remote sensing

MDPI

Technical Note

Road Extraction from Remote Sensing Images Using the Inner Convolution Integrated Encoder-Decoder Network and Directional Conditional Random Fields

Shuyang Wang [1], Xiaodong Mu [1], Dongfang Yang [2,*], Hao He [3] and Peng Zhao [1]

[1] Department of Information Engineering, Xi'an Hi-Tech Research Institute, Xi'an 710025, China; yelvlanshu@163.com (S.W.); wascom4@sina.com (X.M.); zpxhh@163.com (P.Z.)
[2] Department of Control Engineering, Xi'an Hi-Tech Research Institute, Xi'an 710025, China
[3] Department of Measurement and Control, Qingzhou Hi-Tech Research Institute, Qingzhou 262500, China; aruhan@peihua.edu.cn
* Correspondence: yangdf@xjtu.edu.cn; Tel.: +86-188-5753-3199

Abstract: Road extraction from remote sensing images is of great significance to urban planning, navigation, disaster assessment, and other applications. Although deep neural networks have shown a strong ability in road extraction, it remains a challenging task due to complex circumstances and factors such as occlusion. To improve the accuracy and connectivity of road extraction, we propose an inner convolution integrated encoder-decoder network with the post-processing of directional conditional random fields. Firstly, we design an inner convolutional network which can propagate information slice-by-slice within feature maps, thus enhancing the learning of road topology and linear features. Additionally, we present the directional conditional random fields to improve the quality of the extracted road by adding the direction of roads to the energy function of the conditional random fields. The experimental results on the Massachusetts road dataset show that the proposed approach achieves high-quality segmentation results, with the F1-score of 84.6%, which outperforms other comparable "state-of-the-art" approaches. The visualization results prove that the proposed approach is able to effectively extract roads from remote sensing images and can solve the road connectivity problem produced by occlusions to some extent.

Keywords: remote sensing; semantic segmentation; encoder-decoder network; inner convolution; directional conditional random fields

Citation: 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. https://doi.org/10.3390/rs13030465

Received: 12 December 2020
Accepted: 22 January 2021
Published: 28 January 2021

1. Introduction

Road extraction from remote sensing images is of great significance for updating geographic information systems (GIS), urban planning, navigation, disaster assessment, etc. [1]. In the past, the most widely used way to extract roads was through manual vision interpretation, which takes a lot of time and has a high labor cost, and the extracted results may vary due to the differences of interpreters. Automatic road extraction technology can improve the efficiency of road extraction, so it has become a hot issue in this field.

Over the past few decades, many studies on automatic road extraction have been conducted. The traditional road extraction approaches are usually based on traditional computer vision methods, such as prior knowledge [2,3], the mathematical morphology [4,5], the active contour [6,7], the Markov random field (MRF) [8,9], the support vector machine (SVM) [10,11], and so on. These methods can work well for some simple cases, but their performance depends on many threshold parameters that should be elaborately given. The threshold parameters usually vary in different images, so the traditional methods can only work in a small range of data, and cannot be validated in complex circumstances.

Benefiting from the rapid growth of available data and the computing power, deep learning technology represented by convolutional neural networks (CNNs) has achieved a breakthrough in the field of computer vision [12–15]. As road extraction can be considered

a binary segmentation problem, researchers have preferred to adopt CNN-based methods to extract roads from remote sensing images in recent years. Early in 2013, Mnih et al. first introduced a CNN to segment a road from aerial images and established a corresponding dataset, known as the Massachusetts dataset [16,17]. Wang et al. [18] proposed a patch-based CNN to recognize patterns of the road, and then tracked the road by a finite state machine (FSM). Moreover, Alshehhi et al. [19] used a patch-based CNN to extract roads and buildings simultaneously. Additionally, Rezaee et al. [20] designed a patch-based deep neural network to extract roads from images with a 0.15 m spatial resolution. However, these patch-based approaches adopt the sliding window strategy, so their accuracy and speed are limited. With the emergence of a large number of excellent semantic segmentation network frameworks based on fully convolutional network (FCN) [21] or encoder-decoder architecture, such as U-Net [22], SegNet [23], and DeepLab [24,25], the road extraction task has achieved significant progress [26].

Wei et al. [27] proposed a road structure refined CNN for road extraction. It adopted the architecture of FCN and designed fusion layers to obtain a structured output of road extraction. Furthermore, Zhong et al. [28] designed an FCN which combines the output of the shallow fine-grained pooling layer with the deep final-score layer. Cheng et al. [29] presented a cascaded encoder-decoder network to extract roads and centerlines simultaneously. Additionally. Zhang and Wang [30] proposed a network with atrous convolution to produce a large receptive field, so it worked well in both road and building extraction tasks. Moreover, Zhang et al. [31] developed a deep residual U-Net for road extraction, which combined the U-Net architecture and residual units. Xin et al. [32] applied a dense U-net for road extraction. A dense U-net consists of dense connection units and skips connections, which strengthens the fusion of different scales by connections at various network layers. Li et al. [33] proposed an improved D-Linknet for detecting roads from unmanned aerial vehicle images. In addition to the improvement of the network structure, some studies have improved the loss function to obtain better road extraction results. Mosinska et al. introduced [34] pixel-wise loss to capture the higher-order topological features of linear structures. Additionally, Abdollahi et al. [35] developed a VNet model with a new dual loss function called cross-entropy-dice-loss, which can decrease the influence of class imbalance and improve the road extraction result. Moreover, He et al. [36] employed the structural similarity as a loss function to improve the quality of road extraction.

Adding the post-processing method can also help improve the performance of road extraction. Sun et al. [37] proposed the stacked U-net with a hybrid loss function, and improved the recall by post-processing methods, including road map vectorization and a shortest path search. Conditional random fields (CRF) is a widely used post-processing model. Chen et al. introduced the fully connected CRF to optimize the segmentation results [38]. Panboonyuen et al. [39] proposed a SegNet-based deep convolutional neural network (DCNN) to segment roads, and CRF was used as a post-processing step to reduce falsely classified roads. By modeling the nearby pixels with energy terms, CRF makes the spatially proximal and similar-colored pixels more likely to be in the same class. However, traditional CRF only takes the position and color as energy terms, which is often used to optimize the segmentation boundary and remove outliers. Therefore, traditional CRF is useless for solving the problem of incomplete road extraction due to occlusions.

These methods can generally segment roads in remote sensing images well; however, they have difficulties in predicting roads covered by trees, buildings, or other non-road objects. Due to the unusual features of covered roads, a normal CNN method will not be able to represent them correctly.

