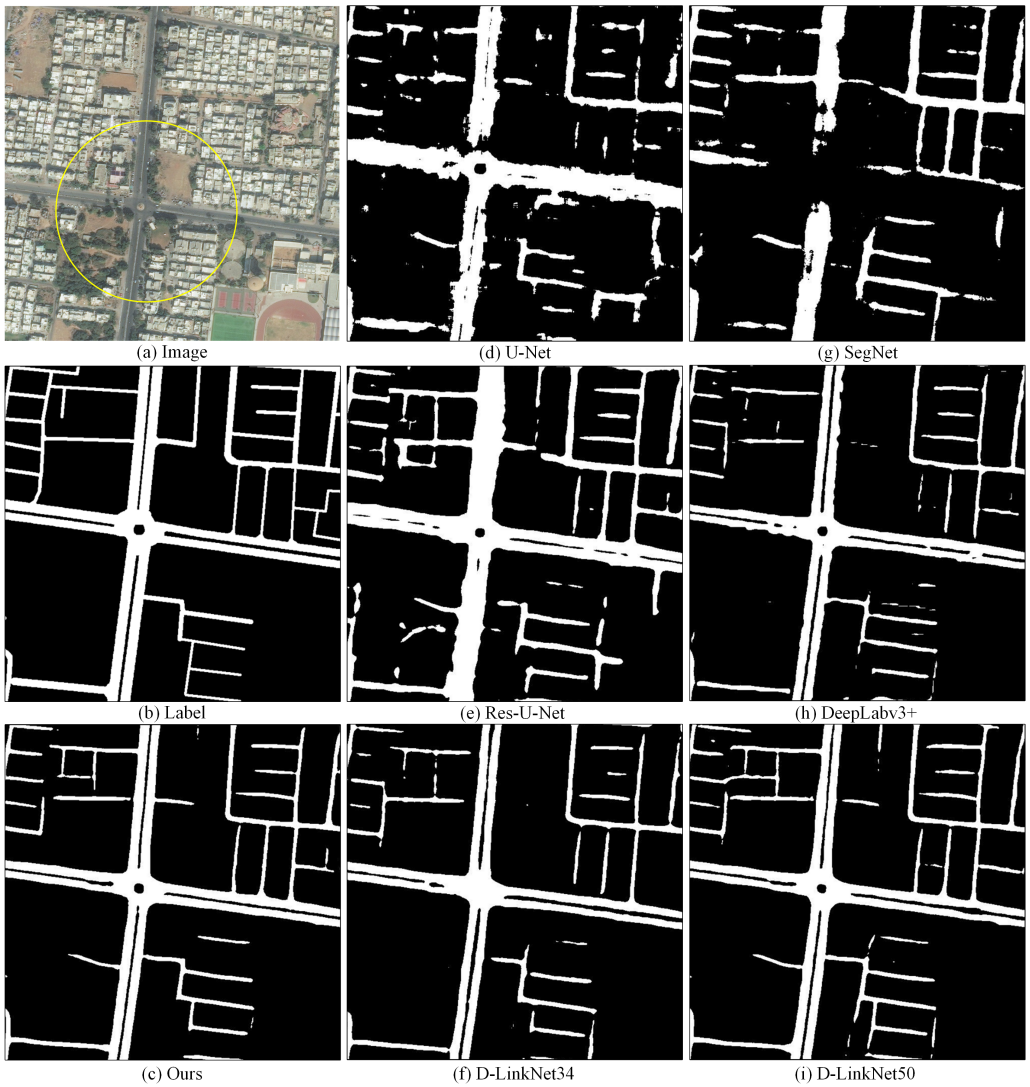


Figure 12. Local detail comparison on the DeepGlobe dataset. Subparts (**d**–**i**) show the results of the related methods corresponding to the area in the original image (**a**) marked with yellow. Subparts (**b**,**c**) are the ground truth and our method.

4.4. Performance on Huawei Cloud

We earned second place in the preliminaries and first place in the finals of the Huawei Cloud competition [59] for road extraction from remote sensing images based on the idea of this works. To validate the advantage of the proposed method and provide a benchmark on this dataset, we designed an experiment to compare our method and the above classical segmentation methods on the training set. The results are shown in Table 4. Our method outperformed the comparison methods on the Huawei test dataset. We made the best result in bold under each metrics.

Table 4. The performance comparison of related methods on the Huawei test dataset.

