remote sensing

MDPI

Article

MRENet: Simultaneous Extraction of Road Surface and Road Centerline in Complex Urban Scenes from Very High-Resolution Images

Zhenfeng Shao [1,*], Zifan Zhou [1], Xiao Huang [2] and Ya Zhang [1]

[1] The State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China; zhoufay@whu.edu.cn (Z.Z.); zhangya_rs@whu.edu.cn (Y.Z.)
[2] Department of Geosciences, University of Arkansas, Fayetteville, AR 72701, USA; xh010@uark.edu
* Correspondence: shaozhenfeng@whu.edu.cn

Abstract: Automatic extraction of the road surface and road centerline from very high-resolution (VHR) remote sensing images has always been a challenging task in the field of feature extraction. Most existing road datasets are based on data with simple and clear backgrounds under ideal conditions, such as images derived from Google Earth. Therefore, the studies on road surface extraction and road centerline extraction under complex scenes are insufficient. Meanwhile, most existing efforts addressed these two tasks separately, without considering the possible joint extraction of road surface and centerline. With the introduction of multitask convolutional neural network models, it is possible to carry out these two tasks simultaneously by facilitating information sharing within a multitask deep learning model. In this study, we first design a challenging dataset using remote sensing images from the GF-2 satellite. The dataset contains complex road scenes with manually annotated images. We then propose a two-task and end-to-end convolution neural network, termed Multitask Road-related Extraction Network (MRENet), for road surface extraction and road centerline extraction. We take features extracted from the road as the condition of centerline extraction, and the information transmission and parameter sharing between the two tasks compensate for the potential problem of insufficient road centerline samples. In the network design, we use atrous convolutions and a pyramid scene parsing pooling module (PSP pooling), aiming to expand the network receptive field, integrate multilevel features, and obtain more abundant information. In addition, we use a weighted binary cross-entropy function to alleviate the background imbalance problem. Experimental results show that the proposed algorithm outperforms several comparative methods in the aspects of classification precision and visual interpretation.

Keywords: multitask learning; convolutional neural networks; road surface extraction; road centerline extraction; VHR remote sensing images

Citation: Shao, Z.; Zhou, Z.; Huang, X.; Zhang, Y. MRENet: Simultaneous Extraction of Road Surface and Road Centerline in Complex Urban Scenes from Very High-Resolution Images. *Remote Sens.* **2021**, *13*, 239. https://doi.org/10.3390/rs13020239

Received: 24 November 2020
Accepted: 8 January 2021
Published: 12 January 2021

1. Introduction

Urban information construction requires the rapid acquisition of a large amount of basic geographic information data. Extracting ground objects using remote sensing images has several advantages, such as large detection range, wide spatial coverage, timeliness, and low cost, making it an important means to construct and update geospatial databases [1]. Road extraction is of great significance for GIS database updates, image matching, target detection, and digital mapping automation, to list a few. It is widely used in traffic management, land use analysis, and other fields [2–5]. With the increasing maturity of remote sensing technology and its applications, more and more scholars started to extract road information directly from very high-resolution (VHR) remote sensing images [5,6].

However, urban roads are generally distributed in a plane shape, especially in VHR images. The rich details of ground features add further complexity to the image information. As for spectral characteristics, there are a large number of the same objects with different

spectra and the same spectra of different objects in the image. For example, the spectral characteristics of roads and buildings are very similar, while the spectral signature inside the road greatly differs. The existence of a large number of geometric topological features around the road leads to more challenges in road extraction [7–9]. In addition, roads tend to be covered by the shadow cast from the adjacent three-dimensional structures in modern cities. Due to the aforementioned influencing factors, extracting road-related information from VHR remote sensing images has been considered to be a rather difficult task [4].

Extracting road-related information from remote sensing images includes two major tasks: (1) road surface extraction and (2) road centerline extraction. Road surface extraction aims to generate pixel-level results, while road centerline extraction aims to extract road skeleton [6].

With the great progress of space technology, mankind entered the era of space remote sensing in the 1970s [10]. Scholars have carried out in-depth studies on road surface extraction models via a variety of approaches that include template matching [11–13], knowledge-driven methods [14,15], and object-oriented methods [5,16–18]. Inspired by the roads seen on satellites, Ruzena et al. [1] designed a computer program for the recognition and description of roads and their intersections in 1976. Helmut et al. [19] used snakes to make up for the gap caused by the shelter of buildings and trees. Trinder et al. [20] proposed a knowledge-based method to extract roads in an automatic manner. Their approaches consist of low-level image processing for edge detection and linking, mid-level processing for the formation of road structure, and high-level processing for the recognition of roads. Based on the anti-parallelism rule and the proximity rules, Dal Poz et al. [21] achieved great results by taking advantage of the characteristics of parallel edges of roads leveraging road seed extraction and road network combination. Rasha et al. [18] extracted roads from VHR images through three steps: feature extraction, graph-based segmentation, and post-processing. Li et al. [5] viewed road areas as binary segmentation trees and combined them with various features to provide an effective method to extract roads from VHR satellite images in densely populated urban areas. Cao et al. [22] chose GPS data for rapid centerline extraction to face challenges such as complex scenes and variable resolution. Liu et al. [23] conducted road surface extraction based on the generalized Hough transform, which has low computational complexity and high time efficiency. In recent years, the advancement of deep learning offers a new solution to road surface extraction tasks. In 2010, Minh et al. [24] attempted to apply neural network technology to road surface extraction tasks with the city-level spatial coverage. Since then, more and more studies have been conducted to extract roads from remote sensing images via convolutional neural networks. Wei et al. [25] built the road-structure-based loss function by embedding the geometric structure of roads and proposed a road structure refined convolutional neural network approach for road surface extraction from aerial images. To facilitate the extraction of tree-blocked roads, Zhang et al. [6] proposed a semantic segmentation neural network that combines the advantages of residual learning and Unet [26] to extract the road area. Cheng et al. [27] proposed a new cascading end-to-end convolutional neural network named CasNet, which handles road surface extraction and centerline extraction tasks simultaneously. Liu et al. [28] proposed RoadNet, a multitask convolutional neural network to predict the road surfaces, edges, and centerlines. Lu et al. [29] adopted U-Net as the basic network of multitask learning and improved the robustness of feature extraction by applying multiscale feature integration. Batra et al. [30] conducted joint learning on the location and division of roads and further improved the connectivity of roads. Focusing on the modeling of road context information, Qi et al. [31] proposed a well-designed spatial information reasoning structure. More recently, Zhang et al. [32] developed a novel road surface extraction method based on improved generative adversarial networks.

In the aspect of road centerline extraction from remote sensing images, the research method mainly focuses on obtaining the linear road skeletons by applying two general steps: (1) thinning and (2) tracking, where the thinning is often carried out after the extraction of road surfaces [33]. Amini et al. [34] used the parallel line theory to obtain the road

skeleton in rural areas after the process of road refinement. Zheng et al. [35] extracted road centerlines from VHR satellite images using support vector machine and tensor voting techniques. Miao et al. [36] first identified potential road sections and then applied multivariate adaptive regression splines to extract the centerlines of the road in VHR images. Cheng et al. [37] targeted the problems of the existence of burr and the inability in the extraction of intersections. They obtained the segmentation results through semisupervised segmentation, multiscale filtering, and multidirection nonmaximum suppression in another study [38]. Gao et al. [39] proposed a semiautomatic road centerline extraction method combining edge constraints and fast marching. Zhou et al. [40] reconstructed roads via boundary distance field and tensor field after obtaining preliminary road and road centerline results. The advance of deep learning has largely facilitated the extraction of road centerlines, especially in the last five years. Wei et al. [41] obtained the confidence map of road centerline based on an end-to-end convolutional neural network and then achieved accurate road centerline extraction by nonmaximal inhibition. Zhang et al. [42] proposed a learning-based road network extraction framework via a multisupervised generative adversarial network, jointly trained by the spectral and topology features of the road networks. Cascading deep learning framework based on multitask networks has been the mainstream idea to solve road-related tasks [27–29], which builds the foundation of our work.

