International Journal of Geo-Information

MDPI

Article

Cascaded Attention DenseUNet (CADUNet) for Road Extraction from Very-High-Resolution Images

Jing Li [1], Yong Liu [1,*], Yindan Zhang [1] and Yang Zhang [2]

[1] College of Earth and Environmental Sciences, Lanzhou University, Lanzhou 730000, China; jingli2018@lzu.edu.cn (J.L.); zhangyd15@lzu.edu.cn (Y.Z.)
[2] Supercomputing Center, Lanzhou University, Lanzhou 730000, China; zhyang@lzu.edu.cn
* Correspondence: liuy@lzu.edu.cn

Abstract: The use of very-high-resolution images to extract urban, suburban and rural roads has important application value. However, it is still a problem to effectively extract the road area occluded by roadside tree canopy or high-rise buildings to maintain the integrity of the extracted road area, the smoothness of the sideline and the connectivity of the road network. This paper proposes an innovative Cascaded Attention DenseUNet (CADUNet) semantic segmentation model by embedding two attention modules, such as global attention and core attention modules, in the DenseUNet framework. First, a set of cascaded global attention modules are introduced to obtain the contextual information of the road; secondly, a set of cascaded core attention modules are embedded to ensure that the road information is transmitted to the greatest extent among the dense blocks in the network, and further assist the global attention module in acquiring multi-scale road information, thereby improving the connectivity of the road network while restoring the integrity of the road area shaded by the tree canopy and high-rise buildings. Based on binary cross entropy, an adaptive loss function is proposed for network parameter tuning. Experiments on the Massachusetts road dataset and the DeepGlobe-CVPR 2018 road dataset show that this semantic segmentation model can effectively extract the road area shaded by tree canopy and improve the connectivity of the road network.

Keywords: deep learning; road; DenseUNet; attention module; semantic segmentation; remote sensing

Citation: Li, J.; Liu, Y.; Zhang, Y.; Zhang, Y. Cascaded Attention DenseUNet (CADUNet) for Road Extraction from Very-High-Resolution Images. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 329. https://doi.org/10.3390/ijgi10050329

Academic Editors: Wolfgang Kainz and Sébastien Lefèvre

Received: 1 March 2021
Accepted: 2 May 2021
Published: 13 May 2021

1. Introduction

Road information is of vital importance in the fields of urban and rural development [1], emergency and disaster relief [2], vehicle navigation [3] and geographic information systems [4]. With the rapid development of remote sensing technology, very-high-resolution (VHR) remote sensing images have been used for extracting road information [5]. In practice, most road data updates still use manual interpretation, which is time-consuming and laborious and lacks quality control. Many road extraction algorithms have been developed [6–8]. These algorithms can be divided into traditional machine learning methods [9–14] and the latest deep learning methods [15–17]. Some traditional road extraction methods mainly use the spectral features of remote sensing images, occasionally supplemented by texture features. However, this method is difficult to effectively use the geometric and context information in remote sensing images [18], and it is easy to produce "salt and pepper" noise [19]. Among the traditional methods, the object-based approach obviously improves the effect on road extraction. Instead of pixels, it uses image objects as the basic unit, utilizing their spectral, geometric, textural and contextual features for information extraction, thereby improving product quality [20,21]. On the one hand, this method is highly dependent on the quality of image segmentation, and how to find suitable parameters for segmentation is itself a difficult problem. On the other hand, there are many spectral, texture, geometric and contextual features, and it is difficult to determine which features are most suitable for road information extraction. When the data source or regional conditions change, the features required for classification need to be adjusted

accordingly [22–24]. The above methods are also difficult to distinguish roads from other artificial surfaces, such as buildings and parking lots, in VHR images.

In recent years, deep learning has been introduced into road extraction from remote sensing images [25]. Methods based on deep learning reveal effective feature expression capabilities and can automatically acquire useful features from images for road extraction [26,27]. The deep learning algorithms used in road extraction are mainly based on convolutional neural networks (CNN). Zhang et al. [28] used CNN for extracting road information in VHR images. By improving the CNN architecture, Gao et al. [17] proposed a deep residual convolutional neural network with post-processing operations, which showed good performance in extracting roads from complex backgrounds involving both urban and rural areas. Fully convolutional network (FCN) is a kind of CNN. Long et al. [29] first proposed a semantic segmentation model based on FCN to road extraction. Later, some new semantic segmentation models were developed on the basis of FCN, such as U-Net [30], SegNet [31], DeepLab V3+ [32], etc., which are used in road extraction from VHR images. Zhang et al. [33] proposed deep residual U-Net for road extraction. This method reduces information loss and effectively improves the accuracy of road extraction. Buslaev et al. [15] proposed an improved U-Net that can shorten the training time and achieved a good result in the CVPR 2018 challenge. The patch-based approaches, like the object-based approach, assign a single label for all pixels within a patch. Mohammad et al. [34] proposed a patch-based deep neural network to detect roads in large-scale datasets. Some studies [23,35] constructed cascaded neural networks to perform multitask learning to simultaneously extract road areas, centerlines and sidelines. To effectively solve the problem of the canopy shading effect and maintain the connectivity of the extracted road network, Tao et al. [36] proposed a spatial information reasoning network to capture and transmit road-specific contextual information. Noticing the low computational efficiency of the D-LinkNet, Li et al. [37] established an improved B-D-Linknet Plus. Experiments show that the improved neural network can reduce the network size and improve the accuracy required for road extraction. Yang et al. [38] designed a recursive convolutional neural network (RCNN) module and integrated it into the U-Net architecture to solve problems such as noise, occlusion and complex background. To preserve boundary information and obtain high-resolution road maps, Abolfazl et al. [39] introduced a new convolutional network, namely the VNet model. Xin et al. [40] proposed DenseUNet for road extraction in complex scenes based on DenseNet, considering its powerful capabilities on multi-level feature extraction and reuse. These deep learning methods perform well, so that roads and buildings and other artificial surfaces are better classified.

Road extraction from VHR images usually faces two difficulties. The first is to maintain the integrity of the road area and the smoothness of the sideline. The large tree canopies and high-rise buildings on the roadside have a shadow effect, often occluding the road area. The second is to maintain the connectivity of the road network so that the road is not missing or interrupted. Some studies have tried to solve the problem of partial shading, which leads to a lack of road area, and complete occluding leads to road interruption. One method is to post-process the segmentation results [17,41,42]. However, the problem is that parameters for post-processing must be manually set and the operation is complicated and difficult for long-distance complex roads. In recent years, the attention mechanism has been introduced into the deep learning model. Lai et al. [43] designed a visual attention unit to locate the focus area more accurately for image fusion. The attention mechanism can improve the effectiveness of the model in target detection [44,45]. With the help of the attention mechanism, deep learning networks can extract more discriminative features for the target task [46–48]. Ye et al. [49] used the attention mechanism to solve the problem of skipping connections for building extraction. Jetley et al. [48] used an attention gating module to generate a contextual attention map at the high level of the network, focusing on the local information useful for middle-level prediction in the form of "global guidance". Jin et al. [50] used the time attention mechanism to adjust the nonlinearity and dynamic adaptability of the electrical network, thereby improving the overall performance of the

prediction model. Oktay et al. [51] proposed an attention module that learns weighted images from a high level to focus on the useful features and suppresses the irrelevant regions in the intermediate feature map, thereby improving the prediction performance.

In response to the problems, we propose an innovative Cascaded Attention DenseUNet (CADUNet) by imbedding two attention modules, such as global attention and core attention, into the DenseUNet framework. We use the core attention module to extract road areas, including the occluded parts, and use the global attention module to enhance global context information about the road network. The main contributions of this article are as follows:

1. The core attention modules and the global attention modules are cascaded in the DenseUNet together to combine road information at different scales, thus improving the connectivity of the road network and the smoothness of the sidelines.
2. An adaptive loss function is introduced to solve the problem of too-small ratio of roads to non-road areas in the training samples.

The rest of the paper is structured as follows: In Section 2, we introduce the CADUNet method. Section 3 specifies data preparation used in the experiment. Section 4 shows the results. Section 5 explores the mechanisms for the effectiveness of the network model and Section 6 provides the conclusions.

2. Methods

The proposed CADUNet is a composite semantic segmentation network established by imbedding global and core attention modules into the DenseUNet framework. The DenseUNet is an integration of two classical networks of UNet and DenseNet [52]. UNet usually consists of two parts: encoder and decoder. DenseUNet normally consists of dense blocks and transition down layers associated with UNet. When making the DenseUNet, the dense block and transition down layers are inserted into the encoder part of UNet to replace the original convolutional layers and pooling layers, thus improving the performance of UNet in semantic segmentation [40,45]. In the CADUNet, global attention modules are further added to the decoder part of UNet (Figure 1). In addition, core attention modules are embedded between the encoder and decoder. To obtain better results, it is necessary to obtain high-level semantic information from images while retaining the low-level detailed information. The information from the lower layers can be transferred to the higher layers along the information transmission path. This compensates for the details of the low-level function and high-level semantic information [44]. The following subsections provide the details.

2.1. Encoder

We use dense blocks and transition down layers in the encoder part of UNet. The dense block is composed of four dense layers (Figure 2), and the output of each dense layer has a feature map of the same channel dimension. In each dense block, all layers maintain dense connections. Dense blocks are connected through transition down layers between them. In a single dense block, the function $F_l(\,)$ is used for nonlinear conversion between layers. The dense connection is defined as Equation (1) [52]:

$$D_l = F_l([D_0, D_1, D_2, \ldots D_{l-1}]) \tag{1}$$

where l is the number of dense layers in each dense block, D_1 is the output feature map of the first layer and $[D_0, D_1, D_2, \ldots, D_{l-1}]$ is a cascade of all previous feature maps of the first layer.

Considering that DenseNet will generate too many feature maps, associated with too many model parameters, we define a growth rate K to control the number of feature maps, where K represents the number of feature layers output by each layer. We set K to 48. It is the same as the size of the feature maps inside each dense block (Figure 2).

To reduce the amount of calculation and increase the receptive field, a down transition layer is used after each dense block. Each transition layer is composed of batch normalization (BN), rectified linear unit (ReLU), bottleneck layer (1 × 1 convolution) and average pooling layer (2 × 2).