Methods	Precision (%)	Recall (%)	F1-Score (%)	IoU (%)
U-Net	60.48	66.01	61.35	46.25
SegNet	51.34	70.54	58.43	42.90
Res-U-Net	71.54	65.07	67.45	52.07
DeepLabv3+	72.28	69.59	70.20	55.57
D-LinkNet34	73.70	68.35	70.35	55.72
D-LinkNet50	73.27	70.44	71.26	56.83
Ours	**74.16**	**71.91**	**72.36**	**57.93**

For the convenience of comparing the extracted results for related methods, we also present some samples and local details in Figure 13 and Figure 14, respectively. The results of the compared works and those of our method are shown in subparts (a) to (h) of Figure 13. In Figure 14, subparts (c) to (i) represent the local details of the proposed method and the results for U-Net, SegNet, Res-U-Net, DeepLabv3+, D-LinkNet34, and D-LinkNet50 corresponding to the yellow circle marked in (a).

Overall, the visualized results of the above three datasets show that the introduced attention-based residual module and the lightweight road optimization network significantly improved the continuity of roads as well as the smoothness of the road boundaries.

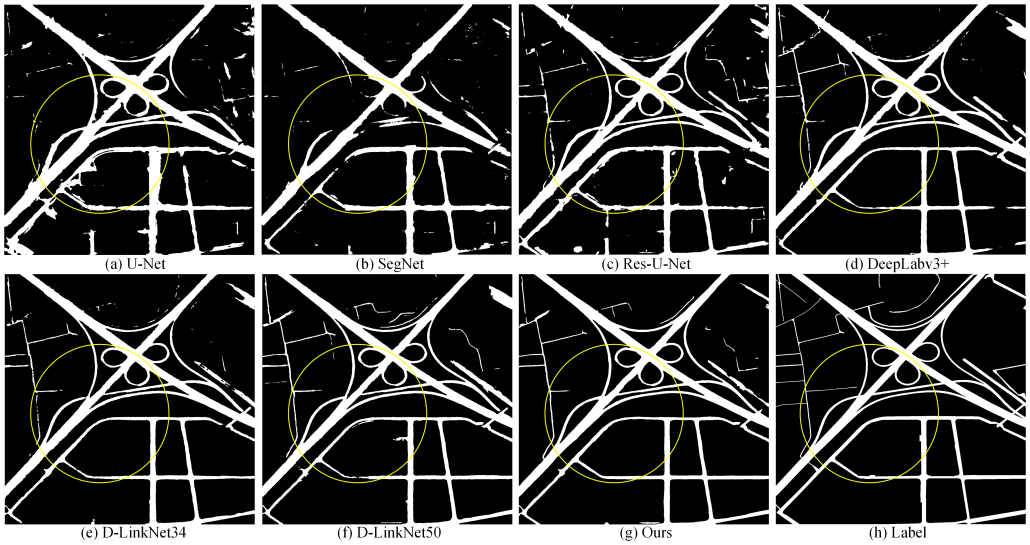


Figure 13. Extracted samples for related methods on the Huawei dataset.