However, most existing road datasets under ideal conditions cannot provide more possibilities for the task of road extraction, and the image information is not fully utilized when two related tasks are extracted separately. In this paper, we propose a two-task and end-to-end convolution neural network, termed Multitask Road-related Extraction Network (MRENet), for road surface extraction and road centerline extraction. Inspired by the main structure of Unet [26], we use atrous convolution to expand the receptive field of feature extraction in the encoder and apply pyramid scene parsing pooling module (PSP pooling) to fuse global context information for pixel-level annotations in the decoder [14,43]. Through information transmission and parameter sharing between the two tasks, the characteristics and the results of road surface extraction are fed into the road centerline network by a concatenating operation to achieve a rapid extraction of single-pixel road centerlines. We select the complex road scenes in China from remote sensing images in GF-2 (details can be found in Section 2.3) and manually annotate the acquired images as the dataset. In terms of the loss function, we use a weighted binary cross-entropy function to alleviate the background imbalance of the datasets. Although Chinese cities were selected as experimental area, the proposed model is expected to be applicable in other complex urban scenes.

The contributions of this paper mainly include the following three aspects:

(1) We introduce a new challenging dataset derived from GF-2 VHR images. The introduced dataset contains complicated urban scenes, which can be better considered as a reflection of the real world, providing more possibilities for road-related information extraction, especially under less ideal situations.
(2) We propose a new network named MRENet that consists of atrous convolutions and a PSP pooling module. The experiments suggest that our approach outperforms existing approaches in both road surface extraction and road centerline extraction tasks.
(3) We conduct a group of band contrast experiments to investigate the effect of incorporating NIR band on experimental results.

The remainder of this paper is organized as follows. Section 2 describes road features and data sources. Section 3 elaborates on the proposed network. Section 4 presents the experiments and analysis, detailing our comparative experiments and analysis of experimental results. Further discussions are arranged in Section 5. Finally, Section 6 presents a summary of our work.

2. Materials

One of the main reasons for the great success of deep learning lies in the massive training data. The performance of deep learning algorithms is largely dependent on the scale and annotation of the training dataset [44]. Open-source databases in the computer vision domain have greatly stimulated the development of deep learning. Unlike conventional natural images, VHR remote sensing images own unique characteristics, such as diverse scales [45], special perspectives [46], and complex backgrounds [47]. Thus, training deep learning models on remote sensing images often requires specialized databases. DOTA [44], AID [48], and other similar datasets, to a certain extent, greatly enriched the database of remote sensing images.

However, it is worth noting that road datasets from VHR remote sensing images are still rare and do not meet the demand. Existing datasets for road extraction, such as the datasets proposed in CasNet [27] and RoadNet [28], contain images that were taken under ideal conditions. Similarly, the datasets used by Das et al. [7] and Cheng et al. [27] use images with simple and clear backgrounds without any occlusion. The dataset proposed by Liu et al. [28] is more challenging as its images include the situation of tree occlusion. The images in the aforementioned road datasets were mainly collected from Google Earth with RGB bands. However, modern urban complex road scenes (e.g., overpass and ring road) and the impact of municipal facilities and road greening are often not included. Thus, their low complexity is inadequate to be considered as a reflection of the real world [44].

Complex urban roads are characterized by composite lane structures, dense traffic, and complex color scheme. In addition, the shadow from the vegetation on both sides of the road, along with the shelter of high-rise buildings, further adds complexity, making road extraction a very challenging task. With the abundant remote sensing data, it is of great importance to establish a challenging dataset with complex scenes, benefiting road extraction tasks in complex environments.

In the following sessions, we summarize the characteristics of the road surface and road centerline under the complex urban scenes of VHR remote sensing images.

2.1. Characteristics of the Road Surface

The urban roads in the VHR remote sensing image mainly include the urban trunk roads and the internal roads of the parcels. The challenges in road extractions from VHR remote sensing images lie in the variance of road width and the existence of traffic management lines, isolation belt, cars, and shadows (cast by poles, buildings, roadside trees, and overpasses).

- In terms of geometric characteristics, urban roads are generally described as a narrow and nearly parallel area with a certain length, stable width, and obvious edge. Both the edge and the centerline have obvious linear geometric features, often with a large length–width ratio;
- In terms of radiation characteristics, roads have distinct spectral characteristics compared with vegetation, soil, and water, but they can be easily confused with artificial structures such as parking lots. The grayscale of the road tends to change uniformly, which generally shows the color of black, white, and gray. However, due to the existence of a large number of vehicles and pedestrians on the surface, such noise interference is inevitable;
- In terms of topological characteristics, urban roads are generally connected with each other, forming a road network with high connectivity;

2.2. Characteristics of the Road Centerline

In remote sensing images, roads are symmetrically distributed in a geometric structure. Road centerlines are important feature lines in the geometric design of road alignment. The extraction of road centerlines is to obtain a linear road skeleton, which is a smooth and complete symmetrical line with single-pixel width.

Road centerlines are generally connected and have unique characteristics that include strong connectivity, complex topology, accurate refinement, and the central axis. Thinning and tracking are the two commonly used methods to extract the road centerlines, where thinning is further expanded on the results of road extraction.

2.3. Description of Datasets

Given the small scale and low diversity in existing datasets [49], we developed a new challenging dataset by collecting VHR images from the GF-2 satellite in complex scenes. We manually marked the accurate reference map of road surface and road centerlines. Our dataset consists of two subdatasets: (1) road surface extraction dataset and (2) road centerline extraction dataset.

The spatial resolution of multispectral bands and the panchromatic band from the GF-2 satellite are 4 m and 1 m, respectively, as shown in Table 1. The multispectral bands for training and testing include four bands: near-infrared (NIR), red (R), green (G), and blue (B). We further discuss the band selection and suitability analysis of remote sensing images in Section 5.1. The road width ranges from 5 pixels to 50 pixels, and the width of the road centerline is 1 pixel.

Table 1. Satellite parameters of GF-2.

Bands	Resolution (m)	Wavelength (μm)
Pan	1	0.45–0.90
Blue	4	0.45–0.52
Green	4	0.52–0.59
Red	4	0.63–0.69
NIRed	4	0.77–0.89

In the process of selecting areas and collecting samples, in order to avoid the problems of poor generalization ability and overfitting caused by excessive imbalance background, we deliberately select the concentrated area and typical area (e.g., overpasses, loops, intersections, etc.) and ignore regions with an excessive background in the image, as shown in Figure 1. Compared with other datasets, the image background of the GF dataset is considered more complex, which can better represent the typical urban road situation in developing countries and in other complex urban fabrics.

We apply data augmentation techniques to increase the amount of training data, fully mine the multidimensional information of data, avoid model underfitting, reduce the imbalance between samples, and further improve the generalization ability of the model. In our experiment, we transform the image geometry without changing the content of the image itself. The data augmentation includes flipping operation and rotation operation. For direction-insensitive tasks such as image classification, the geometric transformation has been proved to be a very effective data augmentation method.

All the testing images are not included in the training dataset, and they are uniformly cut into 256 $\times$ 256 pixels with overlapping of 64 pixels. A total of 13,590 images are divided into the training set, testing set, and validation set. 10,872 images are used to train the network, and the rest are evenly allocated for validation and testing.