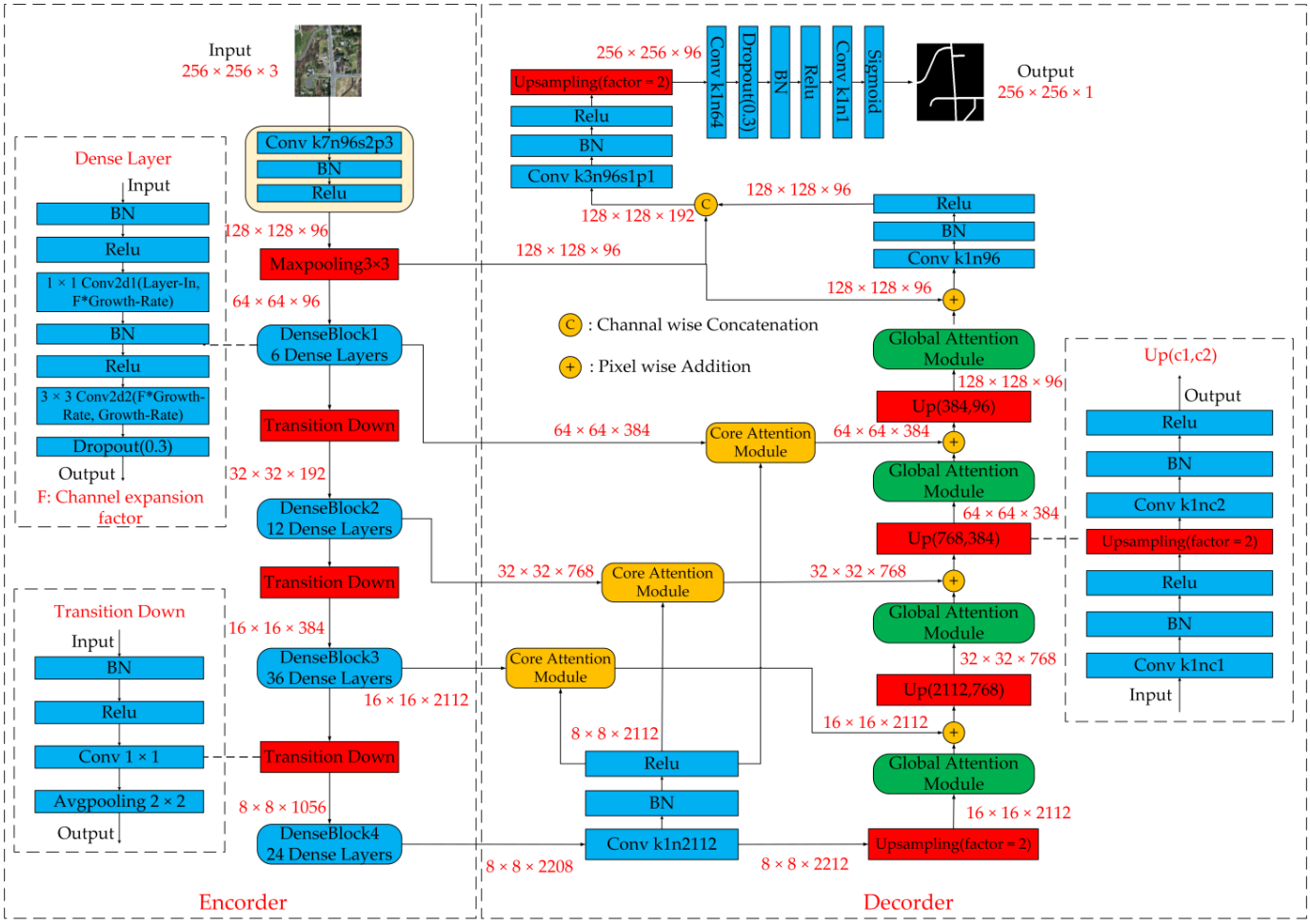

Figure 1. Architecture of CADUNet (The parameters include: *k*, the kernel size; *n*, the number of output channels; *s*, the stride size; *p*, the padding size).

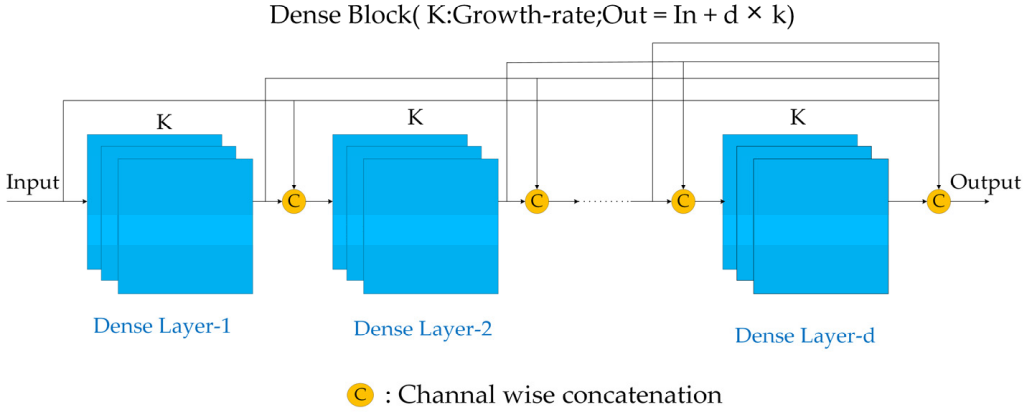

Figure 2. Structure of dense block.

2.2. Attention Mechanism

The attention mechanism can help to focus more attention on interesting targets [44,45]. This study uses two attention modules: core attention module [44] and global attention module [45]. In the core attention module, the input value of the signal is calculated by calculating the output of the last dense block (Figure 3). The core attention module contains two inputs, one is the output to the three dense blocks, and the other is the attention signal input. By connecting the low-level features to the high-level features, the core attention module can weaken the background information and enhance useful local details, thereby reducing the misjudgment of the original jump connection feature and improving the integrity of the extracted road network. The introduction of the core attention module, on the one hand, can ensure the maximum transmission of road information between all layers of the network. On the other hand, it can assist the global attention module to improve the integrity of the road while eliminating the tree canopy occlusion effect.

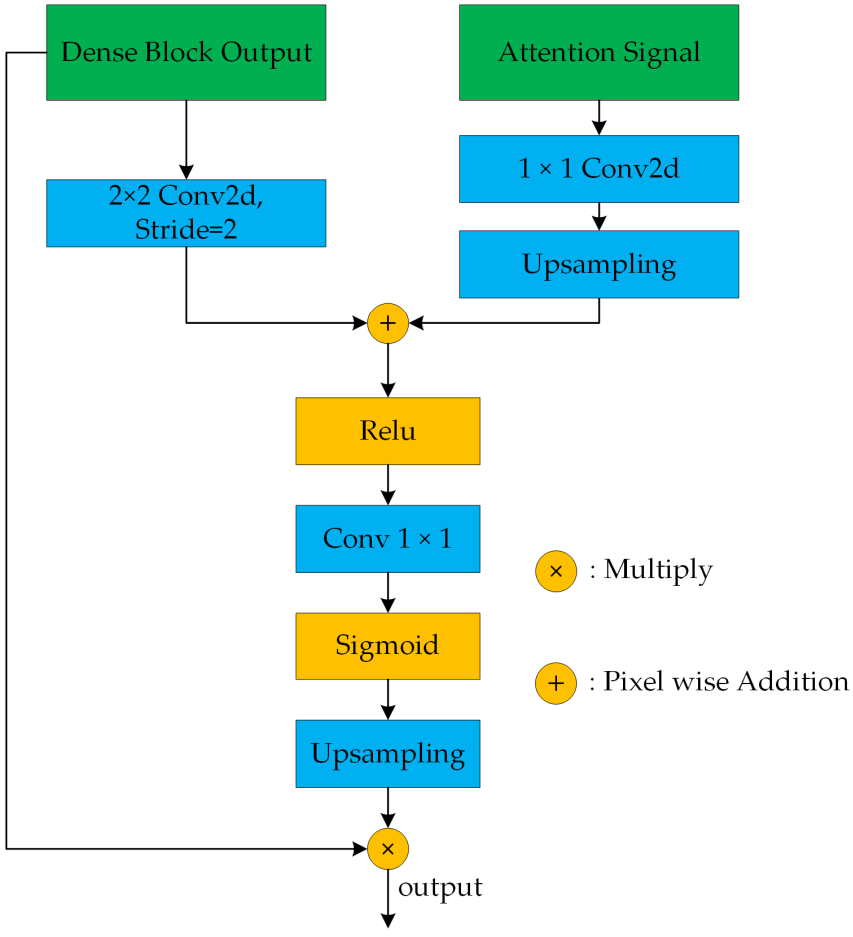

Figure 3. Core attention module.

In the global attention module, the global average pool is first used to extract global context information from the high-level feature map (Figure 4). The global average pool is convenient to obtain global context information in images [45]. Then, the output of global context information is activated through a sigmoid function. Finally, weighted features are added to the feature map to integrate global information. The global attention module uses the global average layer to collect the global context information from the feature map and enhances the global information of the feature map, thereby solving the interruption of road extraction caused by tree canopy occlusion.

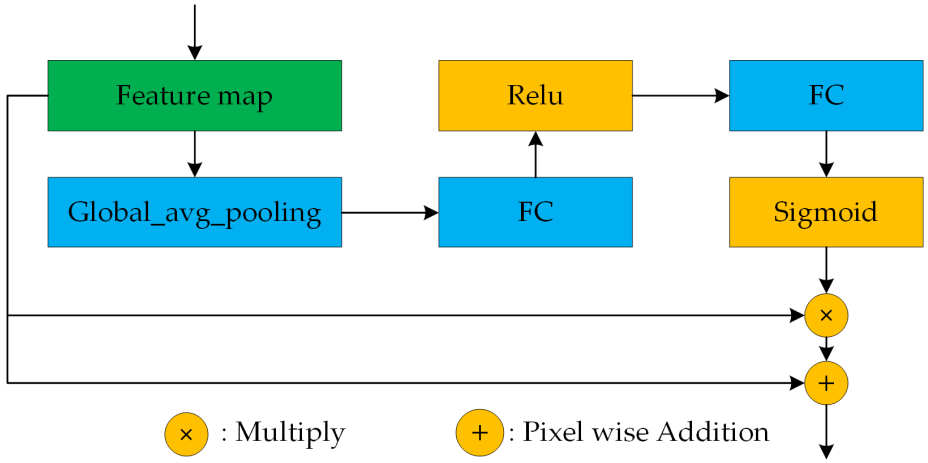

Figure 4. Global attention module (FC means a fully connected layer).

