applied sciences

MDPI

*Article*

# Remote Sensing Road Extraction by Road Segmentation Network

**Jiahai Tan [1,2], Ming Gao [1,*], Kai Yang [2,*] and Tao Duan [2]**

[1] School of Optoelectronic Engineering, Xi'an Technological University, Xi'an 710021, China; tanjiahai@xab.ac.cn
[2] State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China; duantao@opt.ac.cn
* Correspondence: minggao1964@163.com (M.G.); yangkai20@mails.ucas.ac.cn (K.Y.)

**Abstract:** Road extraction from remote sensing images has attracted much attention in geospatial applications. However, the existing methods do not accurately identify the connectivity of the road. The identification of the road pixels may be interfered with by the abundant ground such as buildings, trees, and shadows. The objective of this paper is to enhance context and strip features of the road by designing UNet-like architecture. The overall method first enhances the context characteristics in the segmentation step and then maintains the stripe characteristics in a refinement step. The segmentation step exploits an attention mechanism to enhance the context information between the adjacent layers. To obtain the strip features of the road, the refinement step introduces the strip pooling in a refinement network to restore the long distance dependent information of the road. Extensive comparative experiments demonstrate that the proposed method outperforms other methods, achieving an overall accuracy of 98.25% on the DeepGlobe dataset, and 97.68% on the Massachusetts dataset.

**Keywords:** convolutional neural networks; semantic segmentation; self-attention mechanism



**Citation:** Tan, J.; Gao, M.; Yang, K.; Duan, T. Remote Sensing Road Extraction by Road Segmentation Network. *Appl. Sci.* **2021**, *11*, 5050. https://doi.org/10.3390/app11115050

Academic Editors: Jean Sequeira, Xingfa Gu and Sébastien Mavromatis

Received: 6 April 2021
Accepted: 25 May 2021
Published: 29 May 2021



## 1. Introduction

Remote sensing road extraction aims to identify the road pixels in the images and complete the binary segmentation of the road. Nowadays, road extraction is needed in many applications [1,2] such as traffic navigation and urban planning. However, it is difficult to extract the road from the remote sensing images due to increased noise information [3]. The noise information mainly comes from occlusion or shadows of other ground objects, and similar categories. Moreover, the roads are not always regular, and the width and curvature of the roads varies greatly in different scenes. These factors reduce the accuracy of road extraction methods and prevent the restoration of complete road connectivity.

Recently, numerous methods have been proposed to obtain a road map. According to whether the Convolutional Neural Network (CNN) architecture is used or not, road extraction methods can be divided into two categories: hand-crafted feature methods and CNN methods.

The hand-crafted feature methods divide the pixels into road and non-road by extracting shallow features and using empirical hypothesis. Early road extraction methods used hand-crafted features (geometry, texture, spectrum, etc.), multiple algorithms (edge detection, tracking, area clustering, etc.), and were combined with empirical assumptions to extract the road in the image. M. Barzohar et al. [4] used the width, length, curvature, and pixel intensity of the road as empirical hypothesis information, and the geometric probability model with the maximum posterior probability estimation was established for road extraction. J. Hu et al. [5] first detected the local uniform area around the pixel to generate the road tree structure, and then a Bayesian decision model was used to obtain the

final road. M. Song et al. [6] used support vector machines as a classifier and considered weight combination of spectral information and shape information to identify road pixels. However, the hand-crafted feature methods depend on the quality of feature selection. These hand-crafted feature methods are rough and depend heavily on prior knowledge.

With the rapid development of deep learning [7], CNN methods have become the mainstream methods due to the representation power [8]. Li et al. [9] used CNN to classify pixels, and then a linear integral convolution algorithm was used to enhance the road connection structure. G. Mattyus et al. [10] used the variant of ResNet [11] for road segmentation and then designed a road inference algorithm to correct the segmentation results. Xu et al. [12] utilized local and global attention mechanisms to enhance road information in DenseNet. However, CNN methods do not fully excavate the information of context features extracted from the CNN [13], and they do not consider enough about the strip shape features and long distance dependence of roads.