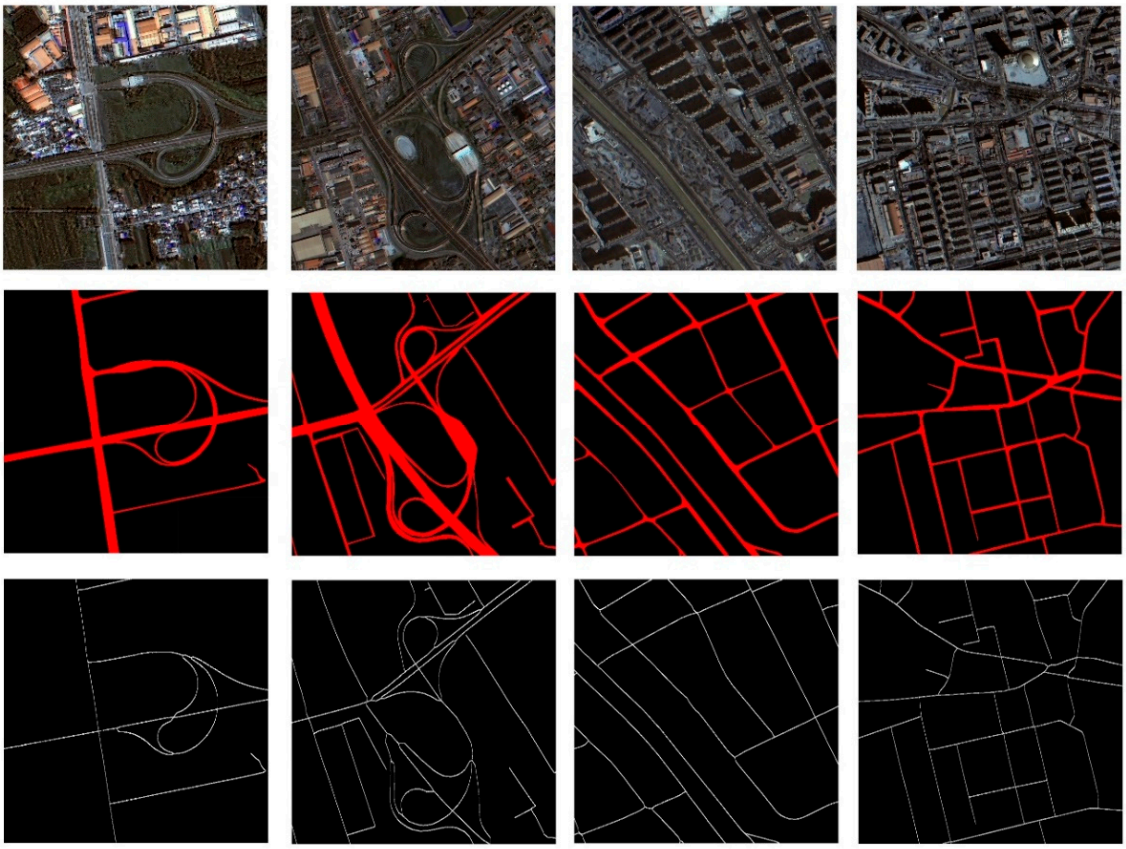

Figure 1. Representative images in our dataset.

3. Methodology

This section details the method proposed in this study. In particular, Sections 3.1 and 3.2 introduce the theory of the Resblock and the pyramid scene parsing (PSP) pooling module. Section 3.3 explains the advantages of multitask learning. Section 3.4 presents the overall architecture of Multitask Road-related Extraction Network (MRENet).

3.1. ResBlock

In the process of semantic segmentation using FCN, the input image is first convoluted and further pooled, similar to the traditional CNN network. The convolution operation extracts image features, and the pooling operation enlarges the receptive field by reducing the image size. Two steps are repeated to obtain the most important features. In the decoding part, it is necessary to enlarge the size of the pooled image to its original size via upsampling. When the image size is reduced, the rich spatial information of pixels in the original images is lost, potentially leading to imbalanced local and global features.

VGGNet [14] proves that using a small convolution kernel can effectively reduce the computational complexity of convolution operation rather than using a large convolution kernel. The deepening network structure and the regularization effect of a small convolution kernel also improve the performance of the model.

To obtain a larger receptive field and more spatial information without pooling operation, a common method is to introduce atrous convolution, a convolution that adds holes to the standard convolution layer to extract features. As shown in Figure 2, compared with the ordinary convolution layer, the atrous convolution can effectively expand the receptive field of feature extraction, retaining the feature size and reducing the loss of spatial information of features without increasing parameters. It has been proved to be an effective feature extraction approach [50]. We use multiple parallel atrous convolution branches in the Resblock module of the network to obtain the global characteristics of

the road surface and road centerline in remote sensing images. As shown in Figure 3, before each convolution operation, batch normalization and Relu activation function are used. The dilation rate is set to 1, 3, 15, and 31. For example, when the dilation rate is 31, the receptive field size is 123 × 123, which fully covers the road width (5–50 pixels). The parameters were derived from Diakogiannis et al. [43].

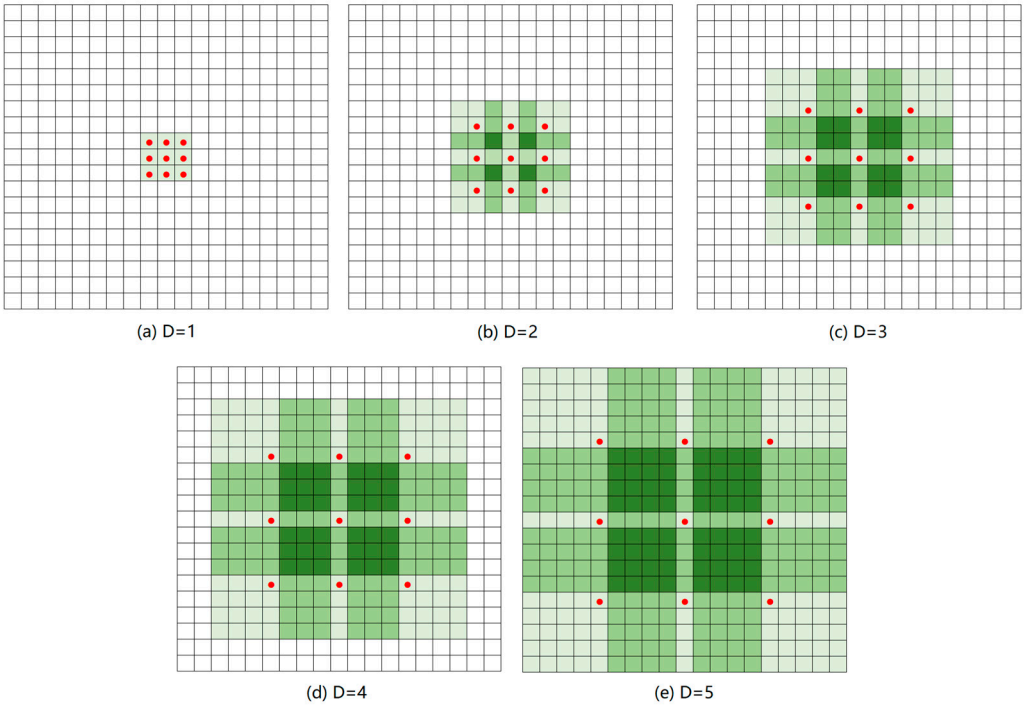

Figure 2. Examples of atrous convolution. The green boxes represent the size of the receptive field under different dilation rates. D represents the dilation rate.

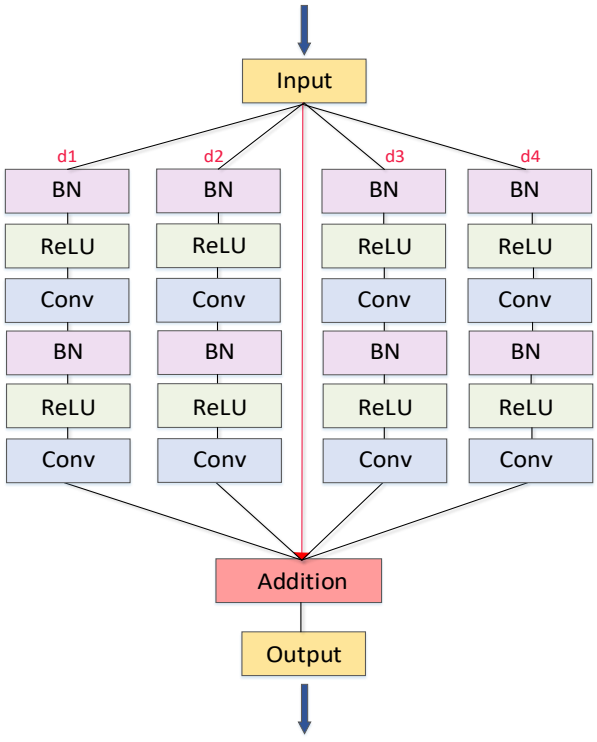

Figure 3. Structure of the Resblock module.