2.3. Decoder

We mainly made two adjustments to the decoder in CADUNet. One is to use a simple up-sampling operation with a step size of 2 in the first layer, and the second is to use 4 global attention modules plus 3 improved up-sampling operations. In the improved up-sampling operations, 1×1 convolution, BN and ReLU operations are performed first, followed by 3×3 convolution operations, BN and ReLU operations, and finally, simple up-sampling. This matches the size of the resulting output by the attention module. We add the output of the last global attention module to the corresponding layer in the encoder. After that, the output relates to the corresponding layer in the encoder. Then, a simple up-sampling operation is added to restore the size of the image to the same as the original input image following 1×1 convolution, BN and ReLU operations. For the final convolution, BN, ReLU and sigmoid operations are used to generate the predicted road map.

2.4. Adaptive Loss Function

In this paper, we consider road extraction as a binary semantic segmentation. The proportion of road area is usually less than 10%, and the proportion of non-road backgrounds is usually greater than 90%. In the case of random sampling, the training efficiency is low since negative samples occupy most of the training samples [24]. To this end, we adopt a new adaptive loss function to adjust the imbalance between positive and negative samples:

$$Loss = P_{road} \times L_{BCE} + P_{background} \times (1 - L_{IoU}) \tag{2}$$

where, P_{road} and $P_{background}$ respectively represent the percentage of roads and non-roads in the entire area. L_{BCE} is the binary cross entropy loss [53], and L_{IoU} is the intersection ratio index [54] and emphasizes the deviation between the predicted road and the actual road. The calculation formula of each is as follows:

$$L_{BCE} = -\frac{1}{n}\sum_{i=1}^{n}(g_i(\log p_i) + (1 - g_i)(1 - \log p_i)) \tag{3}$$

$$L_{IoU} = 1 - \frac{\sum_{i=1}^{n} g_i p_i}{\sum_{i=1}^{n}(g_i + p_i - g_i p_i)} \tag{4}$$

where g_i ($i = 0, 1, 2, \ldots, n$) is the ground truth of the i-th pixel, p_i ($i = 0, 1, 2, \ldots, n$) is the predictions of the i-th pixel and n is the number of pixels.

3. Experiment Preparation

The datasets used in this study are from the Massachusetts road dataset and DeepGLOBE-CVPR 2018 road dataset (CVPR dataset) [55,56]. They are composed of

image datasets for training, validation and test, associated with corresponding reference maps. The Massachusetts road dataset contains a total of 1171 images. Each image in this dataset is 1500 × 1500 pixels, with a spatial resolution of 1.2 m and a coverage area of 2.25 square kilometers. The dataset covers a variety of typical urban, suburban and rural areas, with a total area of more than 2600 square kilometers. The CVPR road dataset contains 6226 satellite images with a size of 1024 × 1024 pixels and a spatial resolution of 50 cm. Accordingly, these datasets can be divided into rural, suburban and urban road datasets, as shown in Figure 5.

To make training, validation and test datasets for this experiment, all image datasets were cropped and augmented. First, the images and the corresponding reference maps were expended by random rotation (90 degrees, 180 degrees and 270 degrees), random horizontal and vertical flips and random brightness adjustment (0.5–1.5). Then, they were randomly cropped to 256 × 256 pixels [36]. Finally, from the Massachusetts dataset, we obtained 50,545 images, of which 42,963 were for training and 7582 were for validating, and the test dataset is 49 original 1500 × 1500 images. From the CVPR road dataset, 84,000 images were obtained, of which 71,400 were for training, 12,600 images were for validating and the test dataset is 105 original 1024 × 1024 images.

We compare this method with UNet [30], DeepLab v3+ [32], DenseUNet [40], the improved DenseUNet (CDenseUNet) with only the core attention modules and the improved DenseUNet (GDenseUNet) with only the global attention modules.

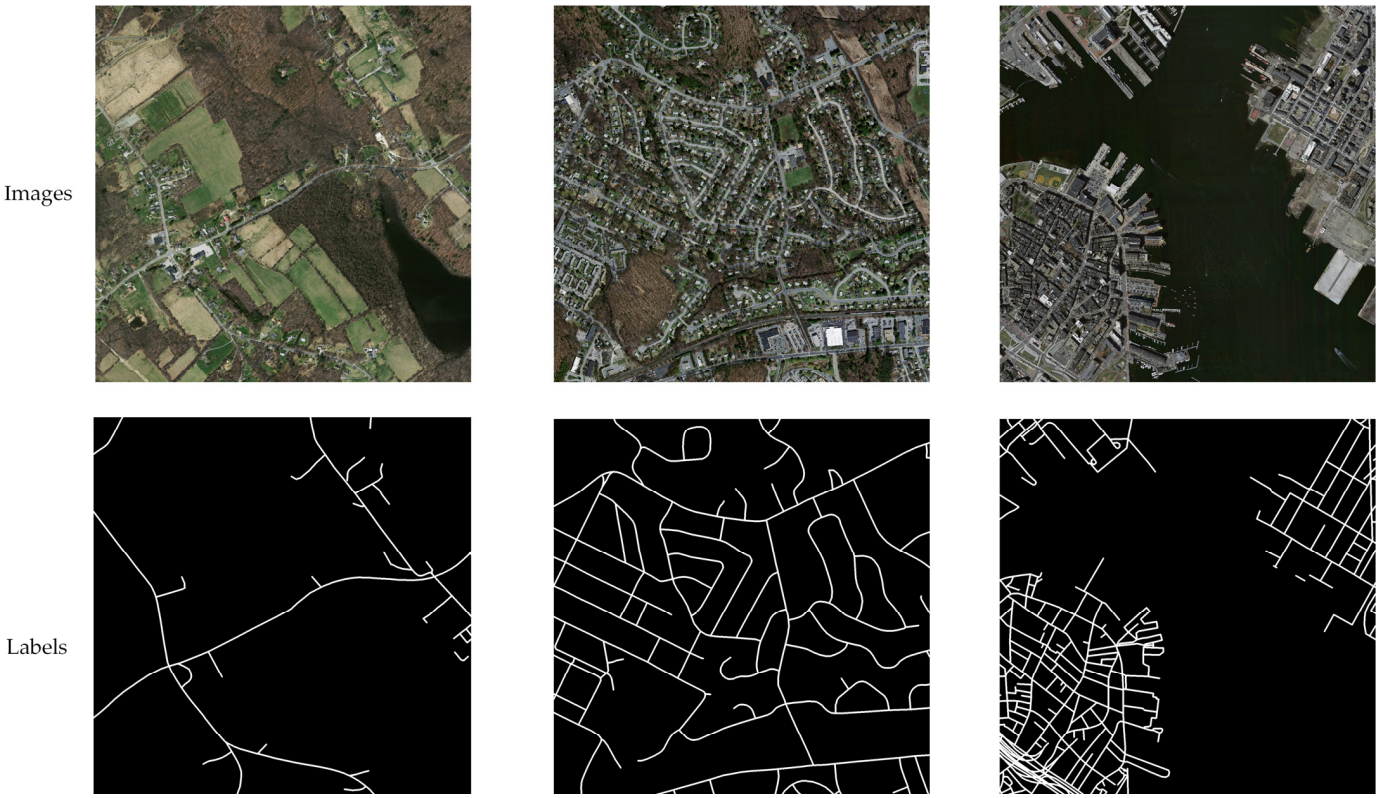

Figure 5. Typical images with roads in rural, suburban and urban areas in the Massachusetts roads dataset. (**a**) The rural roads; (**b**) The suburban roads; (**c**) The urban roads.

This experiment is implemented on a high-performance computing platform: the CPU is composed of 2 groups of Intel Xeon 5120 with 14 cores, associated with 128 GB of working memory, the GPU is 2 groups of NVIDIA P100 with 16 GB of memory and the operating system uses CentOS 7. We used the TensorFlow backend to execute on the deep learning framework of Keras. The Adam function [57] is used for parameter optimization.

Each epoch processed 16 images. The learning rate was initially set to 0.0001, and was reduced by 0.02 times per period, and the number of epochs was set to 50.

In this experiment, we use overall accuracy (*OA*), *precision*, *recall*, $F_{1-score}$, and Intersection over Union (*IoU*) for validation. Equations (5)–(9) [36,54,58] describe these assessment metrics:

$$OA = \frac{TP + TN}{TP + FP + FN + TN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F_{1-score} = \frac{2TP}{2TP + FP + FN} \tag{8}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{9}$$

where, *TP*, *FP*, *FN* and *TN* represent true positive, false positive, false negative and true negative, respectively.

4. Results

4.1. Pavement Integrity and Sideline Smoothness of Roads

4.1.1. Massachusetts Dataset

In the Massachusetts dataset, the road occluding mainly comes from the tree canopy aside the rural and suburban roads, while the images occluded by urban roads are few. Figure 6 shows the partial roads occluded by tree canopies in rural areas (scenes 1–3), the partial roads occluded by tree canopies in suburbs (scenes 4–5) and the partial roads occluded by urban high-rise buildings in urban areas (scene 6).

According to these results, the CADUNet proposed in this paper has achieved good results on the blocked roads in the rural, suburban and urban areas. It can be found that there is a gap between the results of DeepLab v3+ and UNet when the road is occluded by tree canopy and its shadows. The results derived from the proposed CADUNet are closer to the ground truth than those from the other methods. The smoothness of the road edges is significantly improved. UNet performs well in the scenes 3 and 6, but performs poorly in the scenes 1, 2 and 4. DeepLab V3+ performed well in the scenes 1 and 3, and DenseUNet performed well in the scenes 3 and 5 but did not perform well in the remaining scenes. In the scenes 1 and 2, the performance of CDenseUNet and GDenseUNet is poor, and the performance in the other scenarios is better. Finally, the CADUNet has achieved the best results in all six scenes by eliminating the occluding effects of tree canopies aside the road.

4.1.2. CVPR Dataset

Figure 7 shows the extracted results from the CVPR road dataset with these 6 methods. The first and second scenes contain roads occluded by tree canopies in rural areas, and the third and fourth scenes are roads covered by tree canopies in the suburbs. The fifth and sixth scenes show roads in the urban area covered by the shadows of high-rise buildings. Our CADUNet method has achieved good results in the information extraction of rural, suburban and urban roads. The first scene shows that when facing a partially covered road, the results obtained by the UNet method are better than that from DeepLabv3+ and DenseUNet. In the second and sixth scenes, when part of the tree canopy and high-rise building shadows block the road, the performance of DeepLab V3+ is better than UNet and DenseUNet. The DenseUNet only shows better performance in the fourth scene, while CDenseUNet performs better in the third and sixth scenes, merely. GDenseUNet and CADUNet obtained the best results in the first, second, third and fifth scenes occluded by the tree canopy. Obviously, the global attention mechanism plays an obvious role in extracting roads occluded by tree canopy and building shadows. In the CVPR dataset,

the global attention mechanism plays a key role in solving the occluding problem. Other methods show poor effects on the fourth scene, owing to not using the core attention mechanism. Therefore, this CADUNet method has achieved good results with the cascading dual attention mechanism.

