remote sensing

MDPI

Article

Road Extraction from High-Resolution Remote Sensing Imagery Using Deep Learning

Yongyang Xu [1], Zhong Xie [1,2], Yaxing Feng [1] and Zhanlong Chen [1,2,*]

[1] Department of Information Engineering, China University of Geosciences, Wuhan 430074, China; yongyangxu@cug.edu.cn (Y.X.); xiezhong@cug.edu.cn (Z.X.); fyx@cug.edu.cn (Y.F.)
[2] National Engineering Research Center of Geographic Information System, Wuhan 430074, China
[*] Correspondence: chenzl@cug.edu.cn

Received: 20 August 2018; Accepted: 11 September 2018; Published: 13 September 2018

Abstract: The road network plays an important role in the modern traffic system; as development occurs, the road structure changes frequently. Owing to the advancements in the field of high-resolution remote sensing, and the success of semantic segmentation success using deep learning in computer version, extracting the road network from high-resolution remote sensing imagery is becoming increasingly popular, and has become a new tool to update the geospatial database. Considering that the training dataset of the deep convolutional neural network will be clipped to a fixed size, which lead to the roads run through each sample, and that different kinds of road types have different widths, this work provides a segmentation model that was designed based on densely connected convolutional networks (DenseNet) and introduces the local and global attention units. The aim of this work is to propose a novel road extraction method that can efficiently extract the road network from remote sensing imagery with local and global information. A dataset from Google Earth was used to validate the method, and experiments showed that the proposed deep convolutional neural network can extract the road network accurately and effectively. This method also achieves a harmonic mean of precision and recall higher than other machine learning and deep learning methods.

Keywords: road network extraction; deep learning; pyramid attention; global attention; high resolution

1. Introduction

With the rapid development of remote sensing technology, high-resolution remote sensing imagery has been widely used in many applications, including disaster management, urban planning, and building footprint extraction [1–3]. Roads network play a key role in the development of transportation systems, including the addition of automatic road navigation and unmanned vehicles, and urban planning [4], which is important in both industry and daily living. Therefore, developing a new method to extract road networks from high-resolution remote sensing imagery would be beneficial to geographical information systems (GIS) and intelligent transportation systems (ITS) [5–7]. Extracting road networks has become one of the main research topics in the field of remote sensing imagery processing, and high-resolution imagery has become an important data source to update roads network in the geospatial database in real-time [8].

The structure of roads is complex and the road segments are irregular; the shadows of trees or buildings on roadsides and the vehicles on the roads can be observed from high resolution imagery [9], on the other hand, insufficient context of the roads in the remote sensing imagery is similar with the roof of the buildings. The aforementioned issues make it more difficult to extract the road networks from high-resolution imagery.

Traditional artificial extraction methods were generally time consuming and contained abundant mistakes made by human operators [6]. Recently, much research regarding the information extracted from images in computer version fields has been used to successfully extract road networks in remote sensing imagery [10,11]. A variety of road network extraction methods have been proposed by the researchers, including unsupervised and supervised classification methods. Most of these methods are based on classification; these methods extract the roads from remote sensing imagery using their geometric, photometric, and textural feature.

Unsupervised methods usually extract roads using clustering algorithms. Miao et al. proposed a semi-automatic method using the mean shift to detect roads [12]. The method extracts the initial point from the road seed points, and then a threshold is used to separate the road and non-road. Unsalan and Sirmacek extracted the road network based on the probability and graph theory [13]. Wang et al. extracted the roads using object-based methods [14]. Their method contains three steps: texture information extraction, road extraction and post-processing. Compared with unsupervised methods, supervised methods are generally more accurate [15]. These methods, which include SVM, random decision forests, and deep learning, extract the roads based on training using labeled samples [16]. Anwer et al. proposed two-stream deep architecture, where texture coded mapped images and RGB stream were fused for remote sensing scene classification [17]. Yager and Sowmya used features including gradient, intensity, edge length, etc., to train the SVM classifier and extract the roads from remote sensing imagery [18]. Simler used 3-class SVM to detect roads by exploiting both spatial and spectral features [19], which performed better on very high resolution (VHR) multispectral aerial images. Markov random fields (MRFs) are widely used in edge detection, semantic segmentation, etc. [20]. Zhu et al. proposed a MAP-MRF framework using MRF to extract the roads, while an SVM plus Fuzzy C-Mean (FCM) model was proposed for semantic segmentation [21].

Encouraged by the good performance of the deep learning in kinds of applications [22,23], the artificial intelligence (AI) methods have now attracted the interest of researchers to extracted roads from high resolution satellite images [6,24–26]. The Fully Convolutional Network (FCN) and its development model exhibit excellent information extraction capabilities [27–30]. Ramesh et al. designed the a U-shaped FCN (UFCN) for road extraction by a stack of convolutions followed by deconvolutions with skip connections [31]. Zhang et al. combined deep residual networks (ResNet) [32] and U-Net [33], which allow for networks to be designed with fewer parameters, but obtain better results [4].

Recently, some researches about local and global attention have been incorporated into neural network architectures. Luong designed the attention-based neural network for translation [34]. Shang proposed neural responding machine with local and global attention for short-text conversation [35]. Cui designed the attention-over-attention neural networks for reading comprehension [36]. These attention units were designed for language or text processing, and they cannot be used for extracting information form remote sensing imagery, especially for extracting the roads, directly, because the remote sensing imagery has multiple scales and most of the roads are thin but run thought almost all the sample imagery.

Previous studies have provided useful insights into semantic segmentation, which can be used to extract roads from remote sensing imagery. However, high-resolution remote sensing imagery provide not only detailed road information but also significant noise, including building shadows, vehicles on the road, as well as trees along the roads. Furthermore, road structures are complex in reality; different types of roads have different widths, and the surfaces of some buildings have the same spectral characteristics as the roads. These problems make it difficult to extract local detailed road information. A raw remote image has millions of pixels and is difficult to process directly. Therefore, during road extraction, the images are clipped into samples measuring 512 × 512 pixels, 256 × 256 pixels, or other size. As the roads are continuous structures, they will run through the clipped images. The global information in the clipped samples holds key morphological characteristics of road structures. To resolve these mentioned problems, the detail local information including the

the shadows and vehicles on the road as well as the turning information of the road, and the global information about the roads continuity and the morphological structure should be extracted effectively. A semantic labelling method is required to account for the global and local context and to increase the accuracy of extraction of road networks from remote sensing imagery.