To address this problem, we designed an inner convolutional network and a directional CRF for road extraction to segment roads from remote sensing images more accurately. The main contributions of this study are as follows:

1. We designed a novel inner convolutional network (ICN) integrated encoder-decoder network for road extraction. ICN splits the feature map into slices along a row or column and views these slices of feature maps as layers of traditional CNN, and

then applies convolution, activation, etc. to these slice maps sequentially. Therefore, the spatial information can be transmitted in the same layer, which is helpful for enhancing the ability of CNN to extract a road covered by other objects;

2. We proposed the directional conditional random fields (DCRF) as a post-processing method to further improve the quality of road extraction. The DCRF adds the direction of the road as an energy term of CRF, which will favor the assignment of the same label to pixels with similar directions, so it can help to connect roads and remove noise;
3. Ablation studies on the Massachusetts dataset verify the effectiveness of the proposed ICN and DCRF. Experimental results show that the proposed method can improve the accuracy of the extracted road and solve the road connectivity problem produced by occlusions to some extent.

The remainder of this study is organized as follows. In Section 2, we describe the details of our proposed methodology. In Section 3, experimental results and the corresponding analyses are provided. Finally, Section 4 concludes the study.

2. Materials and Methods

In this section, we present an inner convolution integrated encoder-decoder network with the post-processing of directional CRF for segmenting roads from remote sensing images. The proposed model is divided into three parts: An encoder-decoder network as the backbone; an inner convolutional network to enable contextual information to be transmitted between pixels across rows and columns in a layer; and the directional CRF approach to optimize segmentation results.

2.1. Overview of the Proposed Method

An overview of the proposed method is shown in Figure 1. We applied the inner convolutional network to the output of the backbone network's encoder. The preliminary segmentation result is optimized by directional CRF.

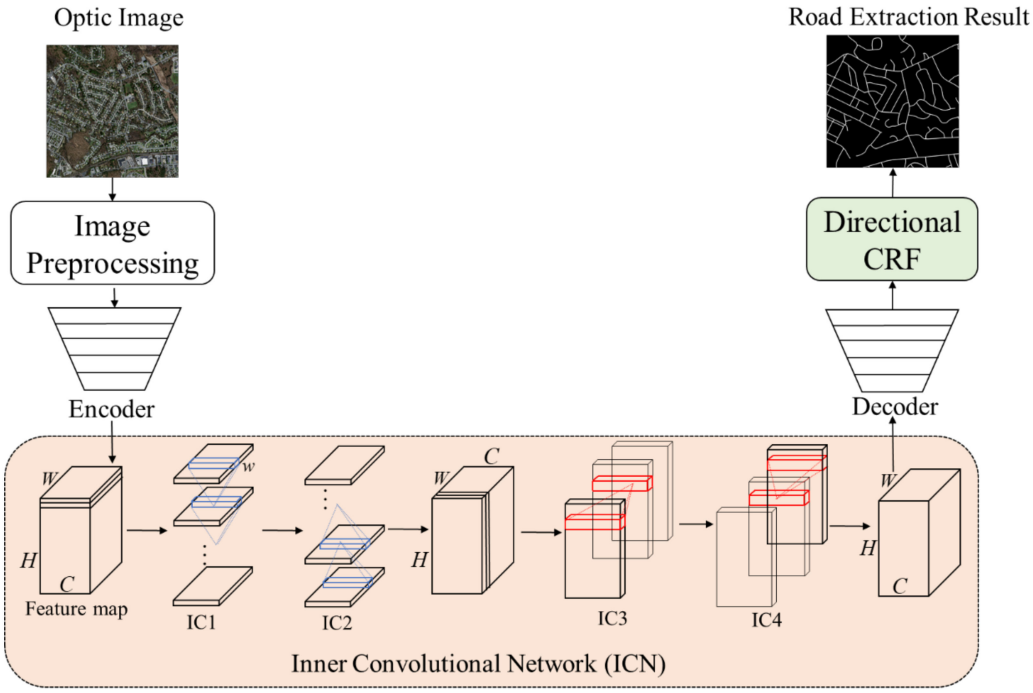

Figure 1. An overview of the proposed method.

In this study, we selected U-net [22] as the backbone for road extraction from remote sensing images. U-net was proposed in 2015, and was first proposed for the segmentation of blood vessels in medical imaging, but it has also been proven to perform excellently in

remote sensing imagery segmentation. Figure 2 briefly introduces the network architecture and parameters of the U-net used in this study.

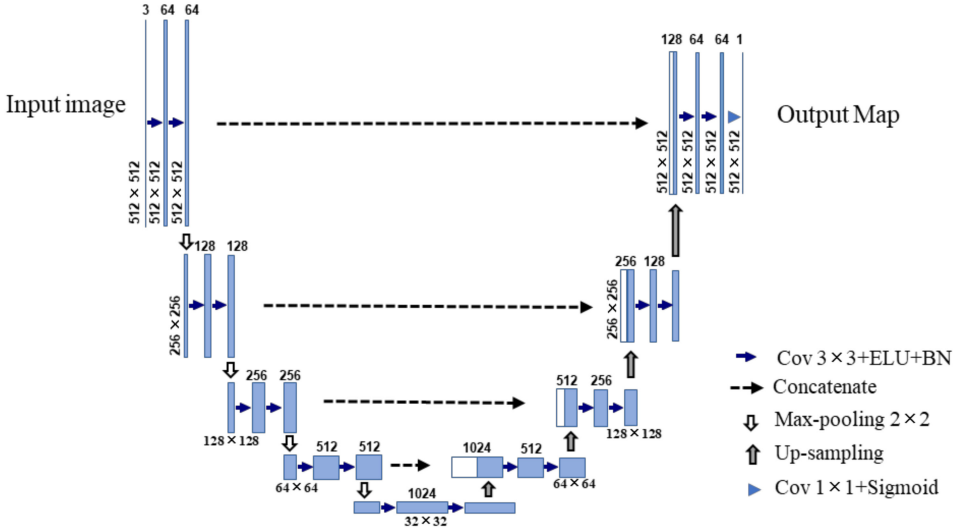

Figure 2. The architecture of the modified U-net.

The U-net consists of an encoder part and a decoder part. The encoder part is a typical convolutional network. It consists of repeated applications of two 3×3 convolutional layers with stride 1 and padding 1. Each convolution follows an exponential linear unit (ELU) activation function and a batch normalization (BN) layer. Moreover, each convolutional block is followed by 2×2 max-pooling with stride 2 for down-sampling. The number of feature channels is doubled after each block. The encoder part carries out four convolutional blocks. The decoder part corresponds to the encoder part. Every block in the decoder part consists of a 2×2 deconvolution, a concatenation with the corresponding feature map from the encoder part, and two 3×3 convolutions followed by ELU activation and a BN layer. The number of feature channels is halved after each up-sampling process. At the final layer, a 1×1 convolution with a Sigmoid function is used to generate the desired prediction.

