applied sciences

MDPI

*Article*

# MSPNet: Multi-Scale Strip Pooling Network for Road Extraction from Remote Sensing Images

**Shenming Qu [1,2,3], Huafei Zhou [1,2,3], Bo Zhang [1] and Shengbin Liang [1,2,3,*]**

[1] School of Software, Henan University, Kaifeng 475001, China; qsm@vip.henu.edu.cn (S.Q.); 104754201764@henu.edu.cn (H.Z.); 1925060181@henu.edu.cn (B.Z.)
[2] Institute of Intelligence Networks System, Henan University, Kaifeng 475001, China
[3] Intelligent Data Processing Engineering Research Center of Henan Province, Kaifeng 475001, China
* Correspondence: liangsbin@henu.edu.cn; Tel.: +86-150-9316-2629

**Abstract:** Extracting roads from remote sensing images can support a range of geo-information applications. However, it is challenging due to factors such as the complex distribution of ground objects and occlusion of buildings, trees, shadows, etc. Pixel-wise classification often fails to predict road connectivity and thus produces fragmented road segments. In this paper, we propose a multi-scale strip pooling network (MSPNet) to learn the linear features of roads. Motivated by the strip pooling being more aligned with the shape of roads, which are long-span and narrow, we develop a multi-scale strip pooling (MSP) module that utilizes strip pooling layers with long but narrow kernel shapes to capture multi-scale long-range context from horizontal and vertical directions. The proposed MSP module focuses on establishing relationships along the road region to guarantee the connectivity of roads. Considering the complex distribution of ground objects, the spatial pyramid pooling is applied to enhance the learning ability of complex features in different sub-regions. In addition, to alleviate the problem caused by an imbalanced distribution of road and non-road pixels, we use binary cross-entropy and dice-coefficient loss functions to jointly train our proposed deep learning model. Then, we perform ablation experiments to adjust the loss contributions to suit the task of road extraction. Comparative experiments on a popular benchmark DeepGlobe dataset demonstrate that our proposed MSPNet establishes new competitive results in both IoU and F1-score.

**Keywords:** road extraction; deep learning; strip pooling; remote sensing images; spatial pyramid pooling



**Citation:** Qu, S.; Zhou, H.; Zhang, B.; Liang, S. MSPNet: Multi-Scale Strip Pooling Network for Road Extraction from Remote Sensing Images. *Appl. Sci.* **2022**, *12*, 4068. https://doi.org/10.3390/app12084068

Academic Editors: Andrea Prati, Luis Javier García Villalba and Vincent A. Cicirello

Received: 1 March 2022
Accepted: 13 April 2022
Published: 18 April 2022



## 1. Introduction

The automatic extraction of roads maps from very high-resolution remote sensing images is an essential and hot research domain, which can be applied to numerous applications that rely on the efficient and real-time updating of road maps, such as navigation, cartography, urban planning, location-based mobile services and autonomous driving. In disaster zones, especially in developing countries, maps and accessibility information are crucial for crisis response. Extracting roads from RSIs is a promising approach and has been studied for decades. Recently, the wide use of convolutional neural networks (CNNs) [1], especially networks with fully-convolutional network (FCN) [2] architecture, has greatly improved the accuracy of road extraction and made the task end-to-end trainable [3]. However, the existing extraction results of road maps are still not satisfactory, which is mainly due to the complex urban traffic environments and special characteristics of roads. Compared with most other ground natural objects with bulk shape, such as buildings and trees, the roads in remote sensing images are narrow, long-span, and can be broadly formulated as elongated regions with similar spectral and texture patterns. Therefore, the road extraction algorithms often produce fragmented road segments leading to road network disconnection due to the occlusion of trees, buildings, cloud, etc. Additionally, the similarities often exist between the roads and other ground objects that are difficult

to identify, as they look visually similar to targets. To match those features' geometric and physical features, the road extraction methods are expected to have a certain level of optimization of the results to reduce the missing connections and false alarms [4].