Considering the feature extractor of densely connected convolutional networks (DenseNet) [37] is powerful enough, and it can take full advantage of features with less training time [37]; the symmetric architecture of U-Net is good at semantic segmentation, this study attempts to improve the performance of road extraction from remote sensing imagery based on DenseNet and U-Net. This work defined the local information as detailed local context, such as the shadows of buildings and trees, vehicles on the road and also the turning information of the road. In a sample, the global information is defined as the continuity and the morphological structure of the roads. To extract the local and global information defined above accurately, the local and global feature attention blocks are proposed, which were designed by operating the convolutional and pooling layers (Sections 2.2 and 2.3) and aim to pay attention to the local and global information respectively. They were designed to extract the local and global road information richly and accurately. The method comprised of two steps: first, all the remote sensing imagery, as well as the corresponding ground truth labels, were pre-processing to prepare the dataset to train the designed model. Then, a deep neural network was designed to extract roads from the pre-processed images. The proposed model produced binary maps, where the roads were treated as the foreground and all the other objects were treated as the background. All the challenges have resulted in an improvement in the extraction accuracy of the road network from remote sensing imagery. The major contribution of this work is proposing a new model, which learnt from the symmetric architecture of U-Net and was designed as contracting and expansive. DenseNet feature extractor was used to extract the road features in contracting. Two units, the local and global feature attention modules, were designed in the model of expansive. The proposed architecture defined as global and local attention model based on U-Net and DenseNet (GL-Dense-U-Net). This paper explores a novel supervised learning framework to extract roads from remote sensing imagery, which was confirmed as accurate and effective by experimental results.

The paper is organized as follows. Section 2 presents the proposed approach. Section 3 describes the experiment results and the parameters. Section 4 is a discussion of the method and Section 5 presents the concluding remarks.

2. Methods for Roads Extraction from High Resolution Remote Sensing Imagery

In this paper, a new deep neural network of image semantic segmentation model was proposed to extract the roads from remote sensing imagery. First, the original remote sensing imagery were pre-processed. To prepare the dataset for training the designed model, all the remote sensing imagery and corresponding ground truth images were clipped by a fixed-size sliding-window. Then, the designed deep neural network GL-Dense-U-Net model was introduced to extract the roads. All the pre-processed samples were treated as the input of the model, and the output of the trained model was the two-category classification maps. The categories were “road” and “others”, which represent the road extraction results.

2.1. The Structure of Deep Convolution Neural Network

The proposed model in this paper was designed based on the DenseNet, which has shown good performance in image classification, and is famous for several advantages, such as: alleviating the vanishing-gradient problem in most deep neural networks; being powerful enough to extract features and strengthen the feature propagation during training and evaluation, at the same time, it can encourage feature reuse for classification or semantic segmentation; most importantly, the DenseNet can reduce the number of parameters, which makes it easy to be trained. Meanwhile, encouraged by the performance of the symmetrical structure of U-Net [33] in semantic segmentation, the GL-Dense-U-Net was designed with two parts (Figure 1). The first part is designed to extract the features using the

DenseNet (top of Figure 1), known as the contracting part. The second (expansive) part is designed to generate the classification map based on the extracted features at different stages of the contracting part (bottom of Figure 1). Each box represents the feature map with size of w × h × c existing in the top part of Figure 1, where w and h represent the width and high, respectively, and c is the channels number of the feature map. To take full advantage of the features in different stages and obtain good performance of road extraction, these two parts were connected by a proposed local attention unit (LAU).

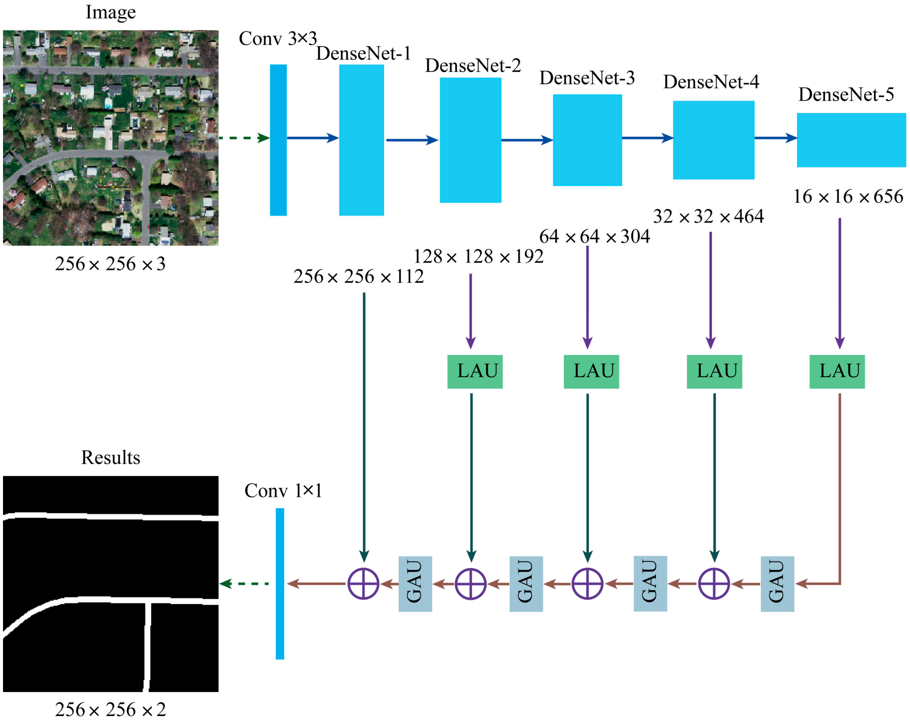

Figure 1. The architecture of the GL-Dense-U-Net used in this work.

Compared with some other traditional deep convolution neural network structures, for each feature extractor layer of the DenseNet, all the preceding layers were treated as the input of the layer. Therefore, there are $L\ (L + 1)/2$ connections in an L-layer instead of only L, as is the case for some traditional structures. In this way, The DenseNet will require fewer parameters during training, because there is no need to re-learn redundant features. At the same time, each layer in the contracting part has access to the gradients which form a loss at both the end of the model and at the beginning of the structure, which improves the flow of information between layers and makes the weights and biases easy to be trained.

The direct connection pattern is used in the dense block, where all the layers are connected, and this structure improves the information flow between layers. To make sure all the layers in a dense block are with same size, a 3 × 3 convolution with a padding operation is used in the block following the BN layer [38] and ReLU layer [39], which is defined as non-linear transformation $T_l()$. Any feature map x_l can be calculated by the preceding layers, including $x_0, \ldots\ x_{l-1}$ by $T_l()$, as follows:

$$x_l = T_l([x_0, x_1, \ldots, x_{l-1}]) \tag{1}$$

where $[x_0, x_1, \ldots, x_{l-1}]$ is the concatenation operation of all the output layers $0, \ldots\ l-1$.

As an important operation in the deep neural network, pooling [40] changes the size of the feature maps, which aids in extracting the information from different levels during training, and acts as a component between the convolution (Conv) layers. To facilitate down-sampling in the DenseNet model and to take full advantage of the architecture, the dense blocks were connected by the pooling operation following a 1 × 1 convolution with padding operation.