The road pixels in remote sensing images may be interfered with by the abundant ground such as buildings, trees, and shadows, which may introduce poor road connectivity. It is important to enhance the context information related to the road and suppress the interference information of non-road. In this paper, a road segmentation method is proposed for road extraction with high-resolution remote sensing images. The road information is extracted by enhancing the context characteristics in the segmentation step and maintaining the stripe characteristics in a refinement step. The contributions are as follows:

(1) To enhance road context features, an end-to-end road segmentation network is designed, since road connectivity is easily disturbed by noise.
(2) To strengthen the context information belonging to the road, an inter-layer self-attention mechanism is introduced to generate weight maps.
(3) To hold the strip information of the road, a refine network with the striping pooling is introduced to refine the results of the segmentation network.

## 2. Materials and Methods

### 2.1. Materials

Two large public datasets were used for experiment: the DeepGlobe dataset and the Massachusetts dataset. A brief introduction is as follows.

The DeepGlobe dataset is a satellite dataset that uses images from Thailand, Indonesia and India. The land area of the dataset is about 2220 $km^2$ and contains a variety of image scenes (urban, rural, wilderness, tropical rainforest, seaside, etc.). The DeepGlobe dataset consists of 6226 images with the size of 1024 $\times$ 1024. The spatial resolution of DeepGlobe is 50 cm/pixel, which is used for the road extraction challenge. In the experiment, 4358 images are used for training, and the remaining 1868 images are used for testing.

The Massachusetts dataset is an aerial dataset that uses images from Massachusetts, US. The dataset has a resolution of 1.2 m/pixel, covers a land area of 2.25 $km^2$, and contains urban, suburban, and rural environments. The size of each image is 1500 $\times$ 1500. The Massachusetts dataset consists of 1108 training images, and 49 testing images.

The experiments were performed on PyTorch library with GPU. Within the limits of computational cost, the size of the input image was cropped to 512 $\times$ 512 pixels in both the DeepGlobe dataset and the Massachusetts dataset. To alleviate the over-fitting problem, data augmentations were used, such as image flipping, image shifting and scaling, and color jittering.

The backbone network of the proposed method was ResNet34, due to its powerful encoder capability. The Adam optimizer was used to train the proposed network. The training epoch was set to 50. The batch size was 24 for the Massachusetts Dataset and 16 for the DeepGlobe dataset. The learning rate was initially set to $2 \times 10^{-4}$ and reduced by a factor of 0.1. During the testing phase, each input image was predicted a total of eight times by flipping different angles, and the final result was the average output of these outputs.

To verify the performance of the proposed method, four widely used metrics were employed: overall accuracy ($OA$), precision ($P$), recall ($R$), and $F1$-score. After the calcula-

tion of true positive (*TP*), true negative (*TN*), false positive (*FP*), and false negative (*NT*), the measures of *OA*, *P*, *R*, and *F*1-score were calculated as follows:

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN}, F1 = 2 \times \frac{P \times R}{P + R}.$$

$$OA = \frac{TP + TN}{TP + FP + TN + FN}$$

The precision and recall were hoped to remain at a high level, indicating improved performance, although the two values are contradictory in some cases and have a negative correlation. *F* score is used to consider the precision and the recall simultaneously. It can be regarded as a weighted average of precision and recall, where its maximum value is 1 and minimum value is 0.

### *2.2. The Proposed Method*

The proposed method is introduced in this section. First, the network architecture is illustrated. Then, a self-attention mechanism is exploited to enhance the inter-layer context features. Finally, a refinement network is designed to refine the road map.

#### 2.2.1. Network Architecture

An overview of the network architecture is shown in Figure 1. The network adopts a UNet-like architecture to extract the road map from a high-resolution remote sensing image. The U-Net architecture concatenates multiple features from high layers to low layers, which is widely used in remote sensing images [14]. The proposed method extracts features by convolution layers and further produces a dense pixel-wise output by deconvolution layers. However, multiple features may contain redundant noise information from non-road pixels. To enhance the context information in adjacent layers, an attention mechanism exploits the hierarchical features to generate attention maps $w_1$, $w_2$, $w_3$. These weight maps are involved in skip connections and adjacent layers to enhance the context information. Then, the low-level features are weighted and fused into the corresponding layers in the deconvolution layers to emphasize the context features of the road. After multiple deconvolutions and sigmoid layers, the features are restored to the size of input image. To enhance the strip information of a road, a network with strip pooling is designed to refine segmentation results.