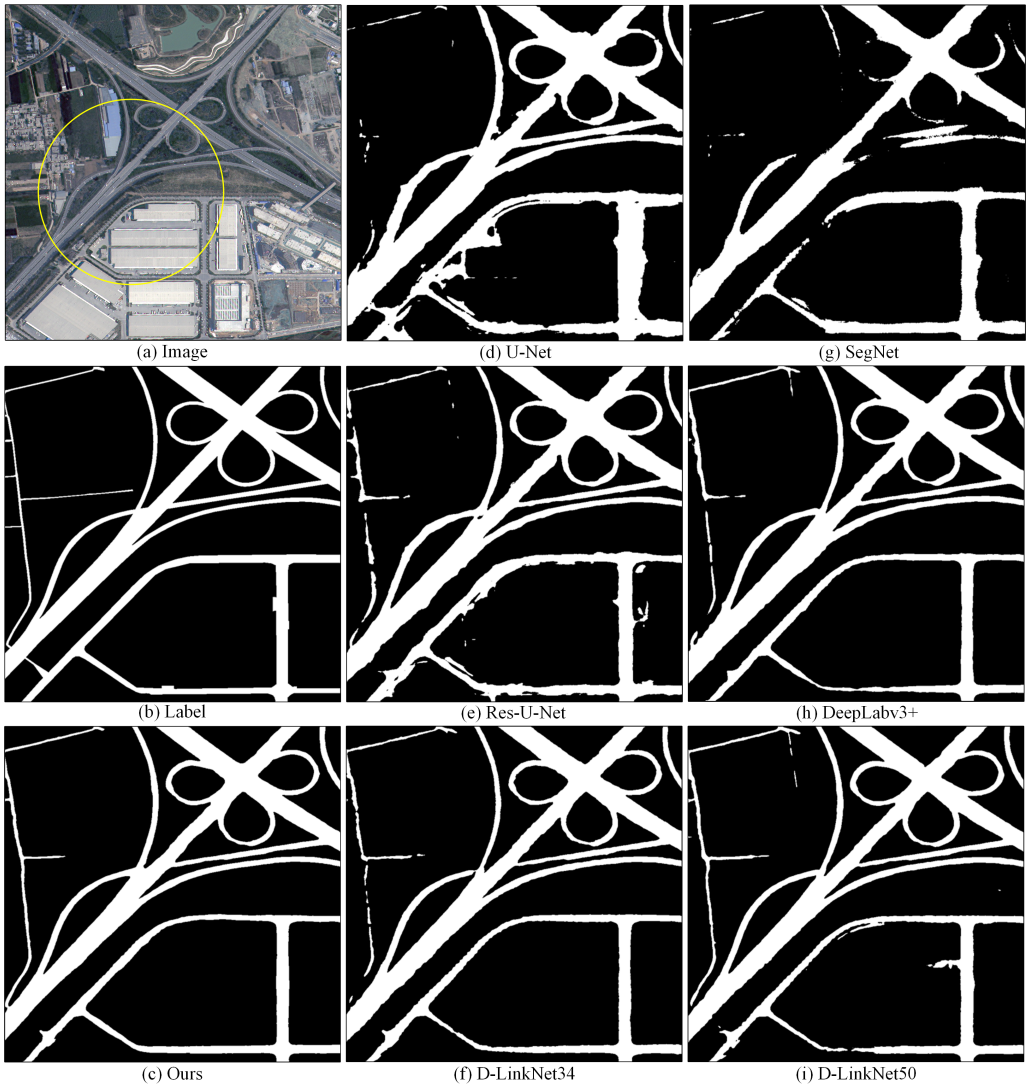


Figure 14. Local detail comparison on Huawei dataset. Subparts (**d**–**i**) are the results of the related methods corresponding to the area in the original image (**a**) marked with yellow. Subparts (**b**,**c**) are the ground truth and our method.

4.5. Ablation Study

Our method introduced attention-based cascaded residual blocks to enhance multi-scale spatial details and semantic features. Furthermore, the lightweight U-Net is connected at the end of the network to optimize the boundary of the roads. Thus, we can extract multi-scale roads from multi-resolution remote sensing imagery accurately. To evaluate the effect of each module of the proposed method, we designed an ablation experiment on the DeepGlobe dataset. We selected the D-LinkNet50 as our baseline because our network is based on it. The attention-based residual block and lightweight refinement network are represented as Att_B and Ref_B, respectively. The experimental result is shown in Table 5.

Table 5. Ablation validation on the DeepGlobe test dataset.

Methods	Precision (%)	Recall (%)	F1-Score (%)	IoU (%)
BaseLine	80.18	77.46	77.72	64.85
BaseLine+Att_B	79.95	78.94	78.25	65.69
BaseLine+Ref_B	81.68	76.82	78.28	65.62
Ours	79.12	80.39	78.75	66.38

The experimental result shows that the introduced modules significantly improved the performance of the network. Att_B obtained a 0.84% IoU improvement relative to the baseline, which indicates that the local detail context in shallow features plays a significant role in road extraction, especially for narrow roads. Ref_B obtained a 0.77% IoU improvement relative to the baseline, which suggests that the boundary optimization module is better than the threshold segmentation method. Overall, our method achieved a 1.53% improvement in IoU metrics over the baseline.

4.6. Efficiency Comparison

Extensive experiments on three public datasets with different image spatial resolutions proved that the proposed method achieved significant improvements over related works. Efficiency is also important in the automatic interpretation of large-scale remote sensing images. Therefore, we discuss the relations between the IoU accurate with trainable parameters and Floating-point Operations (FLOPs) of related works on the DeepGlobe dataset in this section. The experimental result is shown in Table 6.

Table 6. The accuracy and efficiency of the related researches.