The architecture of the proposed deep neural convolution network is symmetrical. The expansive part is used to recover the road networks from feature maps extracted by contracting part. Every dense block in the contracting part corresponds to a local attention unit (LAU) and a global attention unit (GAU) in the expansive part, where a LAU is designed to extract the roads local information from remote sensing imagery during the different neural network training stages. GAUs are used to extract the global road information when recovering the information from deep level feature maps. The designed extraction model is end-to-end, and the size of the output is the same as the input remote sensing image. Therefore, some up-sampling operations are required in the expansive part. During expanding, a LAU and a GAU are regarded as a group, which is followed by a deconvolution layer [41] to enlarge the size of the feature maps. At the end of the model, a 1 × 1 convolution is used to map the feature maps into two classes: road and non-road; the problem of extracting roads can be resolved by binary classification. Training the deep neural convolution network is a process of minimizing the energy function via the gradient descent [42]. Because the output of the softmax function can be used to represent a categorical distribution, the energy function of model is defined as the cross-entropy between the estimated class probabilities and the "true" distribution. To train the convolutional neural layers and classifier coherently, soft-max layer and cross entropy [33] were used in the network, where the soft-max layer is defined to calculate the classification probability, as follows:

$$p_i = \frac{\exp(a_i)}{\sum\limits_{k=1}^{K} \exp(a_k)} \tag{2}$$

where p_i represents the probability that a pixel is predicted to belong to class i. Because the images in this work are classified into the foreground and background, the number of classes K is set as 2 in the experiment. a is the output of the last layer in the model. In this work, the energy function can be defined as follows:

$$E = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{K}[y_i^n \ln p_i^n + (1 - y_i^n)\ln(1 - p_i^n)] \tag{3}$$

where, N represents the number of samples in training dataset, y is the expected output and p is the probability mentioned above.

2.2. Local Attention Unit

Inspired by pyramid feature maps extraction, the local attention unit was designed to provide precise pixel-level attention to the feature maps extracted by DenseNet from deep levels. Benefitting from the special structure, pyramid pooling can extract information from different scale feature maps; at the same time, this design method helps to increase the receptive field, and is widely used in semantic segmentation [43,44]. However, the pyramid structure does not give significant attention to the global context information, and the channel-wise attention vector used in the structure is limited to extract pixel-wise information [45].

Considering the analysis above, in order to extract the local pixel-level information of road networks from remote sensing, the LAU is designed to fuse different scale feature maps and draw attention to the pixel-level information from deep-level of the DenseNet. To improve the performance of the LAU in extracting information from different scale feature maps, this paper applied four different convolution operations with kernel sizes of 1 × 1, 3 × 3, 5 × 5, 7 × 7. The features are integrated by the LAU from bottom to top in a stepwise fashion (Figure 2), in this way, the context information from neighboring scales can be incorporated precisely. At the top of the LAU, the 1 × 1 convolution is designed to be multiplied pixel-wise by the feature information extracted from bottom convolution operations. The pyramid structure to fuse different scale information, while the pixel-wise multiplication allows for better extraction of local pixel-level information for road extraction.

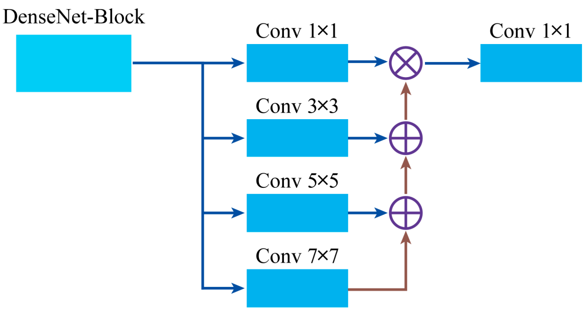

Figure 2. The architecture of the local attention unit.

2.3. Global Attention Unit

A road is a continuum, and so runs through all the images. Therefore, the global information is important in the expansive part of road extraction from remote sensing imagery. There are some semantic segmentation models designed for this part using bilinearly up-sampling to directly generate the results [46,47]. The one-step decoder is limited to recovering location information and will lose some global road features due to a lack of different scale low level feature-maps.

Encouraged by recent research, in which the contracting part was combined with the pyramid structure to improve the performance of semantic segmentation, the global attention unit (GAU) was designed in the expansive part. This GAU introduces global average pooling (GAP) into the unit to extract global road information, which is connected to the result of deconvolution and is used as a guide to recover the information (Figure 3). In detail, firstly, a 1 × 1 convolution and a dense block, which corresponds with the block in contracting part, are applied to operate the feature maps from a high-level; then, the features are operated in two ways, one using a deconvolution layer, and the other employing a GAP operation followed by a 1 × 1 convolution and deconvolution. Finally, features from the two methods are added together as the output of the GAU. This proposed unit considers the feature maps from both low-levels and high-levels, effectively, and provides the global information to guide feature recovery in the expansive part of the network.

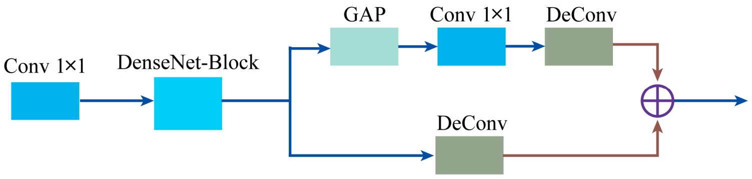

Figure 3. The architecture of the global attention unit.

3. Results

3.1. Dataset

The high-resolution remote sensing imagery collected by Cheng, et al. [48] from screenshots taken from Google Earth [49] were used in this work. The ground truth was manually labeled by the reference maps, and the dataset is publicly available. There are 224 images in the dataset, and at least 600 × 600 pixels in each image. The spatial resolution of every pixel in the remote sensing imagery is 1.2 m as description in [48]. The dataset covers urban, suburban and rural regions. There are waters, mountains and hills and lands covered by the vegetation. Most original images are under complex grounds, which make the road extraction task very challenging. The ground truth images have two kinds of pixels: including road and unknown (clutter); the road widths in the ground truth images are about 12–15 pixels (Figure 4). To avoid the network learns using pixels that are reserved to the independent testing dataset, all the samples and ground truth images were divided into two parts randomly, where eighty percent were used to train the GL-Dense-U-Net model and twenty percent were used to validate the trained model. The whole validating dataset were clipped into 256 × 256 non-overlapping samples. To enhance the training samples, all the original images, as well as the corresponding ground truth images, were clipped by a 256 × 256 sliding window with a stride of 64 pixels. To increase the size of our dataset and avoid padding or null values, all the clipped square samples were rotated by 90°, 180° and 270°.

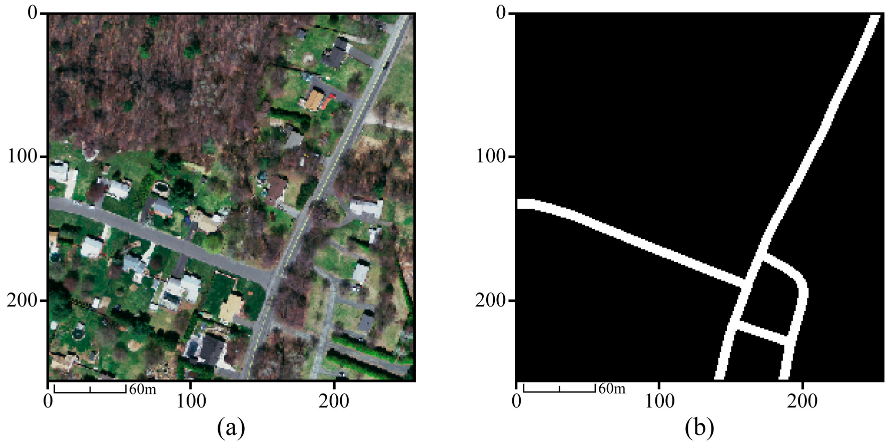

Figure 4. Samples of the remote sensing imagery used in the experiments. The RGB image (**a**) and the corresponding label (**b**).