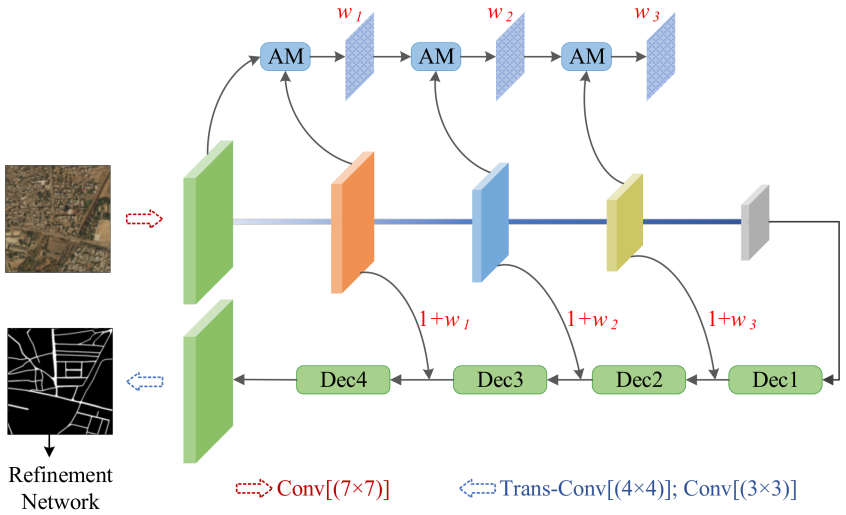


**Figure 1.** The architecture of the proposed network. A self-attention mechanism is exploited to enhance the inter-layer features. The final road maps are obtained through a refinement network.

#### 2.2.2. Segmentation Step

The spatial context information in different convolutions is complementary: the global context in deep layers and the local context in shallow layers. Skip connections from shallow to high layer can introduce location information into semantic information. However, most of the information in images will cause disturbance to the road. Therefore, a self-attention mechanism is proposed to emphasize the target information in adjacent layers.

The attention mechanism (AM) is shown in Figure 2. The feature of the low layer is denoted as $F_i$, while the feature of the high layer is up-sampled with bilinear interpolation to the same size of $F_i$, and then denoted as $F_j$. The concatenation of $F_i$ and $F_j$ is sent to a series of convolution layers which contains $1 \times 1$ and $3 \times 3$. After the sigmoid layer, the AM obtains the weighted map, $w$, which has two effects. The first is used to select the useful information and refine the concatenated feature with residual learning. The refined feature is the '$F_i$' for the next adjacent layer of network in the AM. The second introduces the skip connections to enhance the detailed information of road. Thus, there are two outputs of the AM: weighted map and refined feature. Mathematically, the AM computes as

$$\hat{F} = R\left(Cat\left(F_i, F_j\right)\right)\left[1 + \frac{1}{1 + e^{-H(Cat(F_i, F_j))}}\right]$$

where $\hat{F}$ is the refined feature by AM, and the weight map $w = 1/\left(1 + e^{-H(cat(F_i, F_j))}\right)$. $Cat(\cdot)$ denotes the concatenation on adjacent layer features $F_i, F_j$, $R(\cdot)$ is three consecutive $3 \times 3$ convolution operations, and $H(\cdot)$ means the operation of alternating $3 \times 3$ convolution and $1 \times 1$ convolution.

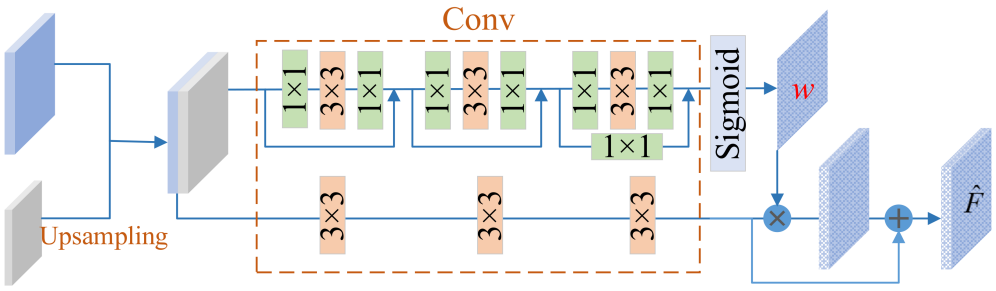


**Figure 2.** Attention mechanism (AM). The features of adjacent layers are first up-sampled to the same size, and are then passed through a series of convolution operations to obtain the attention weight map, $w$, and the refined features, $\hat{F}$.