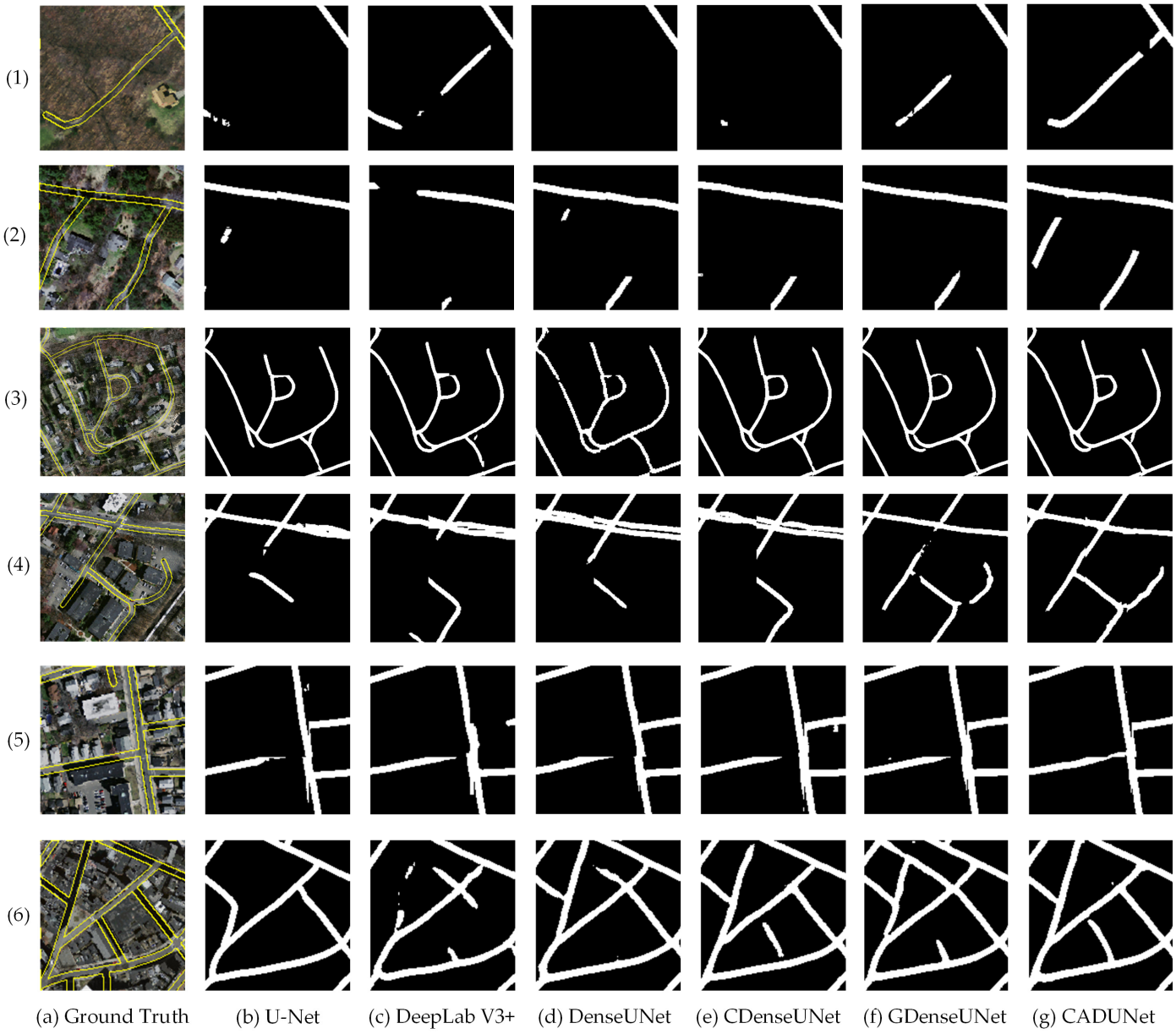

Figure 6. Pavement integrity and sideline smoothness of the extracted roads in the Massachusetts dataset. (**1**–**3**) The partial roads occluded by tree canopies in rural areas; (**4**,**5**) The partial roads occluded by tree canopies in suburbs; (**6**) The partial roads occluded by urban high-rise buildings in urban areas.

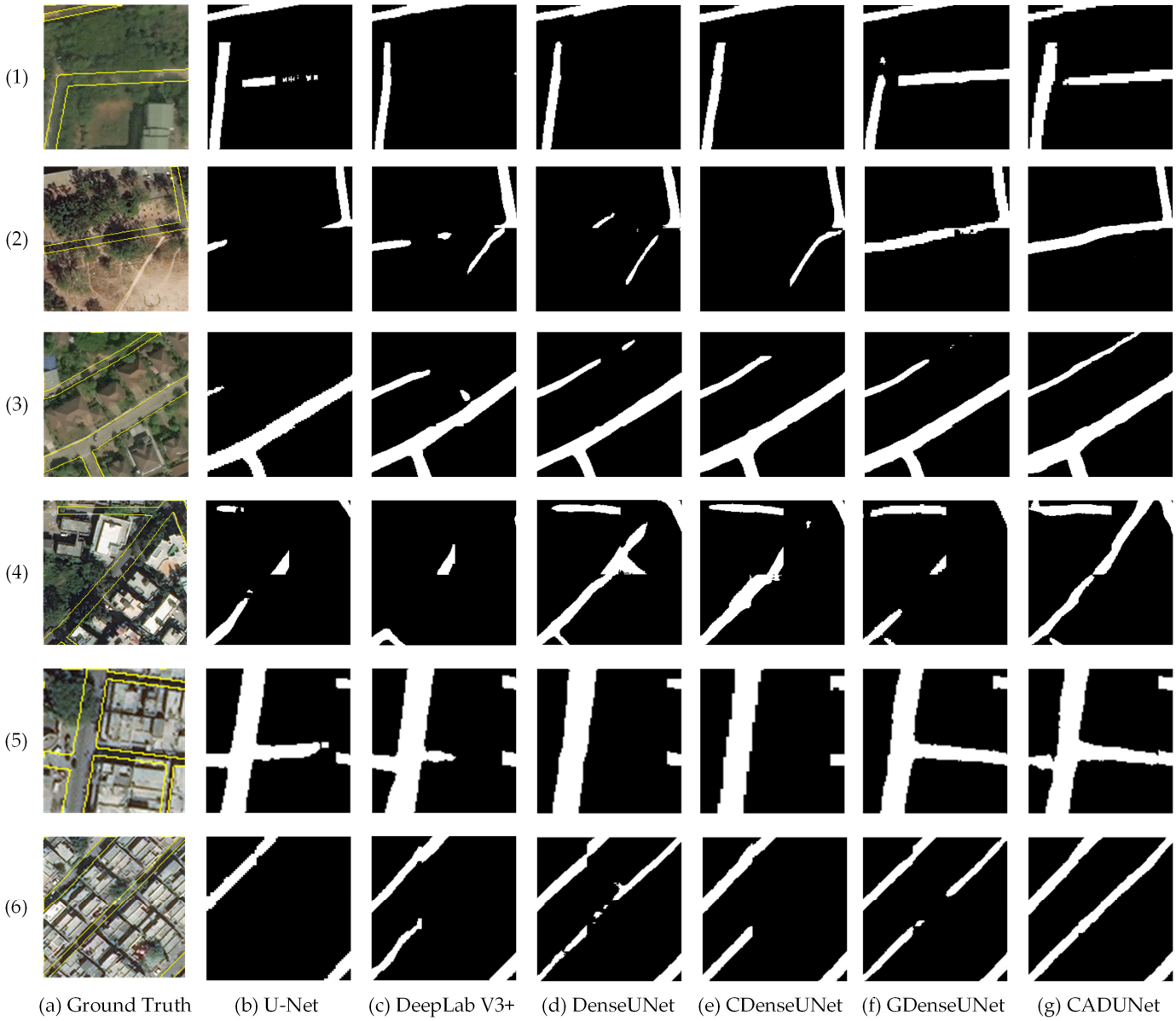

Figure 7. Pavement integrity and sideline smoothness of the extracted roads in the CVPR dataset. (**1**,**2**) The partial roads occluded by tree canopies in rural areas; (**3**,**4**) The partial roads occluded by tree canopies in suburbs; (**5**,**6**) The partial roads occluded by urban high-rise buildings in urban areas.

4.2. Road Network Connectivity

4.2.1. Massachusetts Dataset

For the Massachusetts road dataset, the 6 methods are used to extract complex road networks, including rural (scenes 1–3 in Figure 8), suburban (scenes 4–5 in Figure 8) and urban road networks (scene 6 in Figure 8), and transportation hub (scenes 7–8 in Figure 8). From the extraction results in rural, suburban and urban areas, CADUNet performs well on sparse rural roads, suburban and urban roads neighboring parking lots. When comparing other models, the CADUNet method not only depends on the visual characteristics of the road, but also has a certain reasoning ability by modeling the road context. It can be seen from Figure 8 that the road network obtained by UNet and Deeplab V3+ networks has obvious defects. Compared with UNet and DeepLab V3+, DenseUNet has some improvements. Compared with the standard DenseUNet, CDenseUNet and GDenseUNet reduce road interruption and enhance the connectivity of the road network. Compared with

the previous 5 models, the results obtained by CADUNet perform better road connectivity and fewer road interruptions.