Compared with the original U-net network, the modified U-net network in Figure 2 replaces the Rectified Linear Unit (ReLU) with ELU [40] and adds the BN layer [41] after each convolutional layer. The ELU activation function is defined in Equation (1).

$$f(x) = \begin{cases} x, & x > 0 \\ \alpha(e^x - 1), & x \leq 0 \end{cases} \quad (1)$$

The ELU and batch normalization have become commonly used components of the CNN network, as their superiority in the training of CNN has been widely recognized. It should be noted that the number of convolution kernels of the output layer of the U-net is set to 1, and the classifier is also transformed from Softmax to Sigmoid, as road extraction is a binary segmentation task. The Sigmoid function is shown in Equation (2).

$$\text{Sigmoid}(x) = \frac{1}{1 + e^{-x}} \quad (2)$$

2.2. Inner Convolutional Network

The greatest difficulty faced in a road extraction task is the extraction of the occluded road. A road covered by trees, buildings, etc., will lead to visual invisibility, so one can only extract the occluded road by analyzing and inferring the surrounding pixels of the occluded road. To address this issue, we propose the inner convolutional network, which can make better use of the road-specific contextual information. Its basic structure is shown in the orange part of Figure 1.

The inner convolutional network is applied to the output feature map of the encoder part in the encoder-decoder network. The output feature map is a tensor of the size C × H × W, where C, H, and W denote the number of channels, rows, and columns, respectively. The feature map contains rich high-level semantic information after feature extraction by the encoder part. Firstly, the feature map is split into H slices along rows, and the obtained H slices are then sent to the first unit of the inner convolutional network, IC1. In this unit, the first slice is sent to a convolution layer with C kernels of size C × w. The output of the convolution layer is added to the second slice to generate a new slice, and the new slice is then sent to the next convolution layer. This process continues until the last slice is processed. In IC1, the context information is continuously transmitted from top to bottom, and the output feature map is then sent to IC2. IC2 is similar to IC1, but the direction of convolution is upward. First, the last slice is sent to the convolution layer, and the output is added with the previous slice to generate a new slice, applying this process until the first slice is updated. After that, the new slices are concated in the row dimension to recombine a complete feature map with the size of C × H × W. In the same way, the recombined feature map is re-split into W slices along rows along the column dimension, and these slices are sent to IC3 and IC4 units to apply the processing with the rightward and leftward direction, respectively.

IC1, IC2, IC3, and IC4 modules constitute the whole inner convolutional network, which propagates the spatial relationship from different directions. The inner convolutional network enhances the ability of the network to learn the specific semantic information and continuous prior information of roads, thus helping to extract the occluded road.

2.3. Directional Conditional Random Fields

CRF is a typical undirectional graph model [42]. As a post-processing method of image segmentation, CRF can reduce the false prediction of a target and improve the segmentation result. Fully connected CRF has been the most commonly used CRF model for image semantic segmentation in recent years [43]. By utilizing the energy function of adjacent nodes, pixels with a similar color and intensity are likely to be included in the same category.

However, the traditional fully connected CRF cannot be effectively applied to the post-processing of the road extraction task. The traditional CRF only uses color and location information to calculate the energy function of adjacent nodes, so, if the road is occluded in the remote sensing images due to the influence of light, shadow, and other factors, the color of the occluded road area is very different from that of the normal road area, so the traditional CRF cannot work in this circumstance.

In terms of the road, there is also a significant characteristic in that the road has an extending direction. Previous studies have shown that learning the road direction can improve the connectivity of road segmentation results [44,45]. Based on this, we considered adding the direction of roads to the energy function of CRF to improve the extraction of the occluded road. Notations used in the formalization are presented in Table 1.

Table 1. Description of notations.

Notation	Description
$\boldsymbol{M}_d$	Road direction map
$\mathbf{I}$	An image
$\mathbf{X}$	Segmentation map of $\mathbf{I}$
$\mathbf{G}$	A graph on $\mathbf{X}$
$\mathrm{P}(\mathbf{X}\|\mathbf{I})$	Gibbs distribution
$E(x\|\mathbf{I})$	Gibbs energy
$\phi_u(x)$	Unary potential function
$\phi_p(x_i, x_j)$	Pairwise potential function
$\mu(x_i, x_j)$	Compatibility function
$w^{(m)}$	Linear combination weights
$k^{(m)}$	Gaussian kernels
$\mathbf{f}_i$	Feature vector for pixels i in an arbitrary feature space
$\Lambda^{(m)}$	Positive-definite precision matrix

We generated the corresponding direction maps from binary maps of the segmentation result. The algorithm employed for generating the direction maps is shown in Algorithm 1.

Algorithm 1 Algorithm for Generating the Direction Map

Input: Binary map of road segmentation result $\boldsymbol{M}$
Parameters: angle step $\Delta\theta$, detecting radius r;
Output: road direction map $\boldsymbol{M}_d$
1. **for** $\boldsymbol{M}(i,j)$ **in** $\boldsymbol{M}$
2. **if** ($\boldsymbol{M}(i,j) = 1$)
3. **for** $\theta = 0$ to π step $\Delta\theta$
4. $d_\theta(i,j) = \sum_{\rho=1}^{r} \boldsymbol{M}(\rho\sin\theta, \rho\cos\theta) + \boldsymbol{M}(-\rho\sin\theta, -\rho\cos\theta)$
5. **end for**
6. find θ_{max} :
7. $d_{\theta_{max}}(i,j) = max\{d_{\theta_{max}}(i,j)\}, \theta \in [0,\pi]$
8. $\boldsymbol{M}_d(i,j) = \theta_{max}$
9. **else**
10. $\boldsymbol{M}_d(i,j) =$ null
11. **end if**
12. **end for**
13. **return** $\boldsymbol{M}_d$

Given an input image $\mathbf{I}$, defined over variables $\{\mathbf{I}_1, \cdots, \mathbf{I}_N\}$, $\mathbf{X}$ is its road segmentation result from DCNN. For any pixel $\mathbf{I}_i$, its segmentation result is x_i. Consider $\mathbf{G} = (V, E)$ is a graph on $\mathbf{X}$, where conditional random fields ($\mathbf{I}$,$\mathbf{X}$) can be characterized by a Gibbs distribution $\mathrm{P}(\mathbf{X}|\mathbf{I})$. The conditional probability of a pixel belonging to label x is

$$P(\mathbf{X} = x|\mathbf{I}) = \frac{1}{Z(\mathbf{I})}\exp(-E(x|\mathbf{I})), \tag{3}$$

where $Z(\mathbf{I}) = \sum P(\mathbf{X}|\mathbf{I})$ is the normalized constant and $E(x|\mathbf{I})$ is the energy function.