3.2. Experimental Setup and Results

All the prepared samples, as well as the ground truth images, were treated as inputs to train the proposed GL-Dense-U-Net model. The architecture and the parameters, including the size of every block and number of blocks, etc., used in this work are shown in Figure 1. To improve the speed of the training model, this work restored the weights of DenseNet which were pre-trained by the ImageNet dataset for the contracting part. During training the model, an excellent optimization is required to minimize the energy function and update the parameters of the model algorithm. Adam (Adaptive Moment Estimation), one of the most commonly used algorithms [50], was treated as the network optimizer to minimize the losses and update the parameters, including weights, biases, and so on. To obtain a better performance and speed up the processing, during training, the learning rate of the GL-Dense-U-Net model was set to 0.001, and was divided by 10 every 10,000 iterations.

Edges are important for roads extracting from the remote sensing images. Edge enhancement is designed to reduce the noise by a filter and decrease the computation complexity. There are some filters used in edge enhancing including contour filter, detail filter, edge enhance filter and so on. All of these filters have been implemented by some image processing libraries such as python imaging library (PIL) and the OpenCV. In essence, these filters are convolutional filter, they enhance the edges of images by sum-weighted value of the convolution region. It is known that convolutional layers, the extractor of DenseNet designed in the contracting part of proposed model, tend to extract low-level features in the first layer, such as edges [51]. As shown in Figure 5, the edges can be extracted clearly, which would play the role of edge enhancement and be beneficial for the roads extraction.

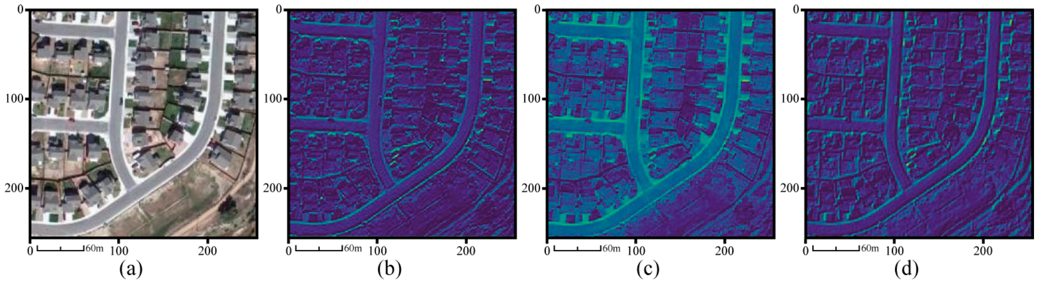

Figure 5. The original image (**a**) and a part of feature maps (**b**–**d**) of first layer of designed model.

To assess the quantitative performance of the proposed GL-Dense-U-Net model in road network extraction from remote sensing imagery, the precision (P) and recall (R) [52] are introduced, as well as the F_1 score [53] and the overall accuracy (OA) metrics. The F_1 score is calculated by P and R, and it is a powerful evaluation metric for the harmonic mean of P and R, and it can be calculated as follows:

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{4}$$

where

$$P = \frac{TP}{TP + FP},\ R = \frac{TP}{TP + FN} \tag{5}$$

Here, R measures the proportion of matched pixels in the ground truth and P is the percentage of matched pixels in the extraction results. TP, FP and FN represent the number of true positives, false positives and false negatives, respectively. OA measures the precision of road and non-road at pixel level and it can be calculated as follows:

$$OA = \frac{Pos_r + Pos_n}{N} \tag{6}$$

where Pos_r and Pos_n represent the positive number for road and non-road at pixel level, and N is the pixels number of test imagery. All the metrics mentioned above can be calculated by the pixel-based confusion matrices [54].

In this work, the proposed GL-Dense-U-Net model was implemented using the open source framework Tensorflow, provided by Google. The code was executed in the Linux platform with four TITAN GPUs (12 GB RAM per GPU). There were 100,000 iterations; the trained model achieved state-of-the-art results (Table 1). Figure 6 shows the changes in accuracy and losses with increasing iterations during the training of the model.

Table 1. The metrics of the model, including overall accuracy for the classification, precision and recall, as well as the F_1 score for road extraction from remote sensing imagery.

Class	Precision	Recall	F_1	OA
R	0.9630	0.9515	0.9572	0.9782
C	0.9936	0.9956	0.9946	

Where R stands for roads, and C represent clutter, and OA represents overall accuracy.

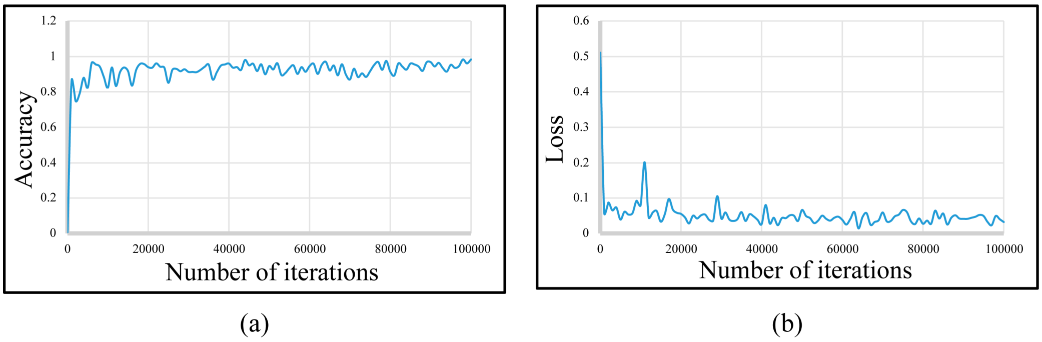

Figure 6. Plots showing the accuracy (**a**) and loss (**b**) of the GL-Dense-U-Net model while training the datasets with increasing iterations.

The trained GL-Dense-U-Net reached a 97.82% overall accuracy, proving that the deep convolutional neural network performs well in extracting roads from high-resolution remote sensing imagery. To prove that the proposed method works in a generic sense, some data in commercial areas, rural areas, deserts and imagery covered by vegetation were used to validate the proposed method, respectively (Figure 7). There are five rows and three columns of subfigures. The first column represents the original image, the second column represents the corresponding ground truth and the last column is the prediction by the proposed method. The performance of designed method in commercial areas was illustrated by the first row. From the second row to the last row were used to illustrate the results of deserts, rural, tropical and residential areas, separately.