Methods	Parameters (M)	FLOPs (%)	IoU (%)
U-Net	28.94	774.09	46.30
SegNet	18.82	467.52	45.17
Res-U-Net	43.57	1263.01	57.36
DeepLabv3+	59.34	354.97	62.33
D-LinkNet34	21.64	54.70	63.98
D-LinkNet50	217.65	240.36	64.85
Ours	49.18	184.31	66.38

According to the results, our method achieved the highest accuracy of the comparison methods. In particular, the proposed method maintains much lower parameters and FLOPs than the ResNet50 pre-trained methods, such as DeepLabv3+ and D-LinkNet50. Although D-LinkNet34 significantly reduced the number of parameters and FLOPs, which benefited from a lighter pre-trained model than our method, the IoU accuracy also dropped. Overall, our method achieved a better trade-off between accuracy and efficiency than related methods. To intuitively compare the results, we visualize the relationship between the accuracy and trainable parameters and FLOPs in Figure 15.

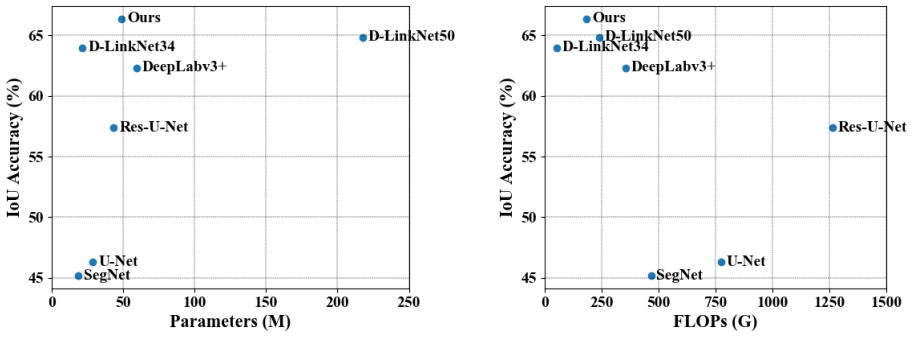


Figure 15. Comparison of accuracy vs. trainable parameters and FLOPs for related methods.

5. Discussion and Conclusions

For problems that exist in road extraction from remote sensing images: (1) The extracted results are discontinuous, and (2) the boundaries of the road are zigzagged and blurred. This work proposed an attention-based cascaded network to optimize road extraction. There are two main parts: The attention-based residual block that is introduced

to maintain the multi-scale spatial details of the roads, and the lightweight optimization network that is designed to refine the road boundaries.

We conducted extensive experiments on three public datasets to evaluate the performance of the proposed method. It outperformed the latest related works and achieved SOTA results. Furthermore, we constructed new benchmarks and provided a detailed description of the data division in the other two datasets, which is convenient for follow-up studies to use for comparison.

These research results suggest that the structure information contained in the shallow convolutional layers plays a vital role in recovering detail, especially for small objects such as buildings and roads. Existing methods learn features from the data end to end, which more or less, ignores prior knowledge such as the structure of the road boundary. We will continue to focus on these structural constraints to address the dependence on data in existing data-driven methods.

Author Contributions: Conceptualization, Xuming Ge and Bo Xu; methodology, Shengfu Li and Cheng Liao; validation, Min Chen, Tianyang Liu, and Di Wu; formal analysis, Yang Jia; investigation, Yang Jia and Yulin Ding; writing—original draft preparation, Shengfu Li and Cheng Liao; writing—review and editing, Han Hu; visualization, Min Chen; supervision, Yulin Ding; project administration, Han Hu; funding acquisition, Shengfu Li. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (nos. 42071355, 41871291), the Sichuan Science and Technology Program (nos. 2020YFG0083 and 2020YJ0010), the Open Innovative Fund of Marine Environment Guarantee (Grand No. HHB002), and the Cultivation Program for the Excellent Doctoral Dissertation of Southwest Jiaotong University (no. 2020YBPY09).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly-available datasets were analyzed in this study. This data can be found here: https://github.com/liaochengcsu/Remote_Sensing_Road_Extraction, accessed on 26 December 2021 .