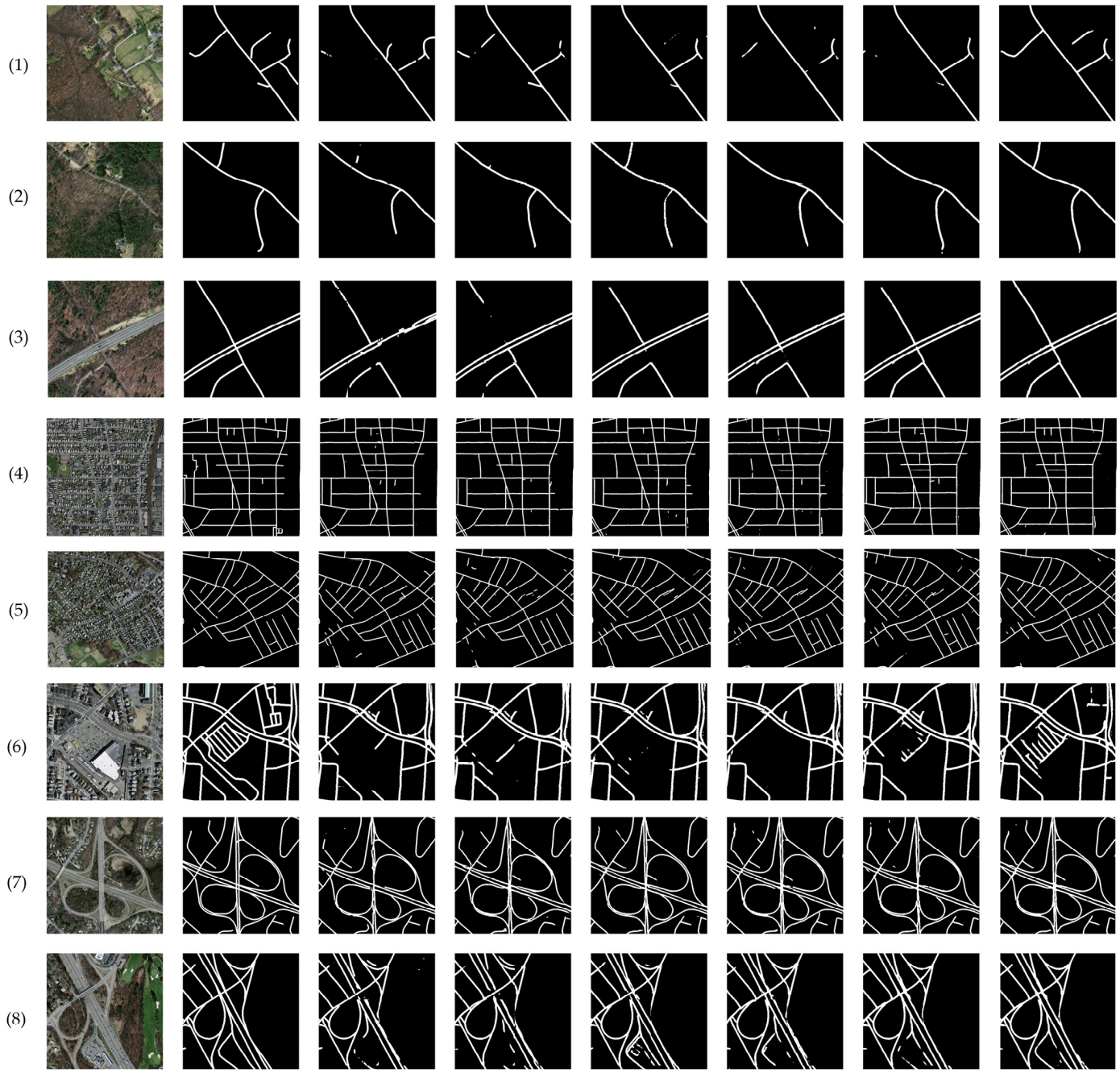

Figure 8. Connectivity of the road networks extracted from the Massachusetts roads dataset. (**1**–**3**) The rural roads; (**4**,**5**) The suburban roads; (**6**) The urban roads; (**7**,**8**) The transportation hub roads.

Accuracy assessment shows that the OA, recall, precision, F1-score and IoU obtained by CADUNet are the highest, reaching 98.00%, 76.55%, 79.45%, 77.89% and 64.12%, respectively (Table 1). Compared with UNet, the F1-score and IoU with the CADUNet method increased by 2.49% and 4.26%, respectively. Compared with the standard DenseUNet, the F1-score and IoU by CADUNet increased by 3.25% and 4.16%, respectively. After adding two attention modules, the intersection ratio by CADUNet increased by 3.04% and 2.21% respectively, compared to CDenseUNet and GDenseUNet.

Table 1. Accuracy assessment of the road extracting results obtained by six methods in the Massachusetts dataset.

Model Name	OA	Recall	Precision	F1-Score	IoU
U-Net	97.82%	73.29%	77.91%	75.40%	59.86%
DeepLab V3+	97.80%	72.30%	78.15%	74.89%	60.23%
DenseUNet	97.64%	76.29%	72.67%	74.64%	59.96%
CDenseUNet	97.80%	74.97%	76.49%	75.55%	61.08%
GDenseUnet	97.84%	75.90%	76.79%	76.17%	61.91%
CADUNet	98.00%	76.55%	79.45%	77.89%	64.12%

4.2.2. CVPR Dataset

As shown in Figure 9, the results based on the CVPR road dataset include rural roads (scenes 1–3), suburban roads (scenes 4–5) and urban roads (scene 6). The best results extracted by CADUNet are rural roads, followed by suburban roads and urban roads. Comparison among the 6 models shows that the results of UNet and DeepLab V3+ have the worst road network connectivity and severe road incompleteness. CDenseUNet and GDenseUNet have made progress based on DenseUNet, but still have their own shortcomings, and the connectivity of the road is poor. Due to imbedding the cascading dual attention mechanism into the DenseUNet, the CADNUnet method has obtained the best results in terms of road network connectivity.

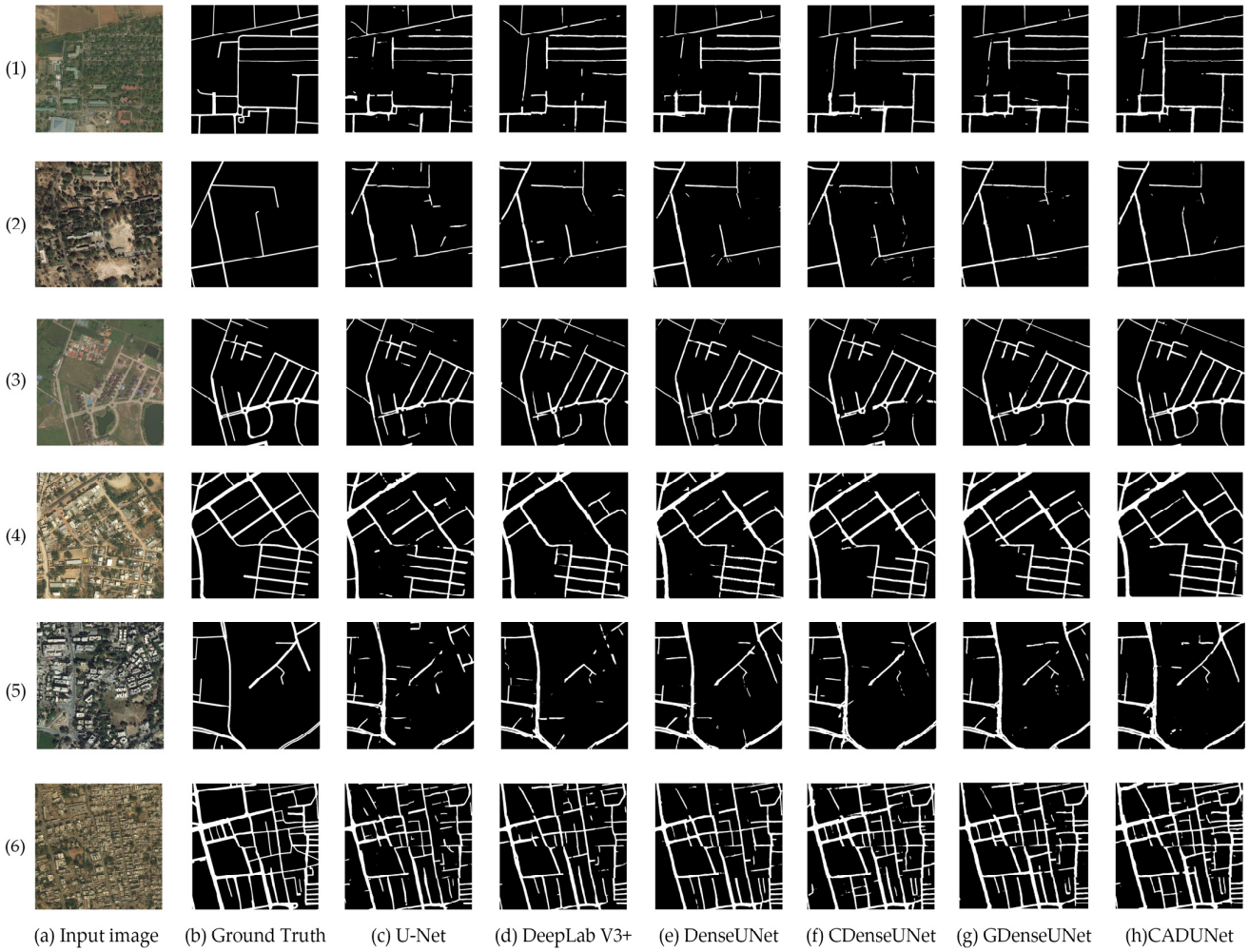

Figure 9. Connectivity of the road networks extracted from the CVPR roads dataset. (**1**–**3**) The rural roads; (**4**,**5**) The suburban roads; (**6**) The urban roads.

In the experiment with the CVPR road dataset, the CADUNet method reached the highest overall accuracy, F1-score and IoU, reaching 97.09%, 76.28% and 62.08%, respectively (Table 2). Compared with UNet, the recall and IoU of this method are increased by 6.14% and 3.83%, respectively. Compared with Deeplab V3+, the CADUNet method increases the IoU by 5.57%. After adding two attention mechanisms, the CADUNet method has increased F1-score and IoU by 1.67% and 2.11% compared to the DenseUNet.

Table 2. Accuracy assessment of the road extracting results obtained by six methods in the CVPR dataset.

Model Name	OA	Recall	Precision	F1-Score	IoU
U-Net	96.89%	72.52%	74.98%	73.16%	58.25%
DeepLab V3+	96.64%	74.27%	70.54%	71.73%	56.51%
DenseUNet	96.94%	77.78%	72.79%	74.61%	59.97%
CDenseUNet	96.96%	77.30%	71.47%	73.63%	58.75%
GDenseUnet	97.04%	77.15%	72.17%	73.93%	59.16%
CADUNet	97.09%	78.66%	74.89%	76.28%	62.08%

4.3. Loss Function

Figure 10a,b reflects the changes of the loss function with epochs on the Massachusetts and CVPR training datasets. As the training epochs increases, the losses of all 6 models gradually decrease with the increased training batches. The CADUNet proposed in this paper shows a better descending rate on the loss function than UNet, DeepLab V3+, CDenseUNet and GDenseUNet. UNet and DeepLab V3+ performed the worst. Figure 6c,d reflects the changes in the loss function corresponding to the training epochs on the Massachusetts and CVPR validation datasets, respectively. The CADUNet proposed in this paper has the lowest loss value verified on the two datasets, that is, the result obtained by the method is the closest to the truth. After 25 epochs of CADUNet, the model tends to be stable.