$$E(x|I) = \sum_i \phi_u(x_i) + \sum_{i<j} \phi_p(x_i, x_j) \tag{4}$$

In Equation (4), the unary potential function $\sum_i \phi_u(x_i)$ is independently computed by the segmentation result of the DCNN, and the pairwise potential function $\phi_p(x_i, x_j)$ can be expressed as

$$\phi_p(x_i, x_j) = \mu(x_i, x_j) \sum_{m=1}^{K} w^{(m)} k^{(m)} f(x_i, x_j), \tag{5}$$

where $\mu(x_i, x_j) = \begin{cases} 0, & x_i = x_j \\ 1, & x_i \neq x_j \end{cases}$, $w^{(m)}$ are the linear combination weights, and $k^{(m)}$ is a Gaussian kernel.

$$k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) = \exp\left(-\frac{1}{2}(\mathbf{f}_i - \mathbf{f}_j)^{\mathrm{T}} \Lambda^{(m)} (\mathbf{f}_i - \mathbf{f}_j)\right), \tag{6}$$

where the vectors $\mathbf{f}_i$ and $\mathbf{f}_j$ are feature vectors for pixels i and j, respectively, which can be chosen arbitrarily, and $\Lambda^{(m)}$ is the positive-definite precision matrix.

For image segmentation, traditional CRF uses two Gaussian kernels: The smoothness kernel and the appearance kernel. The smoothness kernel is defined in terms of positions p_i and p_j, which can remove small isolated regions. The appearance kernel k_α utilizes color vectors C_i and C_j as features, which encourages pixels with a similar color and position to have the same label. In addition to the smoothness kernel and the appearance kernel, we use the direction vectors D_i and D_j on the direction map $\boldsymbol{M}_d$ as the third Gaussian kernel—the direction kernel—which encourages pixels with a similar direction and position to be in the same class.

The entire pairwise potential can be defined as

$$k(\mathbf{f}_i, \mathbf{f}_j) = w^{(1)} \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2}\right) + w^{(2)} \exp\left(-\frac{|p_i - p_j|^2}{2\theta_{\beta_1}^2} - \frac{|C_i - C_j|^2}{2\theta_{\gamma_1}^2}\right) + w^{(3)} \exp\left(-\frac{|p_i - p_j|^2}{2\theta_{\beta_2}^2} - \frac{|D_i - D_j|^2}{2\theta_{\gamma_2}^2}\right), \quad (7)$$

where $w^{(1)}$, $w^{(2)}$, and $w^{(3)}$, as well as θ_α, θ_{β_1}, θ_{γ_1}, θ_{β_2}, and θ_{γ_2}, are the learnable parameters. The size of the Gaussian kernel is controlled by θ_α, θ_{β_1}, θ_{γ_1}, θ_{β_2}, and θ_{γ_2}, respectively.

Since the graph model in the image segmentation task has millions of nodes and edges, we used the mean field to approximate [38].

3. Experimental Results and Discussion

3.1. The Dataset and Preprocessing

The Massachusetts roads dataset [17] was selected for the validation experiments. The Massachusetts roads dataset contains 1171 images in total, including 1108 training images, 14 images for validation, and 49 images for the test. Each image is 1500 × 1500 pixels, with the ground sampling distance (GSD) of 1.2 m/pixel. All the images were captured over Massachusetts, US, containing urban, suburban, and rural regions, with an area of 2.25 km^2 Several samples are shown in Figure 3.

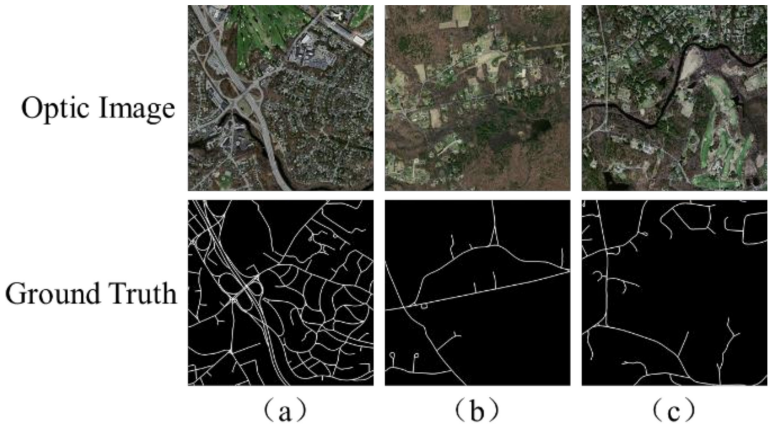

Figure 3. Examples of the Massachusetts roads dataset.: (**a**) town area, (**b**) rural area, (**c**) river area.

We conducted the experiments by using the PyTorch framework, with a GPU of NVIDIA GeForce GTX 1080Ti (11G), which was employed to accelerate the process. Limited to the capacity of the GPU memory, each original image was randomly cropped to 512 × 512 To increase the capacity and generalization of training data, we augmented the training data by rotating (90, 180, and 270 degrees) and flipping (horizontal and vertical) them.

Before training the network, we normalized the images to [−0.5,0.5] by 0-1 normalization and average value subtraction, in order to improve the stability of gradient calculation during the network training.

All of the networks in the experiments use the Adam algorithm to update the parameters. The parameters in the Adam algorithm are recommended in ref. [46], and include the learning rate $\alpha = 10^{-3}$, the attenuation coefficient $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and the constant $\varepsilon = 10^{-8}$. Additionally, the batch size is set to 6.

3.2. Evaluation Method

As road extraction can be viewed as a problem of semantic segmentation, we used the recall, precision, and F1-score to evaluate the extraction results. The precision, which is also called the correctness in remote sensing literature, is the ratio of predicted road pixels that are true roads. The recall, which is also called the completeness, is the ratio of true road pixels that are correctly detected. F1 is a comprehensive metric, which can be calculated by precision and recall.

True-Positive (TP) denotes the number of road pixels that are correctly identified. False-Positive (FP) denotes the number of non-road pixels that are detected as road pixels. False-Negative (FN) denotes the number of road pixels that are detected as non-roads. The metrics are calculated as follows:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{8}$$

$$\text{precison} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{9}$$

$$\text{F1} = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}}. \tag{10}$$

3.3. Experimental Results and Analysis

In comparative experiments, we compared the proposed method with other road extraction approaches. The quantitative comparison of different approaches for the test dataset is listed in Table 2.

Table 2. Results of different methods for the Massachusetts roads dataset.