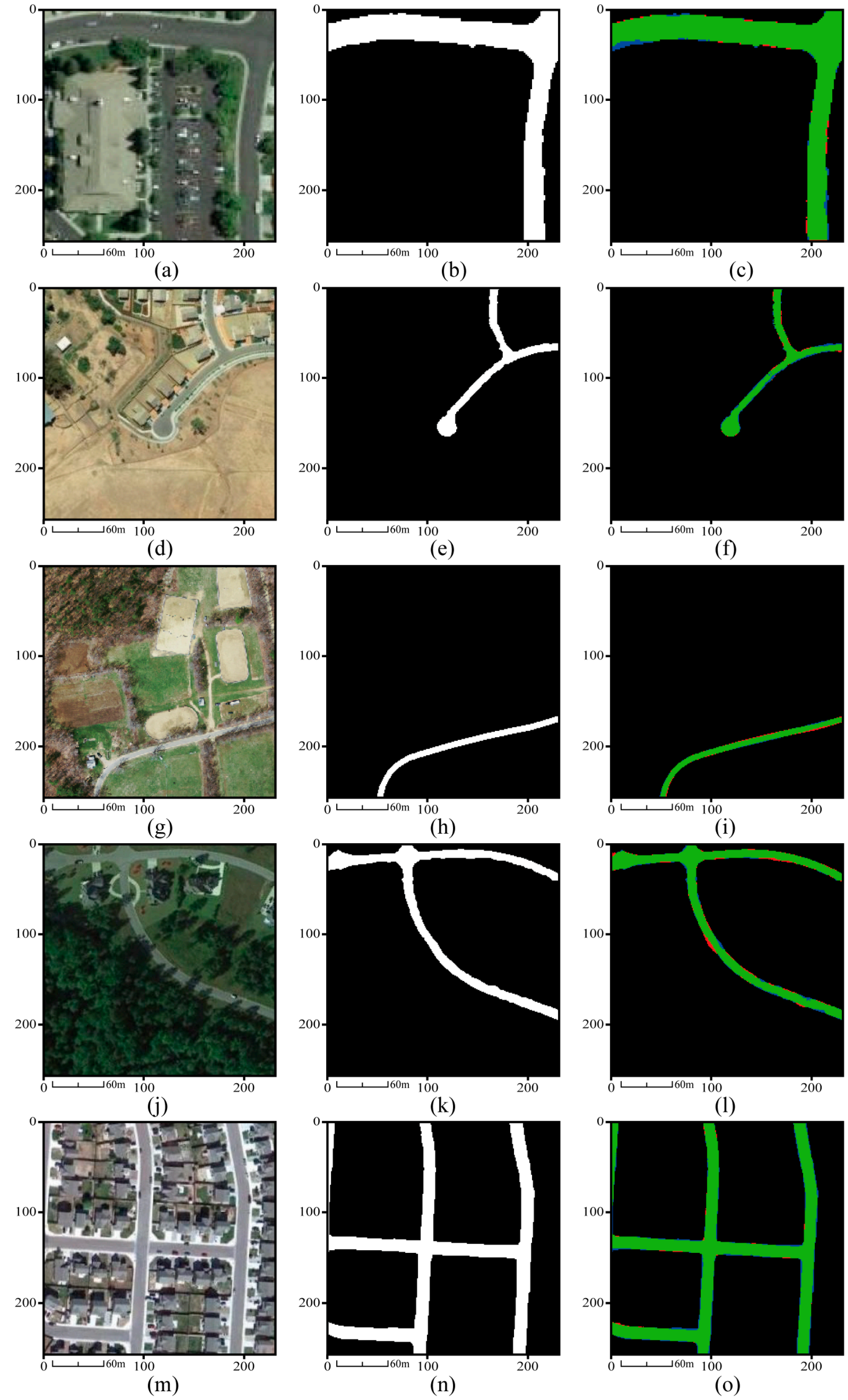

Figure 7. Results of roads extraction using the proposed GL-Dense-U-Net model. The original images (**a**,**d**,**g**,**j**,**m**), while the corresponding ground truths (**b**,**e**,**h**,**k**,**n**) and predictions (**c**,**f**,**i**,**l**,**o**) are also given. Green, red and blue represents the TP, FP and FN, respectively.

3.3. Comparison of the Proposed Method

This work uses the Dense Convolutional Network as the feature extractor in the contracting part, which allows direct connections between two layers in a block, while keeping layers in the same block within the same feature map size. During training, the DenseNet can achieve state-of-the-art results with fewer parameters and less computation. The feature extractor follows a simple connectivity rule, which is beneficial to integrating identity maps, deep supervision and diversified depth. In this way, the deep neural convolutional network can consequently learn more information and reduce feature redundancy for road extraction. The designed structure has shown good performance in image classification. To demonstrate the function of the DenseNet in semantic segmentation, this work compared the model which use DenseNet as the feature extractor with the model using a general convolutional neural network as the feature extractor (U-Net). The comparison results were shown in Figure 8.

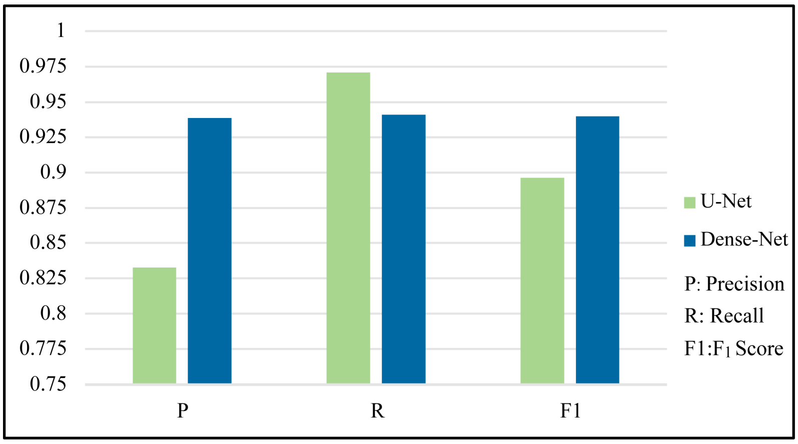

Figure 8. The results of road extraction from high-resolution remote sensing imagery using U-Net and DenseNet as feature extractors in the designed model.

The recall of U-Net is clear higher than the model using DenseNet as the feature extractor, while the road extraction precision and F_1 scores have improved by 10.59% and 4.33%, respectively. By analyzing the results, it is clear that DenseNet is able to extract more road information, allowing more precise pixel classification.

The proposed GL-Dense-U-Net model in this work is powerful for extracting roads from high-resolution remote sensing imagery. The network extracts the road features via the DenseNet in the contracting part, which shows drastic improvements over previous models. Because the original remote sensing image must to be clipped to train the deep neural convolutional network, this paper designed the LAU and GAU to obtain the local and global road information, while combining multiple scales in different blocks in the expansive part. All of the improvements helped to recover the road by the extracted information and to classify roads of different sizes.

To evaluate the effectiveness of the proposed GL-Dense-U-Net model on road extraction from remote sensing imagery, this work was compared with some state-of-the-art methods, including the deep learning methods like FCN [42] and U-Net [33]. At the same time, the newest deep convolutional neural network DeepLab V3+ [55] was used to compare the proposed GL-Dense-U-Net model with. To make the comparisons objective and fair, all the of the methods mentioned were tested with the same set of images. The comparison results are exhibited in Table 2, where the best values of each column were in bold font.

To show the improvements to individual images and the whole dataset, this work chose three sampled images in the first three columns and the all the test dataset in the last column in Table 2. The precision, recall and F_1 score were calculated by state-of-the-art methods. From comparing results of different methods (Table 2) we can see that the U-Net model (based on FCN), achieve higher recall values. However, their comprehensive performances are unsatisfying and obtain a lower F_1 score.