The multiscale information can be obtained using different dilation rates on different scales. Each scale is an independent branch. These branches are combined by the addition operation and then connected by the next convolution layer. Such a design can effectively avoid redundant information, improve the performance by identifying the correlation among objects at positions in the image, effectively expand the receptive field of feature extraction, and avoid the loss of semantic meaning of a single-pixel in distinguishing small-scale feature extraction. In view of the problem of road extraction, more contextual information can be obtained by expanding the acceptance domain of feature extraction, which can well solve the problem of insufficient semantic features caused by the large coverage of ground objects and tree occlusion.

3.2. PSP Pooling

In the last stage of encoding and decoding, we apply PSP pooling to integrate multilevel features [43,51]. As shown in Figure 4, the main idea of PSP pooling is to divide the initial input into four subregions at varying levels in the feature space and obtain the pooled features of these four subregions, respectively. A 1×1 convolution layer is used to reduce the dimension of the context features to maintain the weight of the global features. The size of the original image is further restored by a resize operation, and finally, the pyramid pooling features are obtained by splicing and fusion in the channel dimension.

The size of receptive fields in a neural network represents the range of context information obtained, but extraction errors can also occur after the acceptance domain has been expanded. The emphasis is on the need to fuse the key information received. The PSP pooling module integrates four different pyramidal scale features, leading to better characteristic expression ability. It can largely facilitate the integration of global context information of pixel-level annotation (low-level spatial information and high-level semantic information) and improve the performance of the model to distinguish roads from other ground objects [52].

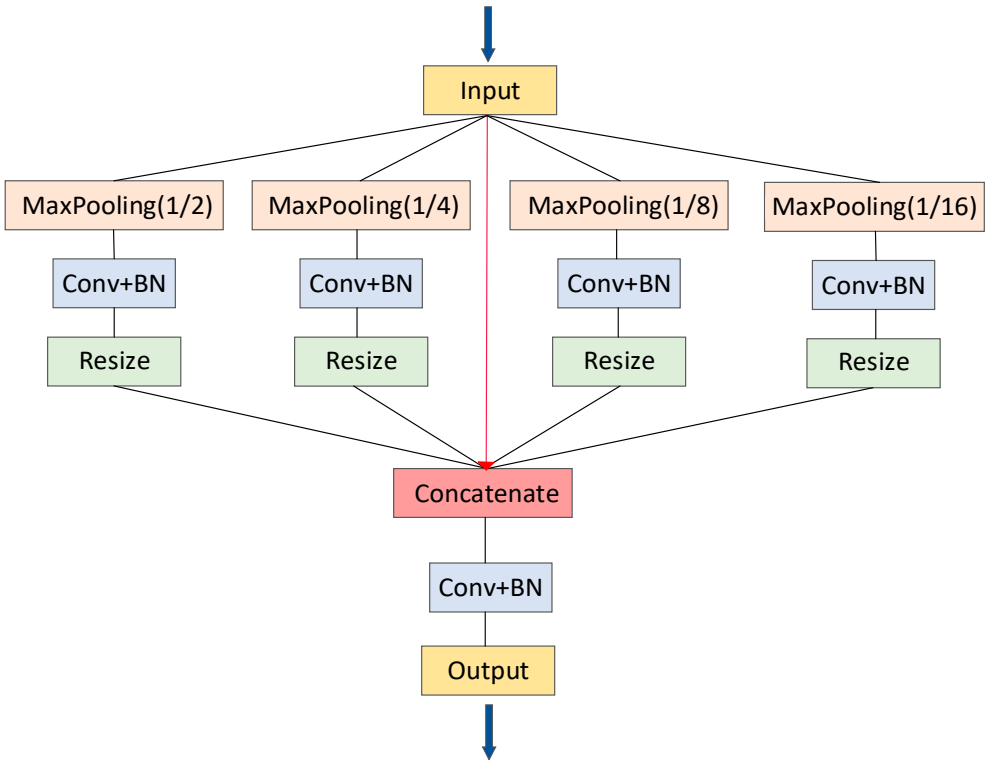

Figure 4. Structure of the PSP pooling module.

3.3. Multitask Learning

The basic idea of multitask learning is that when a task to be learned is similar or related, cross-sharing of information between tasks in the model may be advantageous [53]. Multitask learning is achieved by learning tasks in parallel using shared representations. The shared representations lead to effective joint learning among multiple tasks [54].

The tasks of road surface extraction and road centerline extraction are dependent to a certain extent. The road extraction results play a decisive role in the centerline extraction, while the centerline enhances the typical linear features of the road [29]. Therefore, it is beneficial to introduce the concept of multitask learning towards a simultaneous extraction of the road surface and road centerline. Through information transmission and parameter sharing between the two tasks, massive training samples are not necessarily required, and the risk of overfitting is reduced.

Specifically, the problem of road surface extraction and road centerline extraction is to transfer the knowledge learned in the road surface extraction process to the road centerline extraction process. By cascading of two tasks, the features extracted from road surfaces are taken as the condition of centerline extraction, compensating for the potential problem of insufficient road centerline samples. Given the existence of noises and the unbalanced ratio between backgrounds and targets, deriving the optimal descent direction of the gradient is often computationally demanding. Multitask learning facilitates the feature sharing and transferring between road surface extraction and road centerline extraction, leading to the retrieval of complete semantic features and model robustness.

3.4. MRENet Architecture

In order to better solve the tasks of road surface extraction and road centerline extraction in complex scenes from VHR remote sensing images, we propose a new two-task, end-to-end deep learning network by adopting the Resblock module and PSP pooling module in the network based on the concept of multitask learning. The MRENet architecture is shown in Figure 5. We regard the tasks of road surface extraction and road centerline extraction as two binary classification tasks. For road surface extraction, we use an encoder–decoder structure, i.e., a Resblock module, that uses multiple parallel atrous convolution branches to obtain the relationship between the road and environment backgrounds. The PSP pooling module further aggregates multiscale and multilevel features to obtain rich context information. Meanwhile, our network structure retains the skip connection structure in Unet [26] and transmits the feature map directly from the encoder to the decoder through the combination operation. This information transmission achieves the integration of the deep and shallow features, providing more fine features for segmentation.

For road centerline extraction, we believe that the features extracted by the encoder in road surface extraction (e.g., road location, directionality, etc.) can be transferred and utilized in the task of road centerline extraction. We aim to achieve this convolution layer sharing to supplement the training difficulties caused by the lack of centerline samples. Different from CasNet [27], we feed the road surface extraction results and the four convolution layers of the road network encoder into the road centerline network to achieve a rapid extraction of single-pixel road centerlines, as shown in Figure 5.

Considering the unbalanced proportion of negative (background) and positive samples (road surface and road centerline) in the training dataset, we further improve the binary cross-entropy loss function by adding a corresponding weight for each category, which was determined by the proportion of the category in all samples.