Method	Precision	Recall	F1-Score
Wegner et al. [8]	40.5%	33.2%	35.9%
Wegner et al. [47]	47.1%	67.9%	55.6%
RSRCNN [27]	60.6%	72.9%	66.2%
FCN-4s [28]	71.0%	66.0%	68.4%
DeepLab v3+ [25]	74.9%	73.3%	74.0%
JointNet [30]	85.4%	71.9%	78.1%
Pixel-wiseNet [34]	77.4%	80.5%	78.9%
CasNet [29]	77.7%	80.9%	79.3%
ResUNet [31]	77.8%	81.1%	79.5%
DiResNet [45]	80.4%	79.4%	79.7%
Our method	87.1%	82.2%	84.6%

In Table 2, approaches 8 and 47 are traditional approaches, which do not use the framework of convolutional networks, and their F1-scores are lower. Among deep learning-based methods, the metrics show that the encoder-decoder architectures, especially the UNet-like networks, achieve positive performances, which proves that UNet-like networks are appropriate for road extraction from remote sensing images. Compared with the literature works, our proposed method performs the best.

An ablation study was performed to test the modules. Table 3 provides the quantitative comparison of the ablation study on the Massachusetts dataset. The Baseline network is the modified U-net introduced in Figure 2. Baseline-ICN integrates the ICN module at the top of the modified U-net. Baseline-CRF utilizes conditional random fields in post-processing, while baseline-DCRF utilizes directional conditional random fields in post-processing. Baseline-ICN-DCRF integrates the ICN module and uses DCRF as a post-processing step.

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

Method	Precision	Recall	F1-Score
Baseline	80.4%	78.6%	79.4%
Baseline-ICN	84.9%	81.7%	83.3%
Baseline-CRF	81.8%	77.6%	79.6%
Baseline-DCRF	82.2%	80.3%	81.3%
Baseline-ICN-DCRF	87.1%	82.2%	84.6%

As a widely used encoder-decoder network, U-net displays a good performance in road extraction research. Therefore, the baseline network used in this study also obtains

good road extraction results, and its precision, recall, and F1-score are 80.4%, 78.6%, and 79.4%, respectively. Compared to the baseline network, the precision, recall, and F1-score of Baseline-ICN are approximately improved by 4.5%, 3.1%, and 3.9%, respectively, which proves that the strategy of integrating the ICN module into the encoder-decoder network is effective.

Although CRF is a well-known post-processing method for semantic segmentation, it does not work well in the road extraction task. Compared with the results of the baseline network, the precision of Baseline-CRF is 1.4% higher, but its recall is reduced by 1.0%. This is because the CRF is more conducive to processing massive targets, rather than slender targets. The idea of using color as an energy function means that the regions with similar colors are combined as one category, which is not conducive to the extraction of road edges and occluded areas. The proposed DCRF not only considers the influence of color and location for pixel classification, but also considers the direction of the road, which can effectively improve the segmentation accuracy. The experimental results show that the recall of Baseline-DCRF is greatly improved, and the F1-score is also increased by 1.9%.

The precision, recall rate, and F1-score of Baseline-ICN-DCRF are 87.1%, 82.2%, and 84.6%, respectively. The proposed method achieves significant advantages over the baseline network, with an advantage of 5.2% for the F1-score, 3.6% for the recall, and 6.7% for the precision.

To visually assess the effect of ICN and DCRF, in Figure 4, we compare the extraction results with and without the use of these two models. The use of inner convolution enables the network to better embed the linear features, so as to improve the extraction of roads. DCRF can reduce false alarms and interruptions by utilizing the relationship between adjacent pixels.

In Figure 4a–c, the roads that need to be extracted are roads with complex backgrounds and high curvature. In this case, the road extracted by the baseline U-net is ambiguous, and the results of the ICN-DCRF are more complete and smoother. Figure 4d,e presents some examples of rivers, whose features are similar to those of the road. The baseline network cannot distinguish roads and rivers accurately, so the segmentation result may classify rivers as road targets, while the ICN-DCRF recognizes the river area correctly. Figure 4f,g show two samples in which the road network is partly occluded by trees. The results show that the ICN module enhances the connectivity of the road network, and DCRF can further remove noise points. In summary, compared with the baseline network, the proposed ICN-DCRF obtains more complete results with fewer false alarms. Indeed, ICN-DCRF significantly improves the precision of the results and reduces the broken segments by making better use of the linear features.

Figure 5 shows the extraction results when the road is covered by trees and shadows. As shown in Figure 5, the roads are partly covered by trees, and our proposed method can successfully extract the occluded road.

Figure 6 Shows more details about false positive and false negative predictions by the model. In Figure 6, typical false positive areas are caused by a road boundary or narrow roads. As the annotation of the road width is not completely accurate, it is hard to determine the road boundary accurately. Typical false negative pixels occur in areas where roads are covered by trees and shadows or interior roads in residences. Generally, our proposed method has far fewer red and green pixels than the baseline methods, as ICN integration can obtain road-specific contextual information to reduce FNs and achieve more coherent results, and the application of DCRF can remove false positive predictions to improve the image quality.