Table 2. Comparison of the results of the proposed method with other methods, where the values in bold are the best.

	Image 1			Image 2			Image 3			ALL		
	P	***R***	F_1	***P***	***R***	F_1	***P***	***R***	F_1	***P***	***R***	F_1
FCN [42]	0.8027	0.9397	0.8658	0.8437	0.9624	0.8991	0.9756	0.8459	0.9061	0.8478	0.9307	0.8873
U-Net [33]	0.8303	**0.9848**	0.9010	0.8432	**0.9942**	0.9125	0.7953	**0.9696**	0.8738	0.8326	**0.9708**	0.8964
DeepLab V3+ [55]	0.9287	0.9211	0.9249	0.9552	0.9255	0.9401	0.9407	0.9458	0.9432	0.9415	0.9308	0.9361
Ours	**0.9516**	0.9537	**0.9527**	**0.9797**	0.9420	**0.9605**	**0.9576**	0.9589	**0.9582**	**0.9630**	0.9515	**0.9572**

Where *P* stands for precision, *R* represents recall and F_1 score for the roads extraction, respectively.

U-Net model is robust to occlusions for convolutional operations after up-sampling in the expansive part of the model. Though this method can achieve satisfactory results and improved recall results, some false positives are introduced in some areas, such as the turning of the roads (the region B in Figure 9), so that the comprehensive performances are unsatisfying. FCN and DeepLab V3+ are sensitive to noise, roads in some regions like the shadows of trees (the region A in Figure 9) cannot be extracted accurately. Benefited from the global and local attention units, the proposed GL-Dense-U-Net model achieved relatively satisfactory performances in recall and the best values both in precision and F_1 score, therefore, the designed model generally achieves the best result. Specifically, the F_1 score of our designed model is 2.11% higher than the next best compared method (DeepLab V3+ [55]). The proposed GL-Dense-U-Net model obtained the highest F_1 score and precision in all the separated samples mentioned. All of them demonstrate that the combination of the proposed attention units and the use of DenseNet as the feature extractor in the GL-Dense-U-Net model helped to achieve better performance than other state-of-the-art methods. At the same time, the comparison validated this work in terms of road network extraction from high-resolution remote sensing imagery.

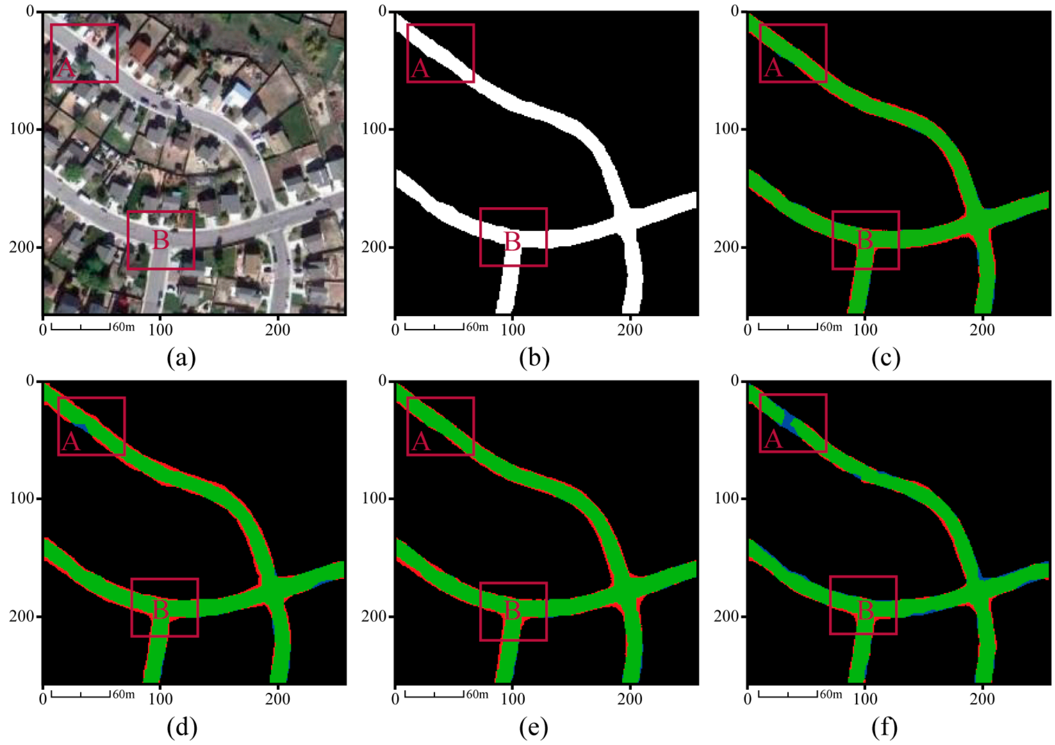

Figure 9. Visual comparisons of road extraction results with different comparing algorithms. (**a**) The original remote sensing image. (**b**) The corresponding ground truth image of this region. (**c**) The results using the GL-Dense-U-Net model. (**d**) Results of FCN. (**e**) Results of U-Net. (**f**) Results of DeepLab V3+. Green, red and blue represents the TP, FP and FN, respectively.

4. Discussion

Local and global road information play important roles for extracting roads from the complex remote sensing imagery. However, most of the road detection methods are limited for information. To improve the performance of road extraction, this work proposes two modules: the local and global attention units, which help to improve the ability to extract local and global information, respectively. Profiting from these powerful units, the precision, recall and the F_1 scores have been significantly improved. To illustrate the effect of the two units, this work compared the results of the proposed model while excluding either the LAU or the GAU; in both cases, the same training dataset was used. All the assessed metrics are shown in Table 3:

Table 3. Comparison of the results while excluding either the LAU or the GAU, where the values in bold are the best.

Elements	OA	Precision (R)	Recall (R)	F_1 (R)	Precision (C)	Recall (C)	F_1 (C)
LAU and GAU	**0.9782**	**0.9630**	0.9515	**0.9572**	0.9936	**0.9956**	**0.9946**
Only LAU	0.9750	0.9565	0.9504	0.9534	0.9941	0.9938	0.9940
Only GAU	0.9699	0.9458	**0.9543**	0.9500	**0.9942**	0.9933	0.9938
Without both	0.9561	0.9385	0.9409	0.9397	0.9914	0.9921	0.9917

Where R stand for buildings and C represent clutter, and OA means overall accuracy.

The OA improved by 1.89% and by 1.38% when the designed model only including the LAU and GAU, respectively. At the same time, the F_1 score for road extraction improved by 1.37% and 1.03% for the models using only the LAU or GAU, respectively. When both the LAU and GAU were designed in the deep convolutional neural network, the OA and F_1 scores did not improve by the addition of the aforementioned results, but did improves slightly over the single attention unit models. The comparison illustrates that some road information in the original remote sensing imagery can be extracted by both of the proposed units during model training, which proves that both of the proposed units have the ability to extract features at different scales.