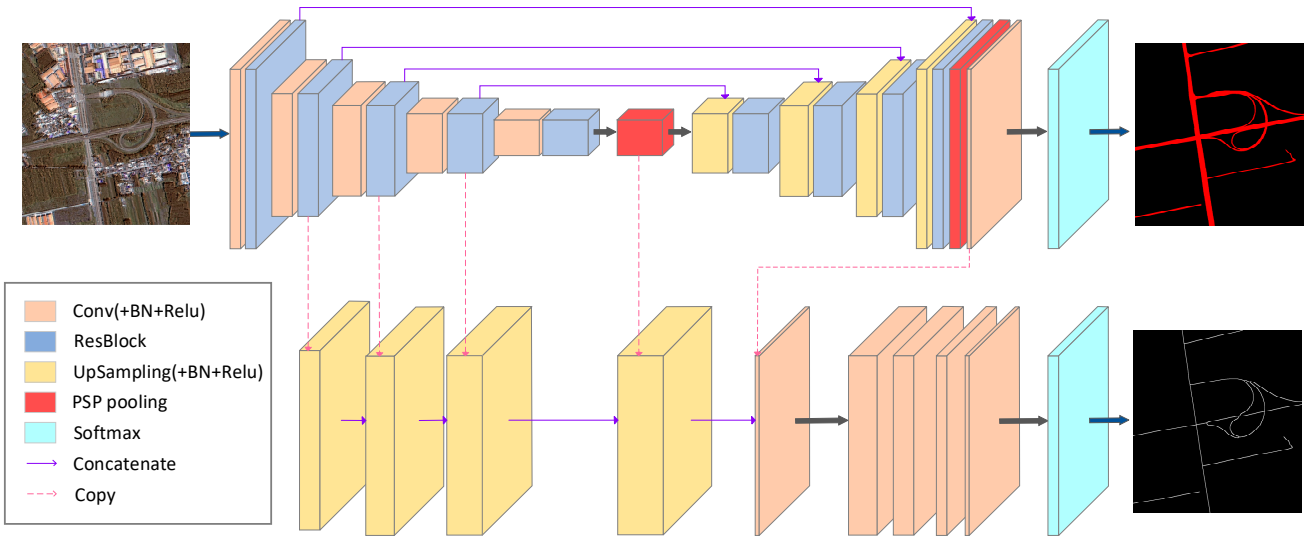

Figure 5. Flowchart of MRENet, which contains two joint convolutional neural networks: road surface extraction network and road centerline extraction network.

4. Experiments and Results

In this section, we evaluate the performance of the proposed MRENet on the developed road datasets. We introduce the evaluation metrics, detail the experimental settings, and present the experimental results.

4.1. Evaluation Metrics

We use a confusion matrix to evaluate the model performance in the binary classification problem. The labels are divided into positive samples and negative samples, while the prediction results are divided into positive results (true) and negative results (false). We use TP (true positive) and TN (true negative) to represent the correct predictions, while we use FP (false positive) and FN (false negative) to represent the wrong predictions. To evaluate the performance of the proposed model in extracting road surface and road centerline, we adopt four evaluation indicators: precision, recall, F1-score and IoU [55–57].

Precision: Precision indicates how many positive samples are predicted to be positive. It represents the proportion of labeled roads extracted from the model:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

Recall: Recall indicates the number of the correct-predicted positive examples in the sample. It represents the proportion of roads that are correctly labeled by the model:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

F1-score: F1-score, a harmonic average of the precision and recall, is a widely used index to measure the accuracy of the binary classification model:

$$\text{F1} - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

IoU: IoU is a common evaluation metrics for semantic segmentation and target detection, which measures the accuracy of the ground objects detected from the dataset and quantifies the fit degree between the extraction results and the true labels. The larger the value of IOU, the more overlapping areas of results.

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{4}$$

Since the road centerline dataset is manually combined with the semiautomatic annotation, we think that certain deviations exist between the derived centerline and the real centerline. Thus, evaluating the accuracy of road centerline with single-pixel width is not a good option. We apply a buffer width ρ as a matching area to evaluate the extraction of road centerlines, as shown in Appendix A. Note that the aforementioned four indexes are also used to evaluate the results of road centerline extraction after transforming the linear precision evaluation to the planar precision evaluation.

4.2. Implementation

All experiments adopt the same parameter initialization method and optimizer in the training process. The batch size is set to 3 with a learning rate of 1×10^{-3}. Adam function, the function with the highest accuracy stored in the validation set, is used for weight initialization. The whole training process takes about 40 h with GPU acceleration. The implementation of our model is based on TensorFlow v2.20 in Windows operating system with Intel (R) Xeon (R) CPU E5-2687 V4 @ 3.00 GHz as CPU. GPU model is NVIDIA GRID RTX8000-12Q with 12 G of memory.

4.3. Comparison of Road Surface Extraction

Figure 6 shows the visual comparison among the proposed model, i.e., MRENet, with FCN, Unet, and SegNet. We observe that incomplete extractions exist in all methods. However, in terms of road integrity, MRENet is able to obtain a more complete road structure. Table 2 records the quantitative results of the comparison among the selected networks. From these experimental indicators, MRENet achieves higher scores than other methods in terms of IOU and F1-score.

We find that all four models can effectively extract straight and clear-cut roads. When it comes to roads with bends and intersections with more shadows, however, their performances differ. MRENet is able to effectively extract the main structure of the road, preserving the details even in some challenging scenes, which proves the effectiveness and superiority of our proposed method.

We attempt to explain the above experimental results, incorporating the specific network structure. U-net and MRENet add a path between encoder and decoder, which is conducive to the transmission of high-level information and low-level information, presumably leading to better robustness. A large number of testing results reveal that the extracted results are relatively better for scenes with smaller road width. We assume that Resblock plays an important role, as it makes the receptive field expand, which is more conducive to extract semantic information among objects. In addition, the extracted information is better fused by the PSP pooling module, leading to faster convergence of the network [43].

Table 2. Experimental results of road surface extraction with different networks.

Methods	Precision	Recall	F1-Score	IoU
FCN	0.7097	0.6455	0.6761	0.5107
SegNet	0.7447	0.6650	0.7025	0.5415
Unet	0.7591	0.6688	0.7111	0.5517
Ours	0.7554	0.6771	0.7141	0.5553