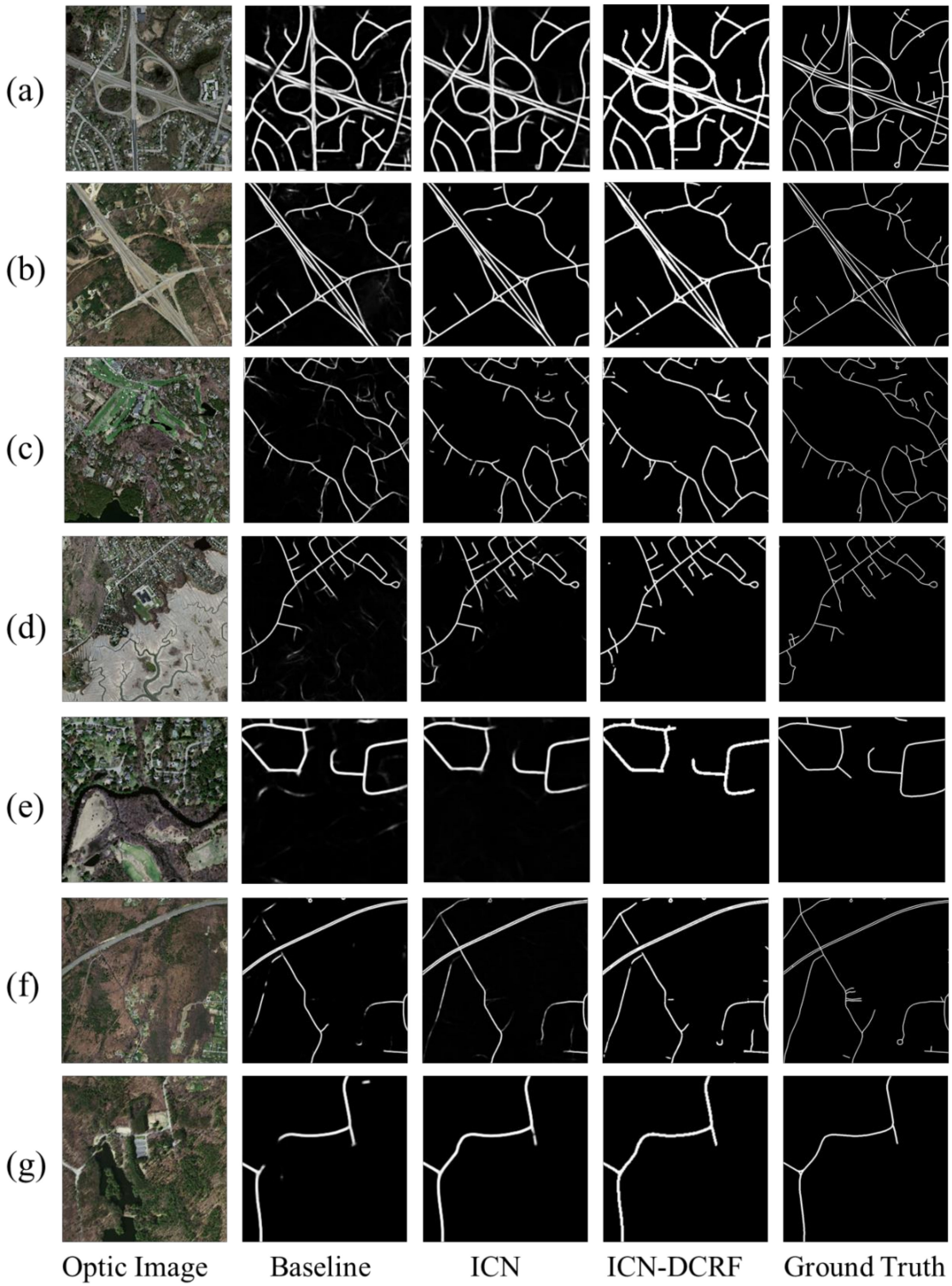

Figure 4. Visual comparisons of the ablation study using test samples from the Massachusetts roads dataset. Each row in (**a**–**g**) contains an RGB image, three result maps obtained by different models, and a ground truth map of a sample area.

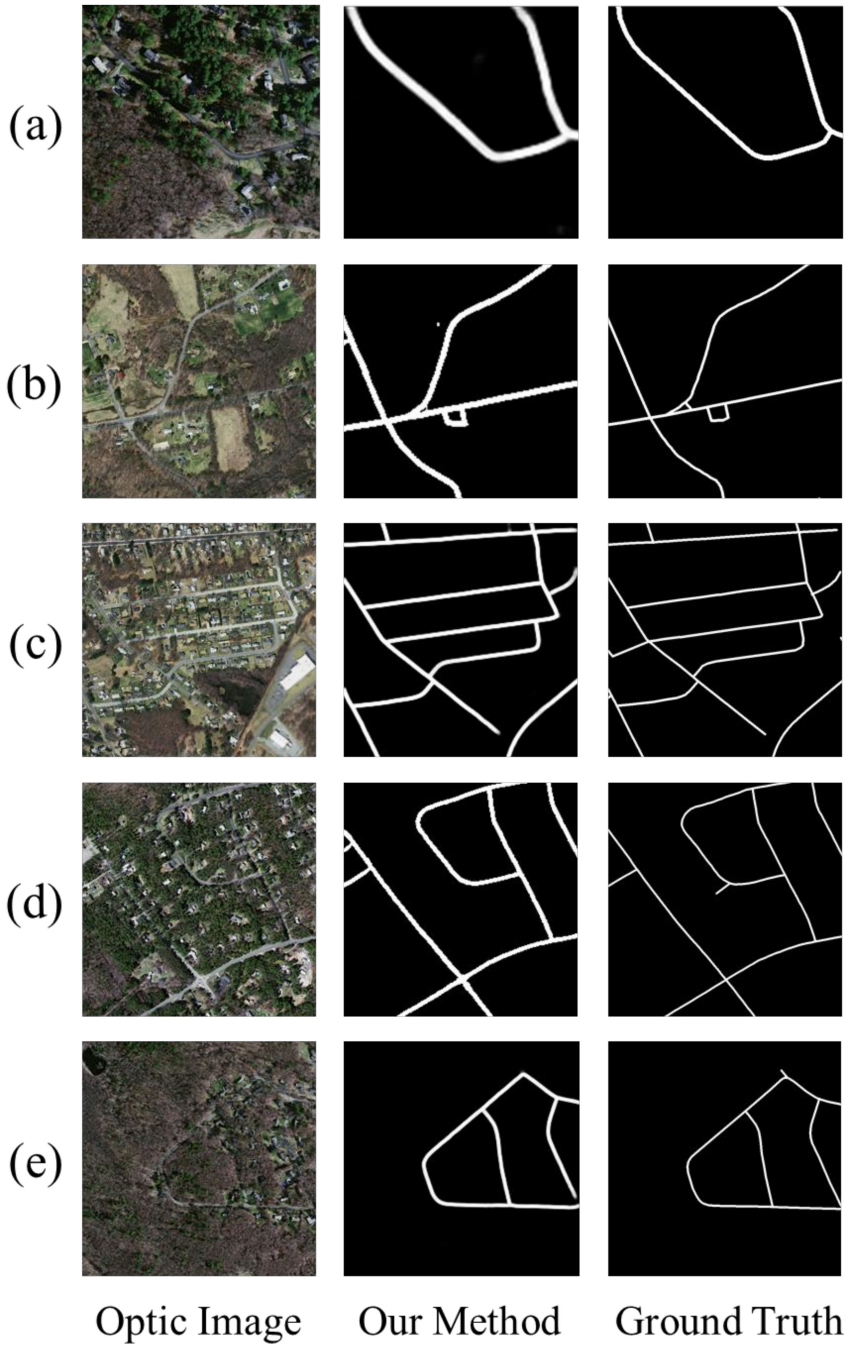

Figure 5. Visualization results of the road area occluded by trees: Each row in (**a**–**e**) contains an RGB image, a segmentation result map of the proposed method, and a ground truth map of a sample area.