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

References

1. Wang, W.; Yang, N.; Zhang, Y.; Wang, F.; Cao, T.; Eklund, P. A review of road extraction from remote sensing images. *J. Traffic Transp. Eng. (Engl. Ed.)* **2016**, *3*, 271–282. [CrossRef]
2. Miao, Z.; Shi, W.; Gamba, P.; Li, Z. An Object-Based Method for Road Network Extraction in VHR Satellite Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 4853–4862. [CrossRef]
3. Abdollahi, A.; Pradhan, B.; Shukla, N.; Chakraborty, S.; Alamri, A. Deep Learning Approaches Applied to Remote Sensing Datasets for Road Extraction: A State-of-the-Art Review. *Remote Sens.* **2020**, *12*, 1444. [CrossRef]
4. 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]
5. Chen, L.; Zhu, Q.; Xie, X.; Hu, H.; Zeng, H. Road Extraction from VHR Remote-Sensing Imagery via Object Segmentation Constrained by Gabor Features. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 362. [CrossRef]
6. 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]
7. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Computer Vision—ECCV 2018*; Series Title: Lecture Notes in Computer Science; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; Volume 11211, pp. 833–851. [CrossRef]
8. 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]
9. 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]
10. 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]

11. Ding, L.; Bruzzone, L. DiResNet: Direction-Aware Residual Network for Road Extraction in VHR Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, 1–12. [CrossRef]
12. Wu, Q.; Luo, F.; Wu, P.; Wang, B.; Yang, H.; Wu, Y. Automatic Road Extraction from High-Resolution Remote Sensing Images Using a Method Based on Densely Connected Spatial Feature-Enhanced Pyramid. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 3–17. [CrossRef]
13. Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef]
14. Zhang, Y.; Zhu, J.; Zhu, Q.; Xie, Y.; Li, W.; Fu, L.; Zhang, J.; Tan, J. The construction of personalized virtual landslide disaster environments based on knowledge graphs and deep neural networks. *Int. J. Digit. Earth* **2020**, *13*, 1637–1655. [CrossRef]
15. Buslaev, A.; Seferbekov, S.; Iglovikov, V.; Shvets, A. Fully Convolutional Network for Automatic Road Extraction from Satellite Imagery. 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; IEEE: Salt Lake City, UT, USA, 2018; pp. 197–1973. [CrossRef]
16. 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]
17. Bastani, F.; He, S.; Abbar, S.; Alizadeh, M.; Balakrishnan, H.; Chawla, S.; Madden, S.; DeWitt, D. RoadTracer: Automatic Extraction of Road Networks from Aerial Images. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; IEEE: Salt Lake City, UT, USA, 2018; pp. 4720–4728. [CrossRef]
18. 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]
19. Grinias, I.; Panagiotakis, C.; Tziritas, G. MRF-based segmentation and unsupervised classification for building and road detection in peri-urban areas of high-resolution satellite images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *122*, 145–166. [CrossRef]
20. 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]
21. Zhu, Q.; Chen, L.; Hu, H.; Pirasteh, S.; Li, H.; Xie, X. Unsupervised Feature Learning to Improve Transferability of Landslide Susceptibility Representations. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 3917–3930. [CrossRef]
22. Zhu, Q.; Liao, C.; Hu, H.; Mei, X.; Li, H. MAP-Net: Multiple Attending Path Neural Network for Building Footprint Extraction From Remote Sensed Imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 6169–6181. [CrossRef]
23. Liao, C.; Hu, H.; Li, H.; Ge, X.; Chen, M.; Li, C.; Zhu, Q. Joint Learning of Contour and Structure for Boundary-Preserved Building Extraction. *Remote Sens.* **2021**, *13*, 1049. [CrossRef]
24. 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. Geo-Inf.* **2019**, *8*, 571. [CrossRef]
25. Ding, C.; Weng, L.; Xia, M.; Lin, H. Non-Local Feature Search Network for Building and Road Segmentation of Remote Sensing Image. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 245. [CrossRef]
26. Zhao, X.; Tao, R.; Li, W.; Philips, W.; Liao, W. Fractional Gabor Convolutional Network for Multisource Remote Sensing Data Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*.
27. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Series Title: Lecture Notes in Computer Science; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; Volume 9351, pp. 234–241. [CrossRef]
28. Zhang, Z.; Liu, Q.; Wang, Y. Road Extraction by Deep Residual U-Net. *IEEE Geosci. Remote. Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
29. 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]
30. Chen, Z.; Wang, C.; Li, J.; Xie, N.; Han, Y.; Du, J. Reconstruction Bias U-Net for Road Extraction From Optical Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2284–2294. [CrossRef]
31. 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]
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE: Las Vegas, NV, USA, 2016; pp. 770–778. [CrossRef]
33. 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]
34. 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; IEEE: Honolulu, HI, USA, 2017; pp. 2261–2269. [CrossRef]
35. Dai, J.; Zhu, T.; Wang, Y.; Ma, R.; Fang, X. Road Extraction from High-Resolution Satellite Images Based on Multiple Descriptors. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 227–240. [CrossRef]
36. Wang, S.; Yang, H.; Wu, Q.; Zheng, Z.; Wu, Y.; Li, J. An Improved Method for Road Extraction from High-Resolution Remote-Sensing Images that Enhances Boundary Information. *Sensors* **2020**, *20*, 2064. [CrossRef]