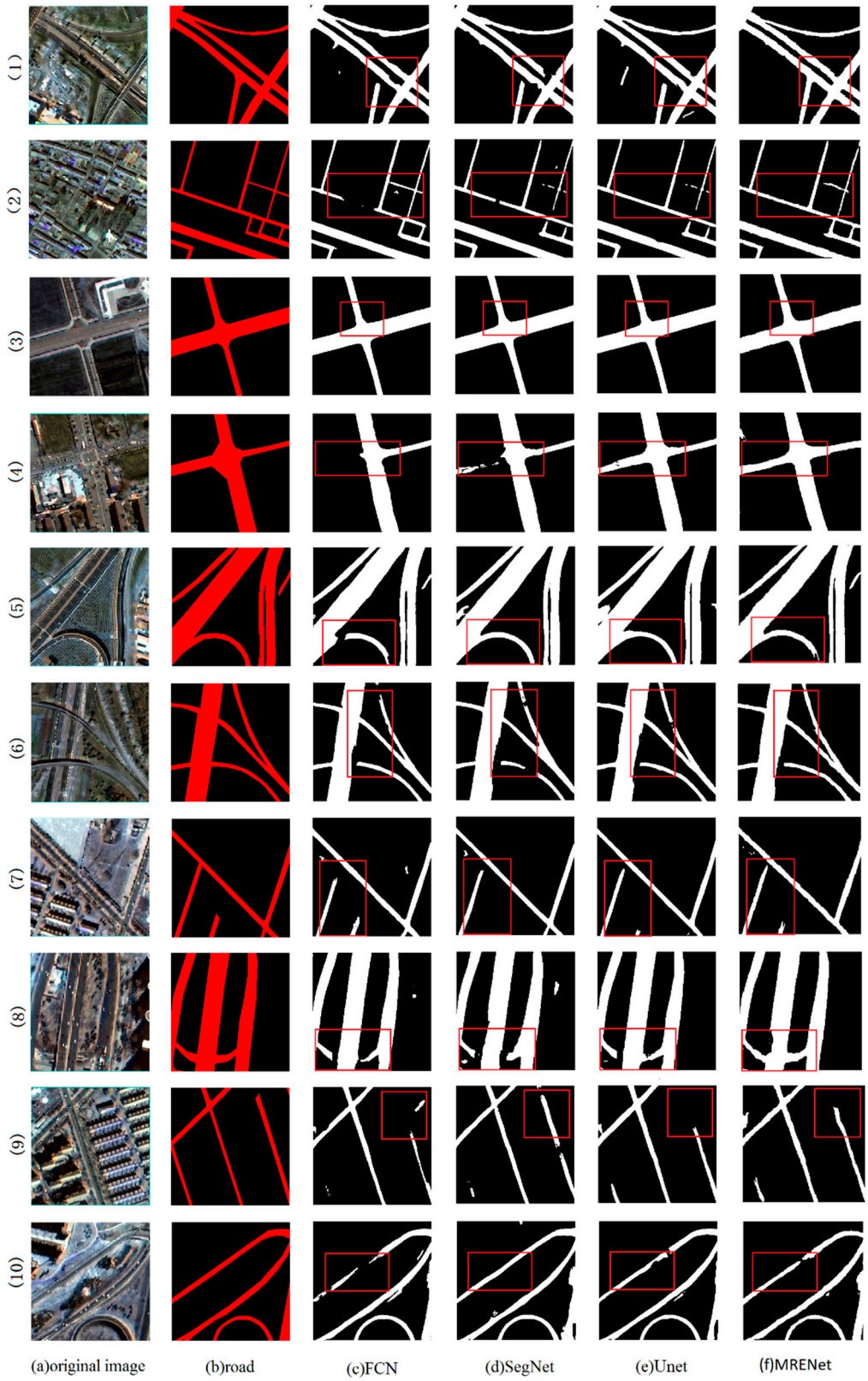

Figure 6. Visual comparisons of road surface extraction results with different networks.

4.4. Comparison of Road Centerline Extraction

As discussed in Section 4.1, we apply a buffer width ρ as a matching area to evaluate the extraction of road centerlines, given that the road centerline dataset is developed using

a semiautomatic approach. Figure 7 shows the visual comparisons of road centerline extraction results with different networks at ρ = 1. We notice that MRENet well maintains connectivity at most intersections, as the extraction results from MRENet contain considerably fewer breakpoints compared to other methods.

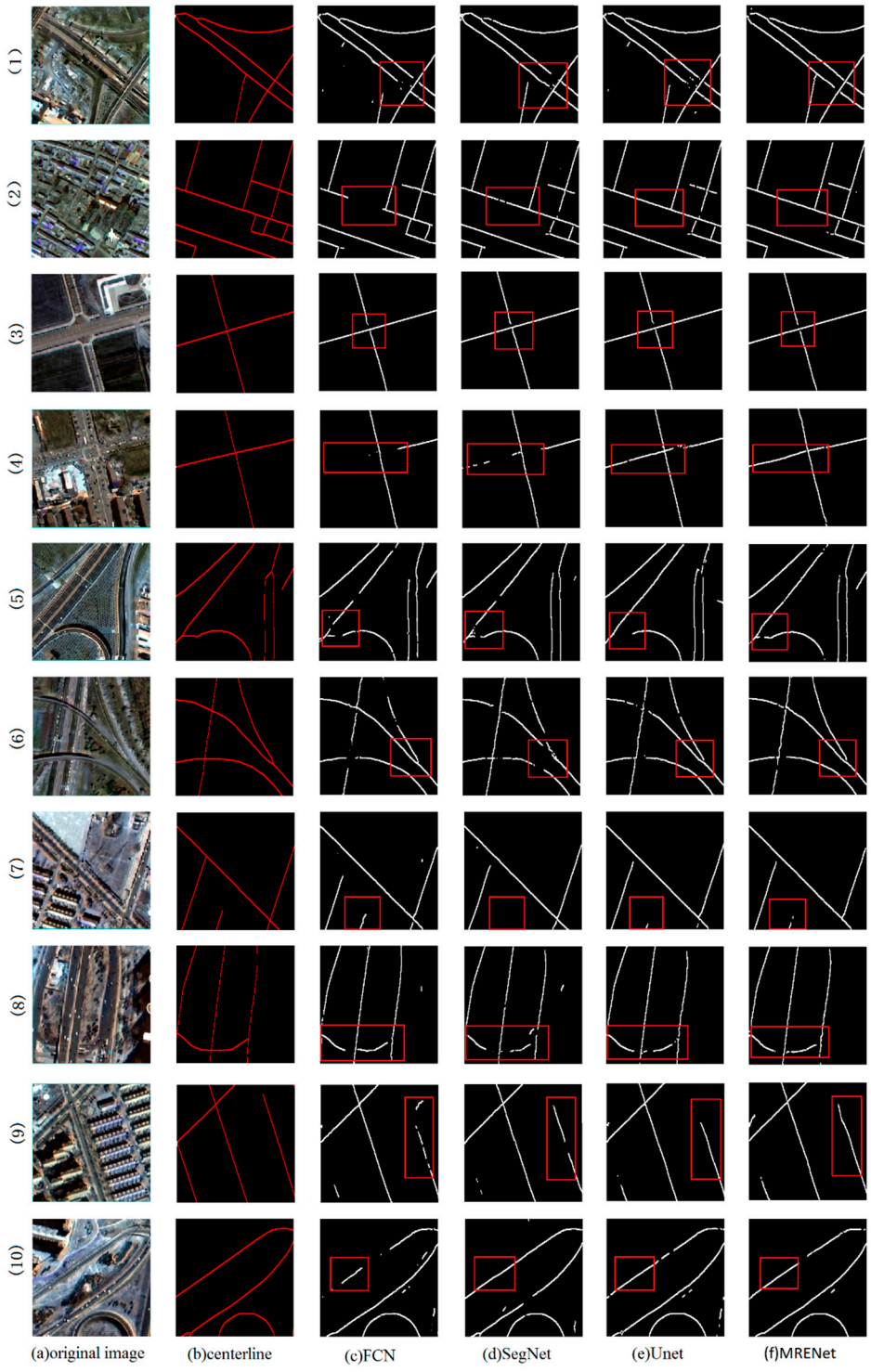

Figure 7. Visual comparisons of road centerline extraction results with different networks.

Table 3 presents the quantitative results of road centerline extraction under different ρ values. The results suggest that MRENet achieves great performance in all evaluating metrics. This phenomenon can be explained by the fact that the image features extracted

by the encoder component are shared in the latter part of the network, leading to the correlation between the two output results to a certain extent.

Table 3. Experimental results of road centerline extraction with different networks under different buffer widths (ρ).

Buffer Xidth	Methods	Precision	Recall	F1-Score	IoU
ρ = 1	FCN	0.6488	0.5727	0.6084	0.4372
	SegNet	0.7004	0.5911	0.6411	0.4718
	Unet	0.7091	0.6164	0.6595	0.4920
	Ours	0.7180	0.6160	0.6631	0.4960
ρ = 3	FCN	0.6820	0.6184	0.6486	0.4800
	SegNet	0.7250	0.6258	0.6718	0.5057
	Unet	0.7321	0.6354	0.6803	0.5155
	Ours	0.7406	0.6377	0.6853	0.5213
ρ = 5	FCN	0.7118	0.6379	0.6728	0.5070
	SegNet	0.7365	0.6465	0.6886	0.5251
	Unet	0.7427	0.6571	0.6973	0.5353
	Ours	0.7516	0.6566	0.7009	0.5395

5. Discussion

In this section, we explore the influence of different band selections and different upsampling connection locations on the experimental results.

5.1. Comparison of Different Band Selection

As the NIR band in remote sensing images is becoming more and more popular, we test the impact of this band on the results in this session. As shown in Tables 4 and 5, compared with the detection results from RGB bands, the inclusion of the NIR band further improves the model performance in both road surface extraction and road centerline extraction tasks. Despite that the reasons behind are complex, it can be inferred that, on the one hand, the NIR band makes the road information of the original data more abundant, facilitating the interpretation of remote sensing image; on the other hand, the NIR band records the spectral information of different ground objects, providing the distinguishing ability between the road from nonroad.