Global information plays an important role in extracting the roads from remote sensing imagery, because the roads' structure can run through almost all of the images. GAP operation followed by a 1×1 convolution and deconvolution were designed in the GAU, which can extract global information and guide feature recovery. Therefore, the proposed GAU helps the model by improving all the metrics of road extraction. The roads in the remote sensing imagery are complex, which makes it easy to miss the local information during road extraction. The LAU is designed using the pyramid structure to fuse different scales feature-maps and draw attention to the pixel-level information, aids the model in road extraction, as shown in Figure 10. The results show that the performance is poor in some local structures like A, C (which includes a tree shadow) and B (where the spectral characteristics are similar to the road), when the LAU is excluded from the designed model. It is clear that the designed GL-Dense-U-Net model is powerful enough to extract the local information of the road, and the pixels were classified into different groups correctly.

As remote sensing technology develops, more satellites are being launched, which make it easier to access high-resolution remote sensing imagery. Semantic segmentation of remote sensing imagery plays an important role in practical applications, such as navigation and urban planning. This work designed a new model, which improved the image classification performance for road extraction. The trained model can be used to extract roads from other remote sensing imagery datasets directly or just fine-tuning. On the other hand, the designed model can be retrained by the datasets of different land uses and then extract their relative information. The model is adapted to imagery with more than 3 bands by modifying the input channels in the first convolutional layer. However, the proposed method is a novel supervised learning framework, and the dataset used in the experiment does not contain the unpaved roads. Therefore, the trained model is invalid for extracting the unpaved roads. The extraction results cannot directly function as vector data for navigation. In the future, a more optimized deep neural network is required to extract the road to a vector format. At the same time, benefited from the performance of this work in extracting local and global information, the semantic segmentation results can be used as additional information, like the road width, for the extracted roads vector data to improve the development of transportation systems. Therefore, future work will focus on how to generate reliable vector road networks using remote sensing imagery with more information, which can be directly used for practical applications.

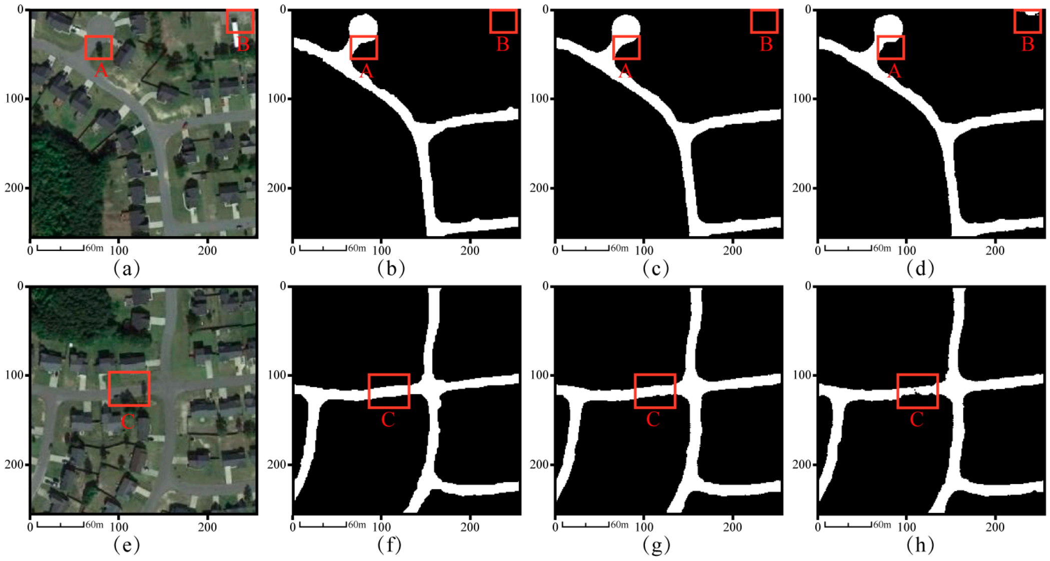

Figure 10. Results of road extraction from the remote sensing imagery with and without LAU. (**a**,**e**) The original remote sensing image. (**b**,**f**) The corresponding ground truth image of this region. (**c**,**g**) The prediction results using the GL-Dense-U-Net model. (**d**,**h**) The prediction results without LAU.

5. Conclusions

In this paper, a novel deep convolutional neural network was presented to perform road extraction from high-resolution remote sensing imagery. The major contribution of this work is the designed road extraction model based on the construction of U-Net and the introduction of the DenseNet as the feature extractor. At the same time, this work designed the local attention unit and global attention unit in the expansive part of the model, which improved the precision and F_1 score. The proposed model aimed to extract the local and global information of roads in the remote sensing imagery and improve the accuracy of road network extraction. The proposed GL-Dense-U-Net model did well in labeling different scale roads in the remote sensing imagery because the DenseNet blocks in different stages were used to guide the feature recovery in the expansive part. Experiments were carried out on a high-resolution remote sensing imagery dataset. The roads were extracted successfully via the deep convolutional neural network proposed in this work, and the results showed the effectiveness and feasibility of the proposed framework in improving the performance of semantic segmentation of remote sensing imagery. Qualitative comparisons were performed with some state-of-the-art methods for semantic segmentation, such as the fully convolutional network (FCN), the U-Net, as well as the new DeepLab V3+ method. Experimental results demonstrated that the proposed model performed better than the other methods. The proposed method in this work can obtain improvements in terms of the comprehensive evaluation metric, the F_1 score, over the aforementioned semantic segmentation systems.

Author Contributions: Y.X. and Z.C. proposed the network architecture design and the framework of extracting roads. Y.X. and Y.F. performed the experiments and analyzed the data. Y.X. and Z.X. wrote the paper. Y.F. revised the paper and provided valuable advices for the experiments.

Funding: This study was financially supported by the National key R & D program of China (No. 2017YFB0503600, 2017YFC0602204, 2018YFB0505500), National Natural Science Foundation of China (41671400), HuBei Natural Science Foundation of China (2015CFA012) and Fundamental Research Funds for National Universities China University of Geosciences (Wuhan).

Acknowledgments: The authors thank Guangliang Cheng providing datasets. The authors also thank Mingyu Xie (University of California, Santa Barbara) for helping improve the language.