#### 2.2.3. Refinement Step

The outputs from the segmentation network are coarse and rough, which will lead to error in road segments caused by the interference information. The tensor voting algorithm and the conditional random field (CRF) are widely used as post-processing to refine the results from segmentation network. However, the former is unstable due to the parameter-dependence, while the latter model is overly complex. A designed refinement network is exploited to refine the results of the segmentation network. The difference is that long-distance dependence of the road is also considered here. The refined network produces more stable road maps and is embedded in the overall network for end-to-end training.

The designed refinement network is shown in Figure 3. The input of DRN is the coarse map, which has the width and height of W and H, respectively, and one channel. The output of DRN is the refined map with the same size. There are three corresponding layers in DRN. In the encoder, the number of channels is changed to 1, 16, 32, and 64 in turn, and the sizes of the features are sub-sampled as 1/2, 1/4, and 1/8. Due to the characteristic of long-distance dependence on the road, maintaining connectivity in the extracted results is challenging. To capture long-distance dependence between different locations in features, the strip pooling [15] that averages all values in a row or column feature is embedded

at the high layer features of 64 channels. It allows the distribution of roads in scattered areas to be connected on the feature map. To capture local context information, two spatial pooling layers are used to collect short distance dependencies. The output features of two pooling layers are concatenated together, and then $1 \times 1$ convolution operations are used to change the channel numbers. Finally, the decoder recovers the characteristic resolution and obtains a refined map of the road.

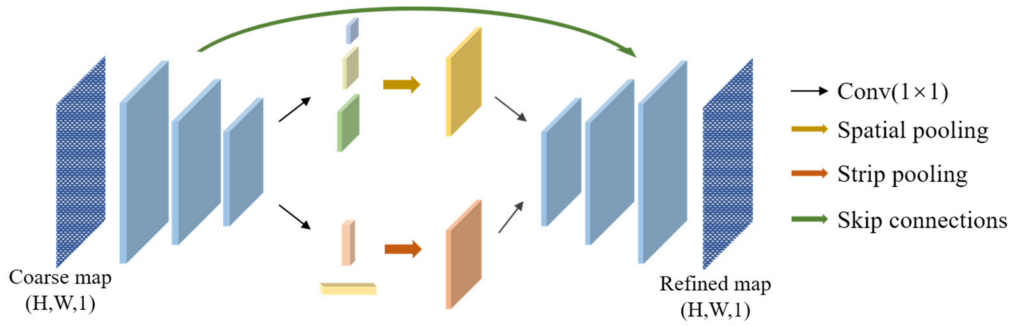


**Figure 3.** Designed refined network (DRN).

Road extraction is a pixel-level recognition and judgment; however, there is a great imbalance between road pixels and non-road pixels in the image. Thus, the constraint on road pixels in the loss function is of great importance. As a binary segmentation task, road extraction most commonly uses the binary cross entropy (bce) loss function. Inspired by [16], the bce is given a dynamic weight probability, which is set by the frequency of the road pixel. Suppose the batch size is represented as $B$, and the size of the image is $H \times W$, then the weight $\alpha$ is calculated as follows:

$$\alpha = \frac{sum(y_n)}{BHW},$$

where $y_n$ is the ground truth of the i-th batch, and $sum(\cdot)$ represents the statistics of the number for road pixels. Thus, the bce loss is modified to be

$$L_{bce} = -\frac{1}{B}[(1-\alpha)y_n \log a_n + \alpha(1-y_n)\log(1-a_n)],$$

$$a_n = \frac{1}{1+\exp(-f_w(x_n))}$$

where $a_n$ is the sigmoid value in the end of model, $f_w(x_n)$ is the output of the model for $x_n$, and $y_n$ is the ground truth for $x_n$, which are binary maps. In order to ensure the loss function still has the ability to constrain the road after multiple iterations, the dice loss is used to measure the similarity of prediction and ground truth:

$$L_{dice} = \frac{\sum_{n=1}^{B} y_n f_w(x_n)}{\sum_{n=1}^{B}(y_n + f_w(x_n))}.$$

The final loss function used in this article is a combination of $L_{bce}$ and $L_{dice}$ to alleviate the performance problems caused by sample imbalance in road extraction.

## 3. Experiments

In the experiments, both the DeepGlobe dataset [17] and Massachusetts dataset [18] are used and five compared methods were considered: the FCN [19], the UNet [20], the CasNet [21], the ResUNet [22], and the D-LinkNet [23]. Among them, FCN has the sequential operations of down-sampling in early layers. CasNet uses the VGG-Net as an encoder network, while ResUNet, D-LinkNet, and the proposed method use the ResNet34

as an encoder network. All the compared methods use cross-entropy as the training loss and employ the same data processing for fairness.