Table 4. Experimental results of road surface extraction with different band selection.

Bands	Precision	Recall	F1-Score	IoU
RGB	0.7526	0.6480	0.6964	0.5342
RGB + NIR	0.7554	0.6771	0.7141	0.5553

Table 5. Experimental results of road centerline extraction with different band selection under different buffer widths (ρ).

Buffer Width	Methods	Precision	Recall	F1-Score	IoU
ρ = 1	RGB	0.7070	0.5933	0.6452	0.4762
	RGB + NIR	0.7180	0.6160	0.6631	0.4960
ρ = 3	RGB	0.7290	0.6251	0.6731	0.5072
	RGB + NIR	0.7406	0.6377	0.6853	0.5213
ρ = 5	RGB	0.7398	0.6448	0.6890	0.5256
	RGB + NIR	0.7516	0.6566	0.7009	0.5395

5.2. Comparison of Different Upsampling Connection Locations

In the architecture of MRENet, the features extracted from the road encoder stage are restored to the original image size via the upsampling operation. The road surface

extraction is involved in the process of road centerline extraction to facilitate the fusion of local and global information. This design of the road centerline extraction network is to avoid that the extraction results of road centerline are completely determined by the surface extraction results. Connecting the extraction results of the first three Resblocks can compensate for the loss of some multilevel detail features in the road extraction process. We notice that upsampling connection location has an unnoticeable influence on the results in road surface extraction, while it has a trivial influence on the performance in road centerline extraction (Table 6).

Table 6. Experimental results of road centerline extraction with different connection locations for upsampling (ρ = 3).

Bands	Precision	Recall	F1-Score	IoU
MRENet_Conv	0.7144	0.6123	0.6594	0.4919
MRENet_Resblock	0.7180	0.6160	0.6631	0.4960

6. Conclusions

In this paper, we summarize the characteristics of the road surface and road centerlines in VHR remote sensing images and propose a new challenging dataset derived from VHR remote sensing images for road-related extraction to overcome the performance saturation problem of the existing benchmark datasets. We propose a two-task and end-to-end convolution neural network, termed Multitask Road-related Extraction Network (MRENet), to bridge the extraction of both road surface and road centerlines by enabling feature transferring. Through the Resblock and the PSP pooling module, the designed network can expand the receptive field and integrate multilevel features, leading to the acquisition of abundant information. Further, we explore the influence of different band selection and different upsampling connection locations on the experimental results. The experimental results show that the proposed model achieves great performance on the proposed datasets compared with other state-of-the-art methods and we believe that our model can also achieve good results in other complex urban scenes.

Author Contributions: Z.S. wrote the manuscript and designed the comparative experiments; Z.Z. designed the architecture, performed the comparative experiments and wrote the manuscript; X.H. gave comments and suggestions in writing the paper; Y.Z. supervised the study and revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Key Research and Development Program of China (2018YFB0505401), the National Natural Science Foundation of China under Grants 41890820, 41771452, 41771454, and 41901340, the Key Research and Development Program of Yunnan province in China (2018IB023), the Research Project from the Ministry of Natural Resources of China under Grant 4201·240100123, the Natural Science Fund of Hubei Province in China under Grant 2018CFA007, National Major Project on High Resolution Earth Observation System (GFZX0403260306).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: [http://www.lmars.whu.edu.cn/prof_web/shaozhenfeng/index.html].

Acknowledgments: We would like to thank the anonymous reviewers for their constructive and valuable suggestions on the earlier drafts of this manuscript.

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

Appendix A

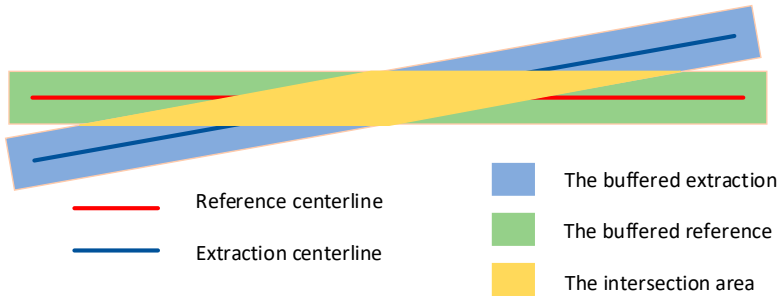

Figure A1. The evaluation of road centerline extraction via buffers.

References

1. Bajcsy, R.; Tavakoli, M. Computer Recognition of Roads from Satellite Pictures. *IEEE Trans. Syst. Man Cybern.* **1976**, *6*, 623–637. [CrossRef]
2. Tunde, A.; Adeniyi, E. Impact of Road Transport on Agricultural Development: A Nigerian Example. *Ethiop. J. Environ. Stud. Manag.* **2012**, *5*, 232–238. [CrossRef]
3. 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*, 1–11. [CrossRef] [PubMed]
4. Wang, W.; Yang, N.; Zhang, Y.; Wang, F.; Cao, T.; Eklund, P. A review of road extraction from remote sensing images. *J. Traffic Transp. Eng.* **2016**, *3*, 271–282. [CrossRef]
5. Li, M.; Stein, A.; Bijker, W.; Zhan, Q. Region-based urban road extraction from VHR satellite images using Binary Partition Tree. *Int. J. Appl. Earth Obs. Geoinf.* **2016**, *44*, 217–225. [CrossRef]
6. Zhang, Z.; Liu, Q.; Wang, Y. Road extraction by deep residual u-net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [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. Lv, X.; Ming, D.; Chen, Y.Y.; Wang, M. Very high resolution remote sensing image classification with SEEDS-CNN and scale effect analysis for superpixel CNN classification. *Int. J. Remote Sens.* **2019**, *40*, 506–531. [CrossRef]
9. Lv, X.; Ming, D.; Lu, T.; Zhou, K.; Wang, M.; Bao, H. A new method for region-based majority voting CNNs for very high resolution image classification. *Remote Sens.* **2018**, *10*, 1946. [CrossRef]
10. Li, D.; Ma, J.; Cheng, T.; van Genderen, J.L.; Shao, Z. Challenges and opportunities for the development of MEGACITIES. *Int. J. Digit. Earth* **2019**, *12*, 1382–1395. [CrossRef]
11. McKeown, D.M.; Denlinger, J.L. Cooperative methods for road tracking in aerial imagery. In Proceedings of the CVPR'88: The Computer Society Conference on Computer Vision and Pattern Recognition, Ann Arbor, MI, USA, 5–9 June 1988; pp. 662–672.
12. Zhang, J.; Lin, X.; Liu, Z.; Shen, J. Semi-automatic road tracking by template matching and distance transformation in Urban areas. *Int. J. Remote Sens.* **2011**, *32*, 8331–8347. [CrossRef]
13. Fu, G.; Zhao, H.; Li, C.; Shi, L. Road Detection from Optical Remote Sensing Imagery Using Circular Projection Matching and Tracking Strategy. *J. Indian Soc. Remote Sens.* **2013**, *41*, 819–831. [CrossRef]
14. Treash, K.; Amaratunga, K. Automatic Road Detection in Grayscale Aerial Images. *J. Comput. Civ. Eng.* **2000**, *14*, 60–69. [CrossRef]
15. Schubert, H.; van de Gronde, J.J.; Roerdink, J.B.T.M. Efficient Computation of Greyscale Path Openings. *Math. Morphol. Theory Appl.* **2016**, *1*, 189–202. [CrossRef]
16. Maboudi, M.; Amini, J.; Hahn, M.; Saati, M. Road network extraction from VHR satellite images using context aware object feature integration and tensor voting. *Remote Sens.* **2016**, *8*, 637. [CrossRef]
17. Maggiori, E.; Manterola, H.L.; del Fresno, M. Perceptual grouping by tensor voting: A comparative survey of recent approaches. *IET Comput. Vis.* **2015**, *9*, 259–277. [CrossRef]
18. Alshehhi, R.; Marpu, P.R. Hierarchical graph-based segmentation for extracting road networks from high-resolution satellite images. *ISPRS J. Photogramm. Remote Sens.* **2017**, *126*, 245–260. [CrossRef]
19. Mayer, H.; Laptev, I.; Baumgartner, A.; Steger, C. Automatic Road Extraction Based On Multi-Scale Modeling, Context, And Snakes. *Int. Arch. Photogramm. Remote Sens.* **1997**, *32*, 106–113.
20. Trinder, J.C.; Wang, Y. Automatic road extraction from aerial images. *Digit. Signal Process. Rev. J.* **1998**, *8*, 215–224. [CrossRef]
21. Dal Poz, A.P.; Zanin, R.B.; Do Vale, G.M. Automated extraction of road network from medium-and high-resolution images. *Pattern Recognit. Image Anal.* **2006**, *16*, 239–248. [CrossRef]
22. Cao, C.; Sun, Y. Automatic road centerline extraction from imagery using road GPS data. *Remote Sens.* **2014**, *6*, 9014–9033. [CrossRef]
23. Liu, W.; Zhang, Z.; Li, S.; Tao, D. Road detection by using a generalized hough transform. *Remote Sens.* **2017**, *9*, 590. [CrossRef]
24. Mnih, V.; Hinton, G.E. Learning to Detect Roads in High-Resolution Aerial Images. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 210–223.