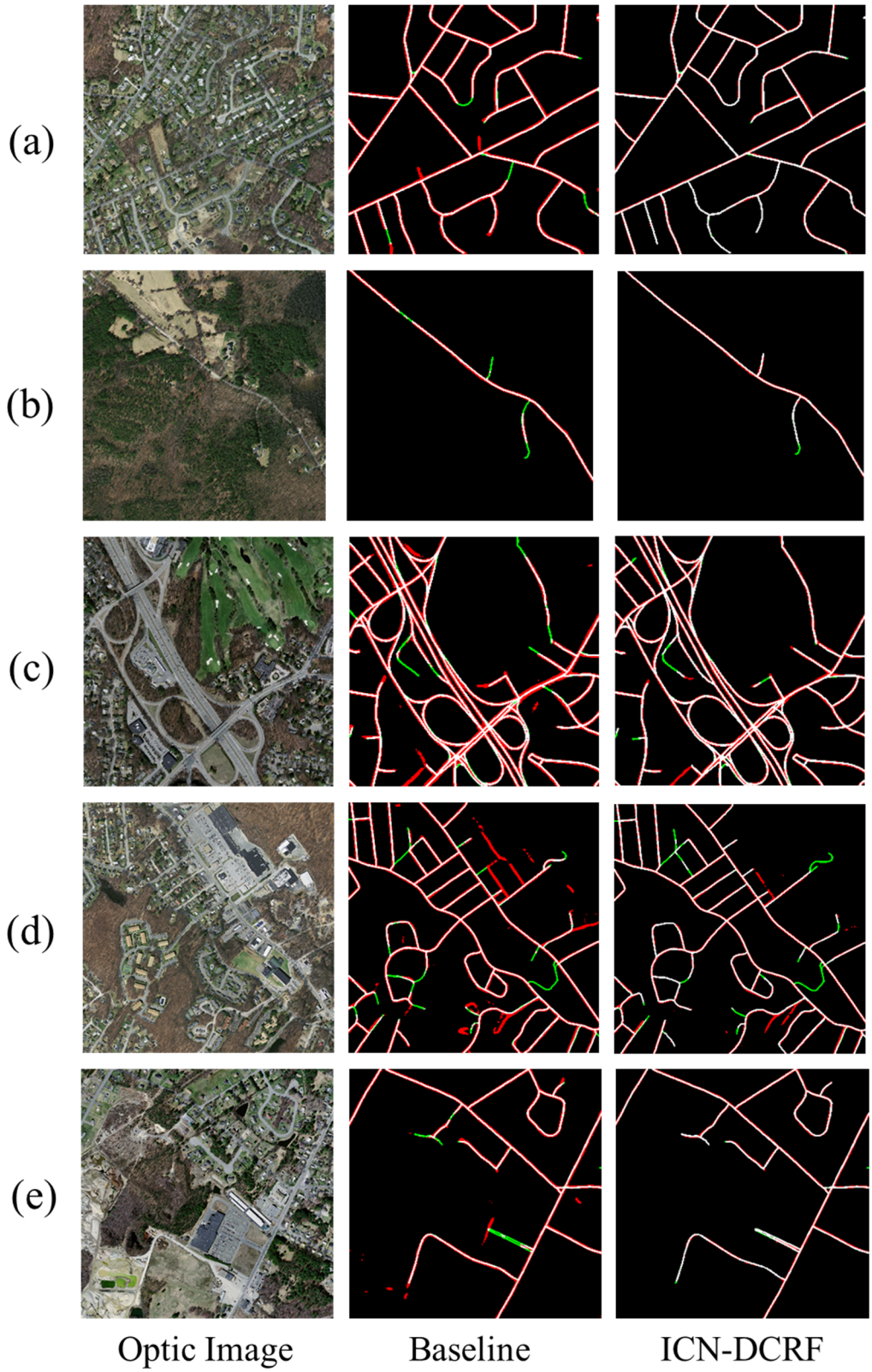

Figure 6. Error analysis of the proposed model. White pixels denote the road area that is detected correctly (TP), red pixels denote the background area which is detected as roads (FP), and green pixels denote the road area which is detected as background (FN). For each row in (**a**–**e**), it contains an optic image, and two result maps of a sample region.

However, for some short blurred roads and long occlusions, our proposed method detects these areas well. Another problem is that DCRF may remove several correct predictions which have low confidence levels.

When the road is occluded by trees, in a normal FCN, the occluded road pixels receive information from their neighboring pixels. However, the neighboring pixels all belong to trees, so these pixels may be classified as non-road areas. Comparatively, in ICN, contextual information propagates slice-by-slice, so the useful contextual information accumulates. The occluded road pixel can gradually receive information from distant road pixels, and it is more likely to be detected as the road pixel. The DCRF adds the direction of the road as an energy term, which favors the assignment of the same label to pixels with similar directions. Therefore, the ICN and DCRF are particularly suitable for linear structures. Further work can be carried out to examine the possibility of applying our proposed model to detect other objects with long continuous shape structures.

4. Conclusions

This study presents a method for road extraction from remote sensing images based on an inner convolution integrated encoder-decoder network and directional conditional random fields. Firstly, we proposed an inner convolutional network and applied it to the top hidden layer of the encoder-decoder network. The inner convolutional network enhances the ability of the network to embed the linear features and to learn the specific semantic information of roads, thus improving the accuracy and connectivity of the extracted road network. Secondly, on the basis of the fully connected CRF, we designed a directional CRF as a post-processing approach to improve the quality of the extracted results. The experiments on the Massachusetts road dataset showed that the proposed method can effectively improve the quality of road extraction, especially in the face of occlusion. The proposed method achieved an F1-score of 84.6%, which is about 5% higher than the F1-score of "state-of-the-art" competitors. Moreover, our approach is robust for extracting roads under occlusions.

As the proposed approach is a pixel-based segmentation method, we cannot obtain the topology of the road directly. In the future, we will investigate constructing a vector map of roads by predicting the nodes and direction and length of roads, which is more useful in GIS applications.

Author Contributions: The concept was proposed by D.Y. X.M. supervised the research and administrated this project. P.Z. conducted the investigation and offered supporting algorithms. The methodology and experiments were conducted by S.W. and H.H. The article was written by S.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (Grant No. 61673017, 61403398), and the Natural Science Foundation of Shaanxi Province (Grant No. 2017JM6077, 2018ZDXM-GY-039).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available, see reference [17] for details.

Acknowledgments: The authors would like to thank Volodymyr Mnih, from University of Toronto, Canada, for providing the Massachusetts Dataset used in the experiments.