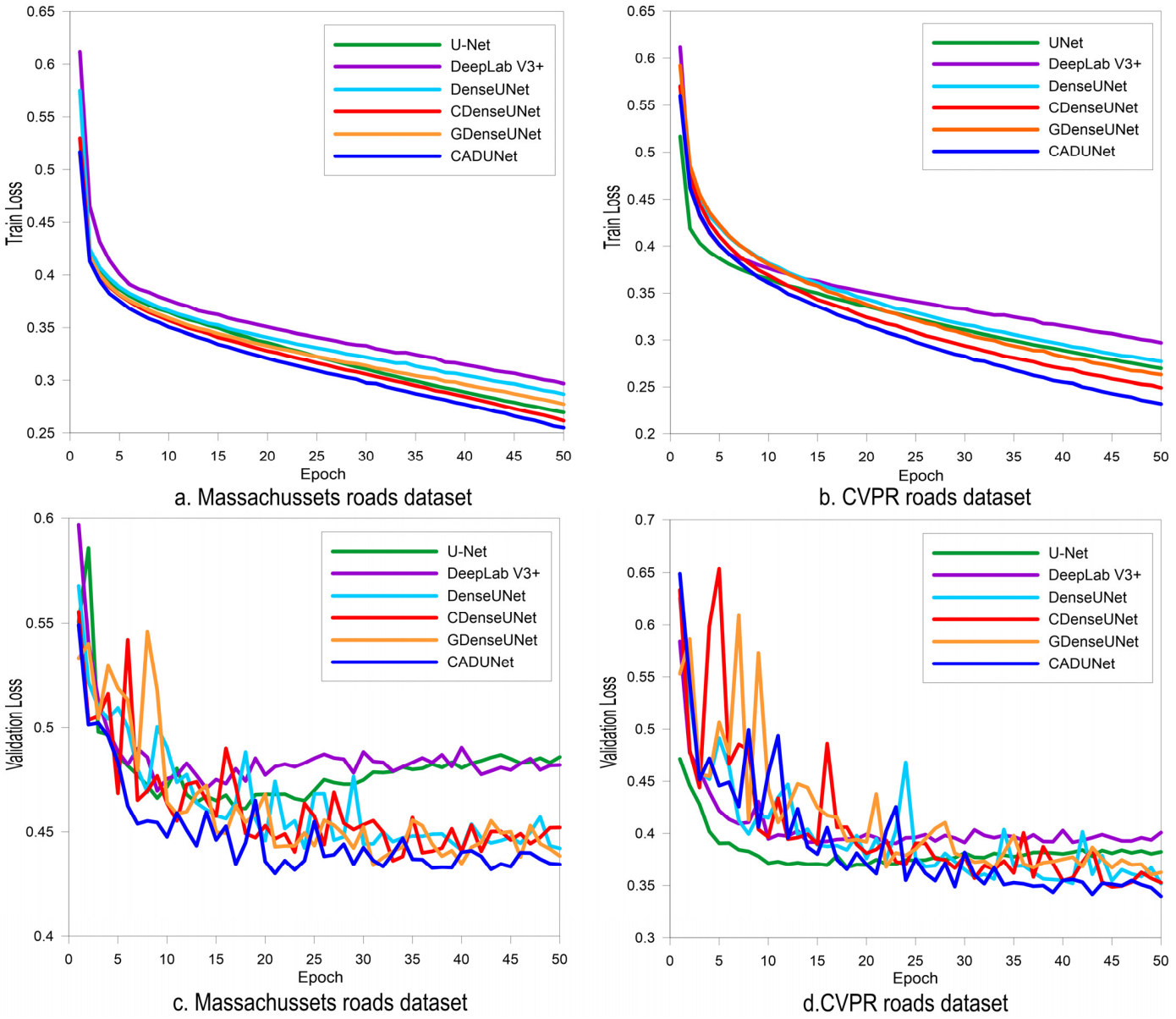

Figure 10. Training and validating curves.

5. Discussion

In the results of road extraction from VHR images, the occluding effect of the tree canopy and high-rise buildings aside the road often leads to the incompleteness of the road surface and even the interruption of the road network. As the basic framework of the proposed CADUNet, the DenseUNet semantic segmentation network performs well on employing the deep features of the image, avoiding gradient dispersion and making the network easy to train. Its feature reuse function can ensure that the most road information is preserved between the network layers, thereby improving the connectivity of the extracted road network. Therefore, it lays a solid foundation for road information extraction. Furthermore, the global attention module that we added to the DenseUNet model can enhance the global context information from the road feature map, thereby reducing the road interruption caused by tree canopy occlusion and building shadows to a certain extent, and the road integrity is significantly improved. We added the core attention module to the DenseUNet model to fuse more low-level features into the high-level feature map, so as to ensure that road information is transmitted to the greatest extent in dense blocks in the network, and further assist the global attention module to obtain more road

information at the encoding part. This module improves the connectivity of the road network, and at the same time restores the integrity of the road surface and the smoothness of the sideline at the decoding part.

Figure 11 shows the accuracy assessment results of six examples using the total six models on the Massachusetts dataset, where the green, red and blue areas represent TP, FP and FN, respectively. The first line in the figure shows an image with a loop road and its extraction results. Only the CDenseUNet and CADUNet models with the core attention mechanism have the most extent of TP and the least area of FP and FN, and the loop is relatively complete. This shows that the core attention mechanism makes up for the deficiency of the global attention mechanism to a certain extent. The second row shows the extraction result of the road that is sheltered by the elevated railway. It can be seen from these panels that the use of CDenseUNet, GDenseUNet and CADUNet models can extract limited roads sheltered by railways. This reflects the superiority of the core and the global attention modules. CADUNet has the most TP areas and the least FP and FN areas due to the use of two cascaded attention modules. The third row shows the extraction result of an image with the intersection of the main road and the minor road. With UNet and the DeepLab V3+ model, only the main road can be identified. Based on DenseUNet, CDenseUNet and CADUNet models, the extraction quality is better than the other three models. The CADUNet model achieves the largest TP areas and the smallest FP and FN areas, which embodies the advantage of the cascaded attention mechanism. The fourth row shows the extraction results of roads that are occluded by dense tree canopies on the roadside. The CDenseUNet, GDenseUNet and CADUNet models obtained good results. The dual attention mechanism integrated in the CADUNet model can solve the problem of roads being occluded by the tree canopy. It can be seen from the panels in row 5 that all the above six models can identify the main road but cannot identify the minor road connected to residential houses. In the Massachusetts dataset, the labeled dataset generally does not include such minor roads, so that they were ignored in the six network models when learning. Therefore, the error is due to the inconsistency of the labeling data and the overall labeling dataset. Row 6 concerns an image with the main and minor road intersection area. In the extraction results, the DeepLab V3+ and DenseUNet models present poor results, while the minor roads are not recognized. However, the main road and one of the minor roads can be well-identified by using CDenseUNet, GDenseUNet and CADUNet models, and the most TP areas and the least FP and FN areas can be achieved with the CADUNet model. At the same time, all six models still missed one of the minor roads labeled in the evaluation dataset. Although the minor road is labeled in the evaluation data, its features as a road are not obvious, which makes it difficult for the six models to recognize.

Figure 12 shows the accuracy assessment results of six examples of road extraction using the total six models on the CVPR dataset, and the color definitions are consistent with the foregoing. The panels in the first row show an image with the intersection of the main road and its minor roads in rural areas. These 6 models merely extract the main road, but not the minor road, which is related to the labeling dataset. In the labeling dataset, only a small part of the roads of this type are labeled, and most are not labeled. As a result, these deep learning models cannot be used to recognize the minor roads of this type. The panels in the second line show the country road that is shaded by trees. For this kind of image with rural roads, DenseUNet, CDenseUNet, GDenseUNet and CADUNet models have achieved good results, obtaining more TP areas and fewer FP and FN areas, which highlights the effectiveness of DenseUNet, as the basis of these networks, and the dual attention mechanisms in road extraction. For the image with parallel roads shown in the third row, the CADUNet model performs well, achieving the most TP area and the least FP and FN area, which reflects the superiority of the cascaded attention mechanism. However, there is still a gap between this extracting result and the labeled dataset, because one of the parallel roads is omitted from the labeling data. The images in the fourth row show the crossing area of the two roads, associated with a roadside canopy occluding effect. For this image, the CADUNet model achieved more TP areas and the least FP and FN

areas, achieving the best recognition effect, thus reflecting the advantages of the cascaded attention mechanism. The fifth and six row reflects the image of a curved road and its extracting effect in an urban area, and part of the road is obviously occluded by the shadow of the buildings. Good results were derived only through the CDenseUNet and CADUNet models, and the results through the other four models are relatively poor, which indicates that the core attention mechanism has a significant role in extracting this type of road.