37. Wegner, J.D.; Montoya-Zegarra, J.A.; Schindler, K. A Higher-Order CRF Model for Road Network Extraction. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; IEEE: Portland, OR, USA, 2013; pp. 1698–1705. [CrossRef]
38. 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]
39. Chen, Z.; Fan, W.; Zhong, B.; Li, J.; Du, J.; Wang, C. Corse-to-fine road extraction based on local Dirichlet mixture models and multiscale-high-order deep learning. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 4283–4293. [CrossRef]
40. Zhang, Y.; Xiong, Z.; Zang, Y.; Wang, C.; Li, J.; Li, X. Topology-Aware Road Network Extraction via Multi-Supervised Generative Adversarial Networks. *Remote Sens.* **2019**, *11*, 1017. [CrossRef]
41. 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]
42. Sghaier, M.O.; Lepage, R. Road Extraction From Very High Resolution Remote Sensing Optical Images Based on Texture Analysis and Beamlet Transform. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 1946–1958. [CrossRef]
43. Alshehhi, R.; Marpu, P.R. Hierarchical graph-based segmentation for extracting road networks from high-resolution satellite images. *ISPRS J. Photogramm. Remote Sens.* **2017**, *126*, 245–260. [CrossRef]
44. Batra, A.; Singh, S.; Pang, G.; Basu, S.; Jawahar, C.; Paluri, M. Improved Road Connectivity by Joint Learning of Orientation and Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; IEEE: Long Beach, CA, USA, 2019; pp. 10377–10385. [CrossRef]
45. Tan, Y.Q.; Gao, S.H.; Li, X.Y.; Cheng, M.M.; Ren, B. VecRoad: Point-Based Iterative Graph Exploration for Road Graphs Extraction. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; IEEE: Seattle, WA, USA, 2020; pp. 8907–8915. [CrossRef]
46. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; IEEE: Long Beach, CA, USA, 2019; pp. 3141–3149. [CrossRef]
47. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local Neural Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; IEEE: Salt Lake City, UT, USA, 2018; pp. 7794–7803. [CrossRef]
48. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef]
49. 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]
50. Li, J.; Liu, Y.; Zhang, Y.; Zhang, Y. Cascaded Attention DenseUNet (CADUNet) for Road Extraction from Very-High-Resolution Images. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 329. [CrossRef]
51. Luc, P.; Couprie, C.; Chintala, S.; Verbeek, J. Semantic Segmentation using Adversarial Networks. *arXiv* **2016**, arXiv:1611.08408.
52. Costea, D.; Marcu, A.; Leordeanu, M.; Slusanschi, E. Creating Roadmaps in Aerial Images with Generative Adversarial Networks and Smoothing-Based Optimization. In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 22–29 October 2017; IEEE: Venice, Italy, 2017; pp. 2100–2109. [CrossRef]
53. 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]
54. 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; IEEE: Salt Lake City, UT, USA, 2018; pp. 192–1924. [CrossRef]
55. Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; Raskar, R. DeepGlobe 2018: A Challenge to Parse the Earth through Satellite Images. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; IEEE: Salt Lake City, UT, USA, 2018; pp. 172–17209. [CrossRef]
56. Bahdanau, D.; Cho, K.H.; Bengio, Y. Neural machine translation by jointly learning to align and translate. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
57. Milletari, F.; Navab, N.; Ahmadi, S.A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; IEEE: Stanford, CA, USA, 2016; pp. 565–571. [CrossRef]
58. Mnih, V. Machine Learning for Aerial Image Labeling. Ph.D. Thesis, University of Toronto, Ottawa, ON, Canada, 2013; ISBN 9780494961841.
59. Huawei Cloud Road Extraction Challenge 2020. Available online: https://competition.huaweicloud.com/information/1000041322/introduction (accessed on 26 December 2021).

60. Geng, K.; Sun, X.; Yan, Z.; Diao, W.; Gao, X. Topological Space Knowledge Distillation for Compact Road Extraction in Optical Remote Sensing Images. *Remote Sens.* **2020**, *12*, 3175. [CrossRef]
61. Zhang, Z.; Wang, Y. JointNet: A Common Neural Network for Road and Building Extraction. *Remote Sens.* **2019**, *11*, 696. [CrossRef]