Table 1 shows the quantitative results on the DeepGlobe dataset. Numerically, the proposed method achieved the best performance on three metrics of OA, P, and F1. D-LinkNet uses dilated convolution to increase the receptive fields on the feature, and it obtained the highest performance on recall. ResUNet had an advantage in P, and was only 0.23% lower than our proposed method. The results of FCN and UNet were worse than that of other methods, which was mainly caused by the loss of road context information after multiple down-sampling operations on different features.

**Table 1.** Results of the comparative experiments on the DeepGlobe dataset.

| Method | OA (%) | P (%) | R (%) | F1 (%) |
|---|---|---|---|---|
| FCN | 97.42 | 77.98 | 51.16 | 61.78 |
| UNet | 97.68 | 74.68 | 60.44 | 66.81 |
| CasNet | 97.66 | 75.72 | 63.41 | 69.02 |
| ResUNet | 98.02 | 79.76 | 62.76 | 70.24 |
| D-LinkNet | 98.10 | 76.57 | **71.46** | 73.93 |
| Our | **98.25** | **79.99** | 70.00 | **74.66** |

Table 2 reports the comparative results on the Massachusetts dataset. Compared with D-LinkNet, the proposed method increased 0.14%, 1.24%, and 0.41% on OA, P, and F1, respectively. The original encoder structure of UNet leads to poor performance in the four metrics. Regarding the improvement of the encoder structure, the ResUNet combines residual learning in its encoder, the CasNet uses VGG-Net, and D-LinkNet uses ResNet as its encoder. They had advantages of 2.16%, 1.15%, and 3.28% compared to UNet on P, respectively.

**Table 2.** Results of the comparative experiments on the Massachusetts dataset.

| Method | OA (%) | P (%) | R (%) | F1 (%) |
|---|---|---|---|---|
| FCN | 97.35 | 74.83 | 53.36 | 62.29 |
| UNet | 97.55 | 75.06 | 55.47 | 63.79 |
| CasNet | 97.49 | 77.22 | 51.60 | 61.86 |
| ResUNet | 97.43 | 76.21 | 54.62 | 63.63 |
| D-LinkNet | 97.54 | 78.34 | **56.69** | 65.78 |
| Our | **97.68** | **79.58** | 56.66 | **66.19** |

To observe the performance of different designs in the proposed method, Table 3 shows the results of the ablation experiment on the Massachusetts dataset. The Base method only uses ResNet as an encoder and multiple deconvolution layers as a decoder. AM is the attention mechanism to enhance the context feature from the inter-layer, and DRN is the network to refine road maps. The last line represents the proposed method, and the reciprocal of road frequency was used as weight in the loss function. Compared to the base method, the AM contributed 3.26% performance improvement on P, and the DRN showed an improvement of 0.6% on P compared to the Base combined with AM method. Our proposed method further promoted 0.21%, while the R score dropped by 0.01%.

**Table 3.** Results of the ablation study on the Massachusetts dataset.

| Method | OA (%) | P (%) | R (%) | F1 (%) |
|---|---|---|---|---|
| Base | 97.43 | 75.51 | 54.64 | 63.40 |
| Base + AM | 97.57 | 78.77 | 55.52 | 65.13 |
| Base + AM + DRN | 97.62 | 79.37 | **56.67** | 66.12 |
| Our | **97.68** | **79.58** | 56.66 | **66.19** |

Additional examples of road extraction are shown in Figure 4. It is obvious from the subgraph (c,e) that road connectivity could not be restored by FCN [19], UNet [20], and ResUNet [22]. The results of CasNet [21] and the DinkNet [23] missed segments on subgraph (f) compared with our proposed method. In the subgraph (a,b), the edge pixels of the trunk road were extracted completely by the proposed method. These examples demonstrate the effectiveness of our proposed method on road extraction.