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

References

1. Weng, Q. Remote sensing of impervious surfaces in the urban areas: Requirements, methods, and trends. *Remote Sens. Environ.* **2012**, *117*, 34–49. [CrossRef]
2. Peng, T.; Jermyn, I.H.; Prinet, V.; Zerubia, J. Incorporating generic and specific prior knowledge in a multiscale phase field model for road extraction from VHR images. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2008**, *1*, 139–146. [CrossRef]
3. Xu, Y.; Wu, L.; Xie, Z.; Chen, Z. Building Extraction in Very High Resolution Remote Sensing Imagery Using Deep Learning and Guided Filters. *Remote Sens.* **2018**, *10*, 144. [CrossRef]
4. Zhang, Z.; Liu, Q.; Wang, Y. Road Extraction by Deep Residual U-Net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
5. Zhu, C.; Shi, W.; Pesaresi, M.; Liu, L.; Chen, X.; King, B. The recognition of road network from high—Resolution satellite remotely sensed data using image morphological characteristics. *Int. J. Remote Sens.* **2005**, *26*, 5493–5508. [CrossRef]
6. 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]
7. 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]
8. Zhang, J.; Chen, L.; Wang, C.; Zhuo, L.; Tian, Q.; Liang, X. Road Recognition From Remote Sensing Imagery Using Incremental Learning. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 2993–3005. [CrossRef]
9. 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 Observ. Remote Sens.* **2016**, *9*, 1946–1958. [CrossRef]
10. Cheng, J.; Ding, W.; Ku, X.; Sun, J. Road Extraction from High-Resolution SAR Images via Automatic Local Detecting and Human-Guided Global Tracking. *Int. J. Antenn. Propag.* **2012**, *2012*, 989823. [CrossRef]
11. Li, G.; An, J.; Chen, C. Automatic Road Extraction from High-Resolution Remote Sensing Image Based on Bat Model and Mutual Information Matching. *JCP* **2011**, *6*, 2417–2426. [CrossRef]
12. Miao, Z.; Wang, B.; Shi, W.; Zhang, H. A semi-automatic method for road centerline extraction from VHR images. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 1856–1860. [CrossRef]
13. Unsalan, C.; Sirmacek, B. Road network detection using probabilistic and graph theoretical methods. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 4441–4453. [CrossRef]
14. Al-Khudhairy, D.; Caravaggi, I.; Giada, S. Structural damage assessments from Ikonos data using change detection, object-oriented segmentation, and classification techniques. *Photogramm. Eng. Remote Sens.* **2005**, *71*, 825–837. [CrossRef]
15. 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]
16. Tian, S.; Zhang, X.; Tian, J.; Sun, Q. Random forest classification of wetland landcovers from multi-sensor data in the arid region of Xinjiang, China. *Remote Sens.* **2016**, *8*, 954. [CrossRef]
17. Anwer, R.M.; Khan, F.S.; van de Weijer, J.; Molinier, M.; Laaksonen, J. Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification. *ISPRS J. Photogramm. Remote Sens.* **2018**, *138*, 74–85. [CrossRef]

18. Yager, N.; Sowmya, A. Support vector machines for road extraction from remotely sensed images. In Proceedings of the International Conference on Computer Analysis of Images and Patterns, Groningen, The Netherlands, 25–27 August 2003; pp. 285–292.
19. Simler, C. An improved road and building detector on VHR images. In Proceedings of the 2011 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Vancouver, BC, Canada, 24–29 July 2011; pp. 507–510.
20. Yousif, O.; Ban, Y. Improving SAR-based urban change detection by combining MAP-MRF classifier and nonlocal means similarity weights. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2014**, *7*, 4288–4300. [CrossRef]
21. Zhu, D.-M.; Wen, X.; Ling, C.-L. Road extraction based on the algorithms of MRF and hybrid model of SVM and FCM. In Proceedings of the 2011 International Symposium on Image and Data Fusion (ISIDF), Taiyuan, China, 3–6 August 2011; pp. 1–4.
22. Razavian, A.S.; Azizpour, H.; Sullivan, J.; Carlsson, S. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. *arXiv*, 2014; arXiv:1403.6382.
23. Xu, Y.; Chen, Z.; Xie, Z.; Wu, L. Quality assessment of building footprint data using a deep autoencoder network. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 1929–1951. [CrossRef]
24. 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.
25. Mnih, V.; Hinton, G.E. Learning to detect roads in high-resolution aerial images. In Proceedings of the European Conference on Computer Vision, San Francisco, CA, USA, 13–18 June 2010; pp. 210–223.
26. Zhang, L.; Yang, F.; Zhang, Y.D.; Zhu, Y.J. Road crack detection using deep convolutional neural network. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 3708–3712.
27. Kampffmeyer, M.; Salberg, A.-B.; Jenssen, R. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 27–30 June 2016.
28. Zhong, Z.; Li, J.; Cui, W.; Jiang, H. Fully convolutional networks for building and road extraction: Preliminary results. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 1591–1594.
29. Fu, G.; Liu, C.; Zhou, R.; Sun, T.; Zhang, Q. Classification for High Resolution Remote Sensing Imagery Using a Fully Convolutional Network. *Remote Sens.* **2017**, *9*, 498. [CrossRef]
30. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Convolutional Neural Networks for Large-Scale Remote-Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 645–657. [CrossRef]
31. Kestur, R.; Farooq, S.; Abdal, R.; Mehraj, E.; Narasipura, O.; Mudigere, M. UFCN: A fully convolutional neural network for road extraction in RGB imagery acquired by remote sensing from an unmanned aerial vehicle. *J. Appl. Remote Sens.* **2018**, *12*, 016020. [CrossRef]
32. 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), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
33. 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.
34. Luong, M.-T.; Pham, H.; Manning, C.D. Effective approaches to attention-based neural machine translation. *arXiv* **2015**, arXiv:1508.04025.
35. Shang, L.; Lu, Z.; Li, H. Neural responding machine for short-text conversation. *arXiv* **2015**, arXiv:1503.02364.
36. Cui, Y.; Chen, Z.; Wei, S.; Wang, S.; Liu, T.; Hu, G. Attention-over-attention neural networks for reading comprehension. *arXiv* **2016**, arXiv:1607.04423.
37. Huang, G.; Liu, Z.; Weinberger, K.Q.; van der Maaten, L. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.

38. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv*, 2015; arXiv:1502.03167.
39. Glorot, X.; Bordes, A.; Bengio, Y. Deep sparse rectifier neural networks. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Ft. Lauderdale, FL, USA, 11–13 April 2011; pp. 315–323.
40. Yann, L.; Léon, B.; Yoshua, B.; Patrick, H. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324.
41. Zeiler, M.D.; Taylor, G.W.; Fergus, R. Adaptive deconvolutional networks for mid and high level feature learning. In Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 2018–2025.
42. 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. 1337–1342.
43. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. *arXiv*, 2017; arXiv:1709.01507.
44. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context encoding for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
45. Li, H.; Xiong, P.; An, J.; Wang, L. Pyramid Attention Network for Semantic Segmentation. *arXiv* **2018**, arXiv:1805.10180.
46. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
47. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv*, 2017; arXiv:1706.05587.
48. 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]
49. Google. Google Earth. Available online: http://www.google.cn/intl/zh-CN/earth/ (accessed on 12 September 2015).
50. Kingma, D.P.; Ba, J. ADAM: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
51. Hwang, J.-J.; Liu, T.-L. Pixel-wise deep learning for contour detection. *arXiv* **2015**, arXiv:1504.01989.
52. Heipke, C.; Mayer, H.; Wiedemann, C.; Jamet, O. Evaluation of automatic road extraction. *Int. Arch. Photogramm. Remote Sens.* **1997**, *32*, 151–160.
53. Martin, D.R.; Fowlkes, C.C.; Malik, J. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Trans. Pattern Anal. Mach. Intell.* **2004**, *26*, 530–549. [CrossRef] [PubMed]
54. Wang, H.; Wang, Y.; Zhang, Q.; Xiang, S.; Pan, C. Gated Convolutional Neural Network for Semantic Segmentation in High-Resolution Images. *Remote Sens.* **2017**, *9*, 446. [CrossRef]
55. 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.