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

Abbreviations

The following abbreviations are used in this manuscript:

BN	Batch Normalization
CNN	Convolutional Neural Network
CRF	Conditional Random Fields

DCRF	Directional Conditional Random Fields
DCNN	Deep Convolutional Neural Network
ELU	Exponential Linear Unit
FCN	Fully Convolutional Network
FP	False Positive
FSM	Finite State Machine
GIS	Geographic Information System
GSD	Ground Sampling Distance
ICN	Inner Convolutional Network
MRF	Markov Random Field
ReLU	Rectified Linear Unit
SVM	Support Vector Machines
TN	True Negative
TP	True Positive

References

1. Wang, W.; Yang, N.; Zhang, Y.; Wang, F.; Cao, T. A review of road extraction from remote sensing images. *J. Traffic Transp. Eng.* **2016**, *3*, 271–282. [CrossRef]
2. Wang, H.; Hou, Y.; Ren, M. A shape-aware road detection method for aerial images. *Int. J. Pattern Recognit. Artif. Intell.* **2017**, *31*, 1–21. [CrossRef]
3. Shao, Y.; Guo, B.; Hu, X.; Di, L. Application of a Fast Linear Feature Detector to Road Extraction From Remotely Sensed Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2011**, *4*, 626–631. [CrossRef]
4. Hu, J.; Razdan, A.; Femiani, J.; 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]
5. Ronggui, M.; Weixing, W.; Sheng, L. Extracting roads based on Retinex and improved Canny operator with shape criteria in vague and unevenly illuminated aerial images. *J. Appl. Remote Sens.* **2012**, *6*, 3610. [CrossRef]
6. Marikhu, R.; Dailey, M.N.; Makhanov, S.; Honda, K. A Family of Quadratic Snakes for Road Extraction. In Proceedings of the 8th Asia Conference on Computer Vision, Tokyo, Japan, 18–22 November 2007; pp. 85–94.
7. Sawano, H.; Okada, M. *A Road Extraction Method by an Active Contour Model with Inertia and Differential Features*; Oxford University Press: Oxford, UK, 2006.
8. Wegner, J.D.; Montoya-Zegarra, J.A.; Schindler, K. A higher-order CRF model for road network extraction. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 1698–1705.
9. Guo, C.; Mita, S.; McAllester, D.A. Adaptive non-planar road detection and tracking in challenging environments using segmentation-based Markov Random Field. In Proceedings of the IEEE International Conference on Robotics and Automation, Shanghai, China, 9–13 May 2011; pp. 1172–1179.
10. Zhou, S.; Gong, J.; Xiong, G.; Chen, H.; Iagnemma, K. Road detection using support vector machine based on online learning and evaluation. In Proceedings of the IEEE Intelligent Vehicles Symposium, La Jolla, CA, USA, 21–24 June 2010; pp. 256–261.
11. Yager, N.; Sowmya, A. Support Vector Machines for Road Extraction from Remotely Sensed Images. In Proceedings of the 10th International Conference on Computer Analysis of Images and Patterns, Groningen, The Netherlands, 25–27 August 2003; pp. 285–292.
12. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
13. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
14. 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 (CVPR), Los Alamitos, CA, USA, 27–30 June 2016; pp. 770–778.
15. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
16. Mnih, V.; Hinton, G.E. Learning to detect roads in high-resolution aerial images. In Proceedings of the European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010; pp. 210–223.
17. Mnih, V. Machine Learning for Aerial Image Labeling. Ph.D. Thesis, University of Toronto, Toronto, ON, Canada, 2013.
18. 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]
19. 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]
20. Rezaee, M.; Zhang, Y. Road detection using Deep Neural Network in high spatial resolution images. In Proceedings of the Joint Urban Remote Sensing Event (JURSE 2017), Dubai, United Arab Emirates, 6–8 March 2017; pp. 1–4.
21. 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.

22. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
23. 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]
24. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
25. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with atrous separable convolution for semantic image segmentation. *arXiv* **2018**, arXiv:1802.02611.
26. Abdollahi, A.; Pradhan, B.; Shukla, N.; Chakraborty, S.; Alamri, A.M. Deep Learning Approaches Applied to Remote Sensing Datasets for Road Extraction: A State-Of-The-Art Review. *Remote Sens.* **2020**, *12*, 1444. [CrossRef]
27. 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]
28. Zhong, Z.; Li, J.; Cui, W.; Jiang, H. Fully convolutional networks for building and road extraction: Preliminary results. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 1591–1594.
29. 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]
30. Zhang, Z.; Wang, Y. JointNet: A Common Neural Network for Road and Building Extraction. *Remote Sens.* **2019**, *11*, 696. [CrossRef]
31. Zhang, Z.; Liu, Q.; Wang, Y. Road extraction by deep residual U-net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
32. 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]
33. Li, Y.; Peng, B.; He, L.; Fan, K.; Li, Z.; Tong, L. Road extraction from unmanned aerial vehicle remote sensing images based on improved neural networks. *Sensors* **2019**, *19*, 4115. [CrossRef]
34. Mosinska, A.; Márquez-Neila, P.; Kozinski, M.; Fua, P. Beyond the pixel-wise loss for topology-aware delineation. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Munich, Germany, 8–14 September 2018; pp. 3136–3145.
35. Abdollahi, A.; Pradhan, B.; Alamri, A.M. VNet: An End-to-End Fully Convolutional Neural Network for Road Extraction from High-Resolution Remote Sensing Data. *IEEE Access* **2020**, *8*, 179424–179436. [CrossRef]
36. 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]
37. Sun, T.; Chen, Z.; Yang, W.; Wang, Y. Stacked U-nets with multi-output for road extraction. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Work, Salt Lake City, UT, USA, 18–22 June 2018; pp. 187–191.
38. Liangchieh, C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *Compute* **2014**, *40*, 357–361.
39. Panboonyuen, T.; Jitkajornwanich, K.; Lawawirojwong, S.; Srestasathiern, P.; Vateekul, P. Road segmentation of remotely-sensed images using deep convolutional neural networks with landscape metrics and conditional random fields. *Remote Sens.* **2017**, *9*, 680. [CrossRef]
40. Clevert, D.A.; Unterthiner, T.; Hochreiter, S. Fast and accurate deep network learning by exponential linear units (ELUs). *arXiv* **2015**, arXiv:1511.07289.
41. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 448–456.
42. Lafferty, J.D.; McCallum, A.; Pereira, F.C.N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the Eighteenth International Conference on Machine Learning, Williamstown, MA, USA, 28 June–1 July 2001; pp. 282–289.
43. Krähenbühl, P.; Koltun, V. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. In Proceedings of the Advances in Neural Information Processing Systems, Granada, Spain, 12–14 December 2011; pp. 109–117.
44. Batra, A.; Singh, S.; Pang, G.; Basu, S.; Jawahar, C.V.; Paluri, M. Improved Road Connectivity by Joint Learning of Orientation and Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 10385–10393.
45. Ding, L.; Bruzzone, L. DiResNet: Direction-aware Residual Network for Road Extraction in VHR Remote Sensing Images. *arXiv* **2020**, arXiv:2005.07232. [CrossRef]
46. Kingma, D.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
47. Wegner, J.D.; Montoya-Zegarra, J.A.; Schindler, K. Road networks as collections of minimum cost paths. *ISPRS J. Photogramm. Remote Sens.* **2015**, *108*, 128–137. [CrossRef]