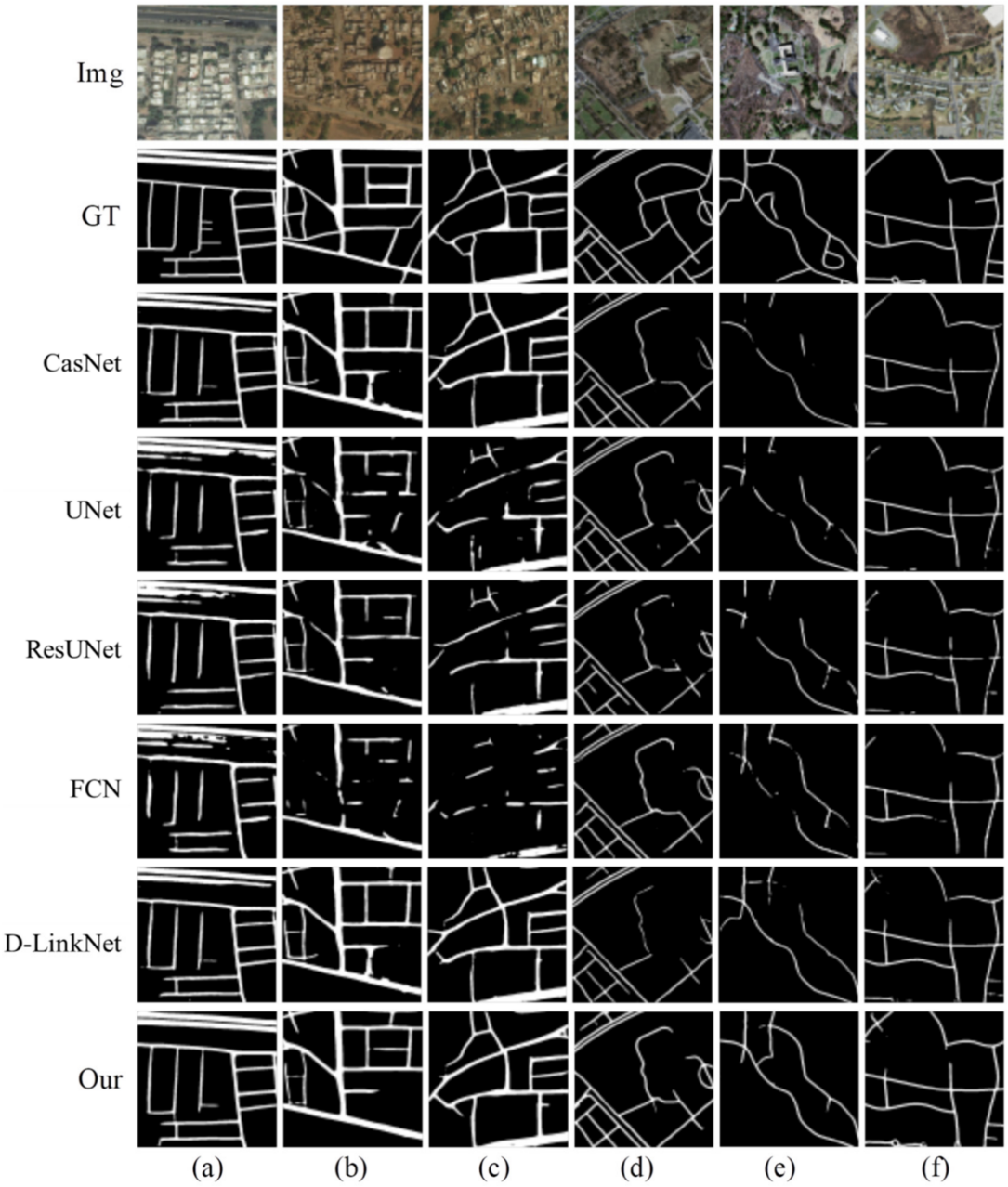


**Figure 4.** Examples of road extraction obtained by the compared methods. (**a**–**c**) Selected from the DeepGlobe dataset, and (**d**–**f**) showed on the Massachusetts dataset.

## 4. Conclusions and Future Work

In this paper, an end-to-end road segmentation network is proposed to extract the road from high-resolution remote sensing images. Although the existing CNN architectures have made achievements, they have not explored enough the context features with adjacent layers of CNN. In this paper, an inter-layer self-attention mechanism was designed to

obtain road context information. The roads in images always have the shape of a thin stripe. The proposed method introduces the strip pooling in the refinement network to better restore the topology and long-distance dependence of roads. The whole network achieved better performance by enhancing the context feature in adjacent layer and strip feature of roads in refined processing. The quantitative results demonstrate the effectiveness of the proposed method and the superiority of road extraction from high-resolution remote sensing images. The improvements of our proposed method obtained on two datasets can be mainly attributed to two designs. The first factor is that the context feature enhanced by the attention mechanism used in adjacent layers guarantees the correct identification of road pixels. The second factor is that the consideration of the strip characteristics of the road in the refinement network makes the extracted roads retain a better topology.