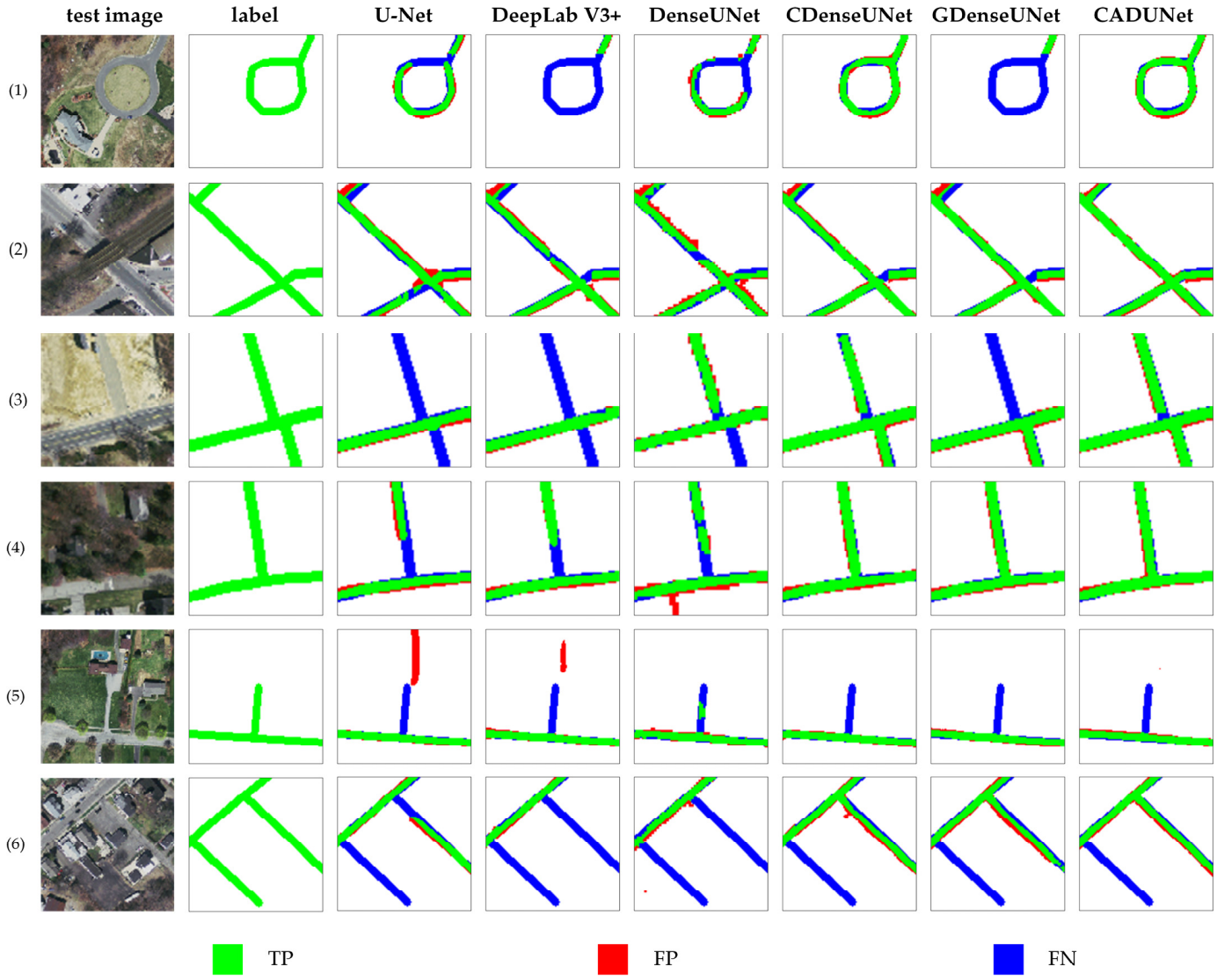

Figure 11. Visual accuracy assessment of road extraction based on the Massachusetts dataset. (**1**) The loop road; (**2**) The road sheltered by the elevated railway; (**3**) The intersection of the main road and the minor road; (**4**) The road occluded by dense tree canopies on the roadside; (**5**) The minor road connected to residential houses; (**6**) The main and minor road intersection area.

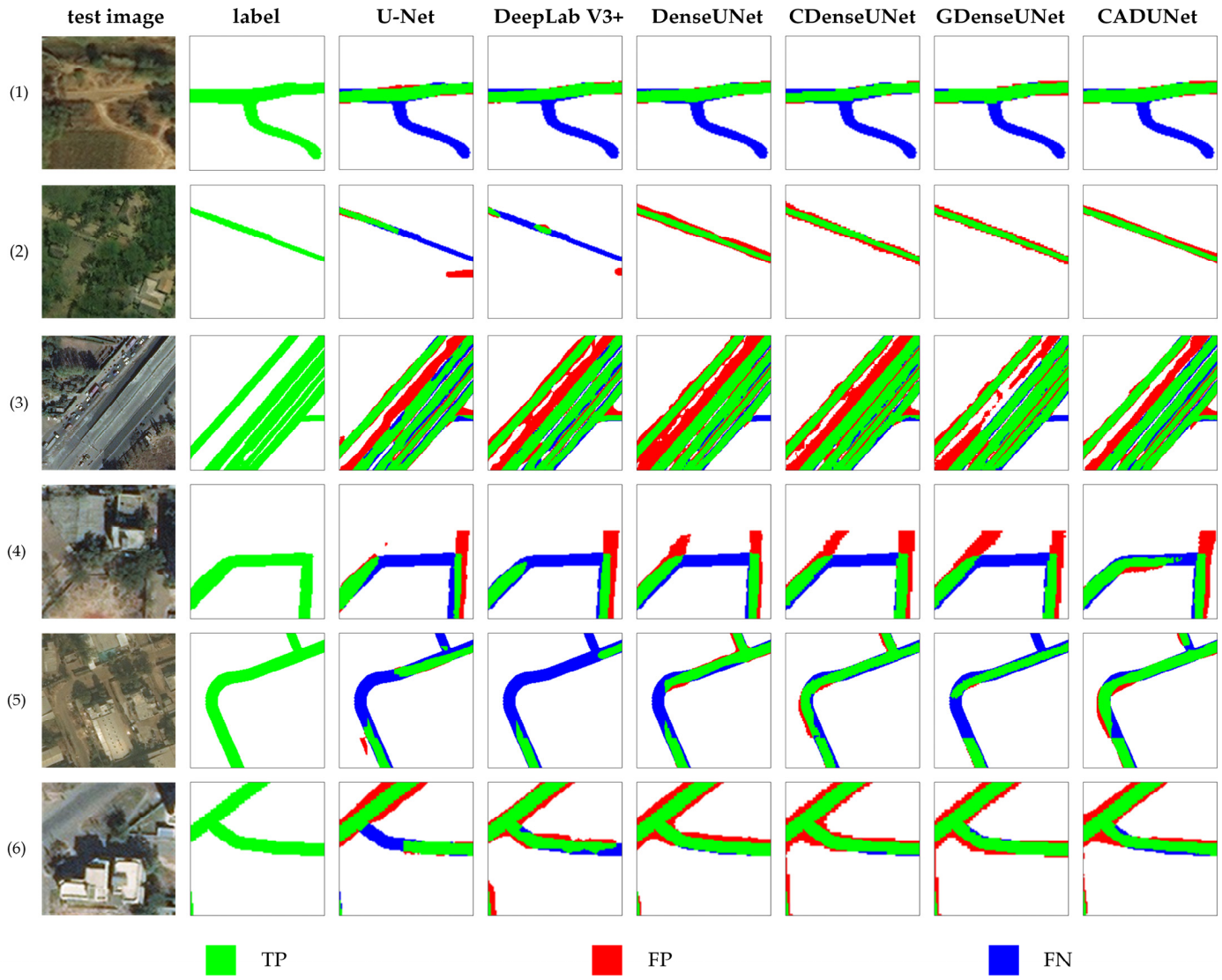

Figure 12. Visual accuracy assessment of road extraction based on the CVPR dataset. (**1**,**2**) The rural roads; (**3**) The parallel roads; (**4**) Intersection roads; (**5**,**6**) The urban road occluded by the shadow of the buildings.

6. Conclusions

In this study, we proposed an innovative CADUNet model based on the DenseUNet framework to solve the problems of incomplete road surface, uneven sidelines and poor road network connectivity due to roadside tree canopy in HRV images. We added global attention modules to obtain the global information of the road and introduced core attention modules to ensure that road information is transmitted to the greatest extent among the various layers of the network in dense ranges. The model can extract more road information from multiple locations to improve road integrity and enhance the robustness of feature extraction under tree canopy and urban high-rise building shadows. Finally, an adaptive loss function was introduced to balance the ratio of road areas to non-road areas in the training samples. This article used the Massachusetts dataset and the DeepGLOBE-CVPR 2018 dataset for comparative experiments. The results showed that the CADUNet model is more encouraging in road extraction from VHR images.

Although our network model has achieved good performance, there is still room for improvement for the problems of insufficient and excessive semantic segmentation of roads concerning sideline smoothness, interruption and the connectivity of the road network. In

addition, it is expected that the quality of the label data set will be further improved in the follow-up work.

Author Contributions: Jing Li and Yong Liu conceived and designed the study program; Jing Li conducted the experiment and wrote the manuscript; Yong Liu and Yindan Zhang provided revision opinions and experimental guidance; Yang Zhang participated in the discussion of project planning and paper revision. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science Foundation of China (NSFC), grant number 41271360.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data generated or analyzed during this study are included in this article.

Acknowledgments: The authors would like to thank the anonymous reviewers for their constructive comments. The authors are also grateful to Hinton et al. for providing the Massachusetts dataset and Demir et al. for providing the DEEPGLOBE-CVPR 2018 road extraction sub-challenge dataset. We would like to express our gratitude to Jianxi Hou of Hebei Changfeng Information Technology Co., Ltd. for participating in this research work and the improvement of the paper. The authors would like to thanks for the hardware support of the Supercomputing Center of Lanzhou University.

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

References

1. Zhang, Y.Y.; Zhang, J.P.; Li, T.; Sun, K. Road Extraction and Intersection Detection Based on Tensor Voting. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 1587–1590.
2. Frizzelle, B.G.; Evenson, K.R.; Rodriguez, D.A.; Laraia, B.A. The Importance of Accurate Road Data for Spatial Applications in Public Health: Customizing a Road Network. *Int. J. Health Geogr.* **2009**, *8*, 24. [CrossRef]
3. Li, D.; Deng, L.; Cai, Z.; Franks, B.; Yao, X. Intelligent Transportation System in Macao Based on Deep Self-Coding Learning. *IEEE Trans. Ind. Inform.* **2018**, *14*, 3253–3260. [CrossRef]
4. Courtrai, L.; Lefevre, S. Morphological Path Filtering at the Region Scale for Efficient and Robust Road Network Extraction from Satellite Imagery. *Pattern Recognit. Lett.* **2016**, *83*, 195–204. [CrossRef]
5. Abdollahi, A.; Pradhan, B.; Shukla, N.; Chakraborty, S.; Alamri, A. Deep Learning Approaches Applied to Remote Sensing Datasets for Road Extraction: A State-Of-The-Art Review. *Remote Sens.* **2020**, *12*, 1444. [CrossRef]
6. Mena, J.B. State of the Art on Automatic Road Extraction for GIS Update: A Novel Classification. *Pattern Recognit. Lett.* **2003**, *24*, 3037–3058. [CrossRef]
7. Das, S.; Mirnalinee, T.T.; Varghese, K. Use of Salient Features for the Design of a Multistage Framework to Extract Roads From High-Resolution Multispectral Satellite Images. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 3906–3931. [CrossRef]
8. Chaudhuri, D.; Kushwaha, N.K.; Samal, A. Semi-Automated Road Detection from High Resolution Satellite Images by Directional Morphological Enhancement and Segmentation Techniques. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2012**, *5*, 1538–1544. [CrossRef]
9. Wan, Y.; Wang, K.; Ming, D. Road Extraction from High-Resolution Remote Sensing Images Based on Spectral and Shape Features. In *MIPPR 2009: Automatic Target. Recognition and Image Analysis*; SPIE: Bellingham, WA, USA, 2009; Volume 7495, pp. 74953R-74951–74953R-74956. [CrossRef]
10. Gaetano, R.; Zerubia, J.; Scarpa, G.; Poggi, G. Morphological road segmentation in urban areas from high resolution satellite images. In Proceedings of the 2011 17th International Conference on Digital Signal Processing (DSP), Corfu, Greece, 6–8 July 2011; pp. 1–8.
11. Mirnalinee, T.T.; Das, S.; Varghese, K. An Integrated Multistage Framework for Automatic Road Extraction from High Resolution Satellite Imagery. *J. Indian Soc. Remote* **2011**, *39*, 1–25. [CrossRef]
12. Miao, Z.L.; Shi, W.Z.; Gamba, P.; Li, Z.B. An Object-Based Method for Road Network Extraction in VHR Satellite Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 4853–4862. [CrossRef]
13. Li, M.M.; Stein, A.; Bijker, W.; Zhan, Q.M. Region-Based Urban Road Extraction from VHR Satellite Images Using Binary Partition Tree. *Int. J. Appl. Earth Obs. Geoinf.* **2016**, *44*, 217–225. [CrossRef]
14. Shanmugam, L.; Kaliaperumal, V. Junction-Aware Water Flow Approach for Urban Road Network Extraction. *IET Image Process.* **2016**, *10*, 227–234. [CrossRef]