25. Wei, Y.; Wang, Z.; Xu, M. Road Structure Refined CNN for Road Extraction in Aerial Image. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 709–713. [CrossRef]
26. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2015; pp. 234–241.
27. 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]
28. Liu, Y.; Yao, J.; Lu, X.; Xia, M.; Wang, X.; Liu, Y. RoadNet: Learning to Comprehensively Analyze Road Networks in Complex Urban Scenes from High-Resolution Remotely Sensed Images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 2043–2056. [CrossRef]
29. Lu, X.; Zhong, Y.; Zheng, Z.; Liu, Y.; Zhao, J.; Ma, A.; Yang, J. Multi-Scale and Multi-Task Deep Learning Framework for Automatic Road Extraction. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 9362–9377. [CrossRef]
30. Batra, A.; Singh, S.; Pang, G.; Basu, S.; Jawahar, C.V.; Paluri, M. Improved road connectivity by joint learning of orientation and segmentation. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 10377–10385.
31. Qi, J.; Tao, C.; Wang, H.; Tang, Y.; Cui, Z. Spatial Information Inference Net: Road Extraction Using Road-Specific Contextual Information. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Yokohama, Japan, 28 July–2 August 2019; pp. 9478–9481.
32. Zhang, X.; Han, X.; Li, C.; Tang, X.; Zhou, H.; Jiao, L. Aerial image road extraction based on an improved generative adversarial network. *Remote Sens.* **2019**, *11*, 930. [CrossRef]
33. Liu, B.; Wu, H.; Wang, Y.; Liu, W. Main road extraction from ZY-3 grayscale imagery based on directional mathematical morphology and VGI prior knowledge in Urban areas. *PLoS ONE* **2015**, *10*, e0138071. [CrossRef]
34. Amini, J.; Saradjian, M.R.; Blais, J.A.R.; Lucas, C.; Azizi, A. Automatic road-side extraction from large scale imagemaps. *Int. J. Appl. Earth Obs. Geoinf.* **2002**, *4*, 95–107. [CrossRef]
35. Zheng, S.; Liu, J.; Shi, W.Z.; Zhu, G.X. Road central contour extraction from high resolution satellite image using tensor voting framework. In Proceedings of the 5th International Conference on Machine Learning and Cybernetics, Dalian, China, 13–16 August 2006; pp. 3248–3253.
36. Miao, Z.; Shi, W.; Zhang, H.; Wang, X. Road Centerline Extraction From High-Resolution Imagery Based on Shape Features and Multivariate Adaptive Regression Splines. *IEEE Geosci. Remote Sens. Lett.* **2013**, *10*, 583–587. [CrossRef]
37. Cheng, G.; Zhu, F.; Xiang, S.; Wang, Y.; Pan, C. Accurate urban road centerline extraction from VHR imagery via multiscale segmentation and tensor voting. *Neurocomputing* **2016**, *205*, 407–420. [CrossRef]
38. Cheng, G.; Zhu, F.; Xiang, S.; Pan, C. Road Centerline Extraction via Semisupervised Segmentation and Multidirection Nonmaximum Suppression. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 545–549. [CrossRef]
39. Gao, L.; Shi, W.; Miao, Z.; Lv, Z. Method based on edge constraint and fast marching for road centerline extraction from very high-resolution remote sensing images. *Remote Sens.* **2018**, *10*, 900. [CrossRef]
40. Zhou, T.; Sun, C.; Fu, H. Road information extraction from high-resolution remote sensing images based on road reconstruction. *Remote Sens.* **2019**, *11*, 79. [CrossRef]
41. Yujun, W.; Xiangyun, H.; Jinqi, G. End-to-end road centerline extraction via learning a confidence map. In Proceedings of the 2018 10th IAPR Workshop on Pattern Recognition in Remote Sensing, PRRS 2018, Beijing, China, 19–20 August 2018; pp. 1–5.
42. Zhang, Y.; Xiong, Z.; Zang, Y.; Wang, C.; Li, J.; Li, X. Topology-aware road network extraction via Multi-supervised Generative Adversarial Networks. *Remote Sens.* **2019**, *11*, 1017. [CrossRef]
43. Diakogiannis, F.I.; Waldner, F.; Caccetta, P.; Wu, C. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 94–114. [CrossRef]
44. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.
45. Zhang, Y.; Yuan, Y.; Feng, Y.; Lu, X. Hierarchical and Robust Convolutional Neural Network for Very High-Resolution Remote Sensing Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5535–5548. [CrossRef]
46. Zhang, R.; Shao, Z.; Huang, X.; Wang, J.; Li, D. Object detection in UAV images via global density fused convolutional network. *Remote Sens.* **2020**, *12*, 3140. [CrossRef]
47. Wang, C.; Bai, X.; Wang, S.; Zhou, J.; Ren, P. Multiscale Visual Attention Networks for Object Detection in VHR Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 310–314. [CrossRef]
48. Xia, G.S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [CrossRef]
49. Hu, F.; Xia, G.S.; Yang, W.; Zhang, L. Recent advances and opportunities in scene classification of aerial images with deep models. In Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS), Valencia, Spain, 22–27 July 2018; pp. 4371–4374.
50. Shao, Z.; Tang, P.; Wang, Z.; Saleem, N.; Yam, S.; Sommai, C. BRRNet: A fully convolutional neural network for automatic building extraction from high-resolution remote sensing images. *Remote Sens.* **2020**, *12*, 1050. [CrossRef]
51. 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. 6230–6239. [CrossRef]

52. Shao, Z.; Pan, Y.; Diao, C.; Cai, J. Cloud detection in remote sensing images based on multiscale features-convolutional neural network. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 4062–4076. [CrossRef]
53. Qi, K.; Liu, W.; Yang, C.; Guan, Q.; Wu, H. Multi-task joint sparse and low-rank representation for the scene classification of high-resolution remote sensing image. *Remote Sens.* **2017**, *9*, 10. [CrossRef]
54. Caruana, R.; Mitchell, T.; Pomerleau, D.; Dietterich, T.; State, O. Multitask Learning. *Mach. Learn.* **1997**, *28*, 41–75. [CrossRef]
55. Davis, J.; Goadrich, M. The relationship between precision-recall and ROC curves. In Proceedings of the ICML 2006: 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 233–240. [CrossRef]
56. Khryashchev, V.; Larionov, R. Wildfire Segmentation on Satellite Images using Deep Learning. In Proceedings of the 2020 Moscow Workshop on Electronic and Networking Technologies, MWENT 2020, Moscow, Russia, 11–13 March 2020; pp. 1–4. [CrossRef]
57. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.