Although the proposed method is successfully applied to remote sensing road extraction, there are still some possible extensions in the future work. This paper only exploits remote sensing images to extract the road, which is difficult with complex scenarios [22] (dense urban area, different viewpoints). Future work will tackle the complex scenarios by integrating multiple data sources, such as multi-temporal images [24], and multi-modal data [25].

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

**Funding:** This research was funded by the Open Research Fund of State Key Laboratory of Transient Optics and Photonics, Chinese Academy of Sciences under Grant SKLST202005.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** The authors would like to thank the editors and reviewers for their insightful suggestions.

**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.* **2016**, *3*, 271–282. [CrossRef]
2. Cao, Y.; Wang, Z.; Yang, L. Advances in method on road extraction from high resolution remote sensing images. *Remote Sens. Technol. Appl.* **2017**, *32*, 20–26.
3. Zheng, X.; Yuan, Y.; Lu, X. Hyperspectral Image Denoising by Fusing the Selected Related Bands. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 2596–2609. [CrossRef]
4. Barzohar, M.; Cooper, D. Automatic finding of main roads in aerial images by using geometric-stochastic models and estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **1996**, *18*, 707–721. [CrossRef]
5. Hu, J.; Razdan, A.; Femiani, J.C.; Cui, M.; Wonka, P. Road Network Extraction and Intersection Detection from Aerial Images by Tracking Road Footprints. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 4144–4157. [CrossRef]
6. Song, M.; Civco, M. Road extraction using SVM and image segmentation. *Photogramm. Eng. Remote Sens.* **2004**, *70*, 1365–1371. [CrossRef]
7. Zheng, X.; Zhang, Y.; Lu, X. Deep balanced discrete hashing for image retrieval. *Neurocomputing* **2020**, *403*, 224–236. [CrossRef]
8. 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]
9. Li, P.; Zang, Y.; Wang, C.; Li, J.; Cheng, M.; Luo, L.; Yu, Y. Road network extraction via deep learning and line integral convolution. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 1599–1602. [CrossRef]
10. Mattyus, G.; Luo, W.; Urtasun, R. DeepRoadMapper: Extracting Road Topology from Aerial Images. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 3458–3466. [CrossRef]

11. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
12. Xu, Y.; Xie, Z.; Feng, Y.; Chen, Z. Road Extraction from High-Resolution Remote Sensing Imagery Using Deep Learning. *Remote Sens.* **2018**, *10*, 1461. [CrossRef]
13. Zheng, X.; Chen, X.; Lu, X. A Joint Relationship Aware Neural Network for Single-Image 3D Human Pose Estimation. *IEEE Trans. Image Process.* **2020**, *29*, 4747–4758. [CrossRef] [PubMed]
14. Zheng, X.; Yuan, Y.; Lu, X. A Deep Scene Representation for Aerial Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 4799–4809. [CrossRef]
15. Hou, Q.; Zhang, L.; Cheng, M.-M.; Feng, J. Strip Pooling: Rethinking Spatial Pooling for Scene Parsing. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 4003–4012.
16. Mostajabi, M.; Yadollahpour, P.; Shakhnarovich, G. Feedforward semantic segmentation with zoom-out features. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3376–3385.
17. Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; Raskar, R. DeepGlobe 2018: A Challenge to Parse the Earth through Satellite Images. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 172–181.
18. Mnih, V. Machine Learning for Aerial Image Labeling. Ph.D. Thesis, University of Toronto, Toronto, ON, Canada, 2013.
19. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
20. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
21. 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]
22. Zhang, Z.; Liu, Q.; Wang, Y. Road Extraction by Deep Residual U-Net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
23. Zhou, L.; Zhang, C.; Wu, M. D-LinkNet: LinkNet with Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 182–186. [CrossRef]
24. Zheng, X.; Chen, X.; Lu, X.; Sun, B. Unsupervised Change Detection by Cross-Resolution Difference Learning. *IEEE Trans. Geosci. Remote Sens.* **2021**.
25. Zheng, X.; Wang, B.; Du, X.; Lu, X. Mutual Attention Inception Network for Remote Sensing Visual Question Answering. *IEEE Trans. Geosci. Remote Sens.* **2021**.