15. Buslaev, A.; Seferbekov, S.; Iglovikov, V.; Shvets, A. Fully Convolutional Network for Automatic Road Extraction from Satellite Imagery. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 197–1973.
16. Zhou, L.C.; Zhang, C.; Wu, M. D-LinkNet: LinkNet with Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 192–1924.
17. Gao, L.; Song, W.; Dai, J.; Chen, Y. Road Extraction from High-Resolution Remote Sensing Imagery Using Refined Deep Residual Convolutional Neural Network. *Remote Sens.* **2019**, *11*, 552. [CrossRef]
18. Blaschke, T. Object Based Image Analysis for Remote Sensing. *ISPRS J. Photogramm. Remote Sens.* **2010**, *65*, 2–16. [CrossRef]
19. Chen, M.J.; Sui, W.; Li, L.; Zhang, C.; Yue, A.Z.; Li, H.X. A Comparison of Pixel-based and Object-oriented Classification Using SPOT5 Imagery. In Proceedings of the 13th WSEAS International Conference on Applied Mathematics, Puerto De La Cruz, Spain, 14–16 December 2008; pp. 321–326.
20. Huang, X.; Zhang, L.P. Road Centreline Extraction from High-Resolution Imagery Based on Multiscale Structural Features and Support Vector Machines. *Int. J. Remote Sens.* **2009**, *30*, 1977–1987. [CrossRef]
21. Shi, W.Z.; 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]
22. Huang, Z.J.; Xu, F.J.; Lu, L.; Nie, H.S. Object-based Conditional Random Fields for Road Extraction from Remote Sensing Image. *Iop Conf. Ser. Earth Environ. Sci.* **2014**, *17*, 012276. [CrossRef]
23. Maboudi, M.; Amini, J.; Malihi, S.; Hahn, M. Integrating fuzzy object based image analysis and ant colony optimization for road extraction from remotely sensed images. *ISPRS J. Photogramm. Remote Sens.* **2018**, *138*, 151–163. [CrossRef]
24. Lu, X.Y.; Zhong, Y.F.; Zheng, Z.; Liu, Y.F.; Zhao, J.; Ma, A.L.; Yang, J. Multi-Scale and Multi-Task Deep Learning Framework for Automatic Road Extraction. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 9362–9377. [CrossRef]
25. Lian, R.; Huang, L. DeepWindow: Sliding Window Based on Deep Learning for Road Extraction From Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 1905–1916. [CrossRef]
26. Liu, Y.F.; Zhong, Y.F.; Qin, Q.Q. Scene Classification Based on Multiscale Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 7109–7121. [CrossRef]
27. Zhu, Q.Q.; Zhong, Y.F.; Zhang, L.P.; Li, D.R. Adaptive Deep Sparse Semantic Modeling Framework for High Spatial Resolution Image Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 6180–6195. [CrossRef]
28. Zhang, Y.; Xia, W.; Zhang, Y.Z.; Sun, S.K.; Sang, L.Z. Road Extraction from Multi-source High-resolution Remote Sensing Image Using Convolutional Neural Network. In Proceedings of the 2018 International Conference on Audio, Language and Image Processing (ICALIP), Shanghai, China, 16–17 July 2018; pp. 201–204.
29. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
30. 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–MICCAI 2015, Munich, Germany, 5–9 October 2015; pp. 234–241.
31. 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] [PubMed]
32. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
33. Zhang, Z.X.; Liu, Q.J.; Wang, Y.H. Road Extraction by Deep Residual U-Net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
34. Rezaee, M.; Zhang, Y. Road Detection Using Deep Neural Network in High Spatial Resolution Images. In Proceedings of the 2017 Joint Urban Remote Sensing Event (JURSE), Dubai, United Arab Emirates, 6–8 March 2017; pp. 1–4.
35. Cheng, G.L.; Wang, Y.; Xu, S.B.; Wang, H.Z.; Xiang, S.M.; Pan, C.H. Automatic Road Detection and Centerline Extraction via Cascaded End-to-End Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3322–3337. [CrossRef]
36. Tao, C.; Qi, J.; Li, Y.; Wang, H.; Li, H. Spatial Information Inference Net: Road Extraction Using Road-Specific Contextual Information. *ISPRS J. Photogramm. Remote Sens.* **2019**, *158*, 155–166. [CrossRef]
37. Li, Y.X.; Peng, B.; He, L.; Fan, K.L.; Li, Z.X.; Tong, L. Road Extraction from Unmanned Aerial Vehicle Remote Sensing Images Based on Improved Neural Networks. *Sensors* **2019**, *19*, 4115. [CrossRef] [PubMed]
38. Yang, X.F.; Li, X.T.; Ye, Y.M.; Lau, R.Y.K.; Zhang, X.F.; Huang, X.H. Road Detection and Centerline Extraction Via Deep Recurrent Convolutional Neural Network U-Net. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 7209–7220. [CrossRef]
39. Abdollahi, A.; Pradhan, B.; Alamri, A. 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]
40. 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]
41. Wegner, J.D.; Montoya-Zegarra, J.A.; Schindler, K. A Higher-Order CRF Model for Road Network Extraction. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Portland, OR, USA, 23–28 June 2013; pp. 1698–1705.

42. Chen, L.; Zhu, Q.; Xie, X.; Hu, H.; Zeng, H. Road Extraction from VHR Remote-Sensing Imagery via Object Segmentation Constrained by Gabor Features. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 362. [CrossRef]
43. Lai, R.; Li, Y.X.; Guan, J.T.; Xiong, A. Multi-Scale Visual Attention Deep Convolutional Neural Network for Multi-Focus Image Fusion. *IEEE Access* **2019**, *7*, 114385–114399. [CrossRef]
44. Wei, Z.; Song, H.; Chen, L.; Li, Q.; Han, G. Attention-Based DenseUnet Network with Adversarial Training for Skin Lesion Segmentation. *IEEE Access* **2019**, *7*, 136616–136629. [CrossRef]
45. Hu, H.J.; Li, Z.; Li, L.; Yang, H.; Zhu, H.H. Classification of Very High-Resolution Remote Sensing Imagery Using a Fully Convolutional Network With Global and Local Context Information Enhancements. *IEEE Access* **2020**, *8*, 14606–14619. [CrossRef]
46. Mnih, V.; Heess, N.; Graves, A.; Kavukcuoglu, K. Recurrent Models of Visual Attention. *Adv. Neural Inf.* **2014**, *27*, 2204–2212.
47. Zhao, H.S.; Shi, J.P.; Qi, X.J.; Wang, X.G.; Jia, J.Y. Pyramid Scene Parsing Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6230–6239.
48. Jetley, S.; Lee, N.; Torr, P.; Golodetz, S. Learn to Pay Attention. In Proceedings of the ICLR 2018, Vancouver, BC, Canada, 30 April–3 May 2018.
49. Ye, Z.R.; Fu, Y.Y.; Gan, M.Y.; Deng, J.S.; Comber, A.; Wang, K. Building Extraction from Very High Resolution Aerial Imagery Using Joint Attention Deep Neural Network. *Remote Sens.* **2019**, *11*, 2970. [CrossRef]
50. Jin, X.-B.; Zheng, W.-Z.; Kong, J.-L.; Wang, X.-Y.; Bai, Y.-T.; Su, T.-L.; Lin, S. Deep-Learning Forecasting Method for Electric Power Load via Attention-Based Encoder-Decoder with Bayesian Optimization. *Energies* **2021**, *14*, 1596. [CrossRef]
51. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; Mcdonagh, S.; Hammerla, N.Y.; Kainz, B. Attention U-Net: Learning Where to Look for the Pancreas. In Proceedings of the 1st Conference on Medical Imaging with Deep Learning (MIDL 2018), Amsterdam, The Netherlands, 4–6 July 2018.
52. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
53. Ma, Y.D.; Liu, Q.; Qian, Z.B. Automated Image Segmentation Using Improved PCNN Model Based on Cross-Entropy. In Proceedings of the 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, Hong Kong, 20–22 October 2004; pp. 743–746.
54. Zhou, M.; Sui, H.; Chen, S.; Wang, J.; Chen, X. BT-RoadNet: A Boundary and Topologically-aware Neural Network for Road Extraction from High-resolution Remote Sensing Imagery. *ISPRS J. Photogramm. Remote Sens.* **2020**, *168*, 288–306. [CrossRef]
55. Mnih, V.; Hinton, G.E. Learning to Detect Roads in High-Resolution Aerial Images. In Proceedings of the Computer Vision–ECCV 2010, Crete, Greece, 5–11 September 2010; pp. 210–223.
56. Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Bast, S.; Hughes, F.; Tuia, D.; Raskar, R. DeepGlobe 2018: A Challenge to Parse the Earth through Satellite Images. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 172–181.
57. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the ICLR 2015, San Diego, CA, USA, 5–8 May 2015.
58. He, H.; Yang, D.F.; Wang, S.C.; Wang, S.Y.; Li, Y.F. Road Extraction by Using Atrous Spatial Pyramid Pooling Integrated Encoder-Decoder Network and Structural Similarity Loss. *Remote Sens.* **2019**, *11*, 1015. [CrossRef]