Conventional expert-knowledge-based methods, such as knowledge-driven-based, template-matching-based, and object-oriented-based, are mainly adopt to extract roads from remote sensing images [5,6]. Those hand-crafted methods are often cumbersome in steps and lead to error accumulation problems, since they usually combine multiple algorithms to match the complex features of roads. The fixed hand-crafted criteria based on geometric and physical features are usually unsuitable and inefficient to large volumes of remote sensing data, and they also cannot guarantee the connectivity of roads. With the rapid development of deep learning technologies, the CNN-based methods make it possible to obtain the results of large-scale remote sensing data [4]. Accordingly, the CNN-based methods improve the accuracy of road extraction significantly compared with conventional methods and have been the mainstream in road extraction due to the great feature learning power of CNNs [7].

Road extraction can be viewed as a binary semantic segmentation problem in CNN-based methods. Several works have been proposed and applied successfully for road segmentation tasks, especially some with an encoder–decoder architecture. In [8], a Dense-UNet model with similar architecture of encoder–decoder and dense connection units is proposed to extract the road network from remote sensing images. Combining the deep residual network, pyramid pooling module, and deep decoder, Han et al. [9] propose a novel deep residual and pyramid pooling network (DRPPNet) for extracting road regions from high-resolution remote sensing images. A dual-attention capsule U-Net (DA-CapsUNet) in [10] is designed for road maps extraction by combining the properties of capsule representations and the features of attention mechanisms. Some other typical works based on modified encoder–decoder structure include [4,11]. However, there are still challenges in accurately extracting road maps, which are mainly due to the special characteries of thin and elongated roads that are easily interpreted by tress, buildings and shadows, etc. These difficulties lead to fragmentation of the road segmentation, and the above methods cannot guarantee the connectivity of roads. A possible solution to these problems is to enhance the embedding of linear features within the CNN architectures.

In this paper, we propose a multi-scale strip pooling network (MSPNet) to address the above-mentioned problems. We first introduce an encoder–decoder architecture network to learn the feature of roads, where the Pyramid Pooling module (PPM) is adopted to increase the receptive field of feature points and learning ability of complex features in different sub-regions. Inspired by the strip pooling being more aligned with the shapes of roads that are long-span, narrow, and distributed continuously, we propose a multi-scale strip pooling (MSP) module to learn the linear features of roads, which is placed in the skip-connection paths. The MSP focuses on establishing relationships in the elongated road region between road and occluded road pixels to guarantee the connectivity of roads. Extensive experiments on popular benchmarks in terms of several metrics demonstrate the superiority of our MSPNet compared with several state-of-the-art methods.

The main contributions of this work are summarized as follows.

- We propose an end-to-end multi-scale strip pooling network (MSPNet) with symmetric encoder–decoder network design for the task of road extraction. This network design can preserve spatial detailed information and therefore optimize the smoothness of roads. In addition, it is also suitable for processing large-scale images.
- We develop a multi-scale strip pooling (MSP) module that utilizes strip pooling layers to aggregate multiple long-range contextual information. The linear features of roads are enhanced within CNN architecture, which thus improves the road connectivity.
- Ablation studies and comparative experiments on a benchmark DeepGlobe data set are performed to verify the effectiveness of our proposed MSPNet.

The remaining of this article is organized as follows. Section 2 introduces the related works of road extraction. In Section 3, we describe datasets, evaluation metrics, and

implementation details, and we illustrate our proposed MSPNet in detail. Extensive experiments are performed to evaluate the performance of the proposed method for road extraction in Section 5. The conclusion and discussion are presented in Section 6.

## 2. Related Work

The literature research on automatic road extraction can be divided into two categories: expert knowledge-based methods and CNN-based ones. Although the CNN-based methods improve the accuracy significantly due to the powerful feature-embedding abilities, traditional methods on how to utilize the geometric and physical properties of roads provide inspiration for future research. In this section, we briefly introduce these two categories of methods.

### 2.1. Expert Knowledge-Based Methods

Conventional expert knowledge-based algorithms for road extraction usually utilize geometric and physical features to match. Fu et al. [12] propose a road detection method based on a Circular Projection (CP) matching and tracking strategy, which is beneficial for twisty roads detection. Xu et al. [5] present a morphological method by combining the automatic thresholding and morphological operation techniques to extract roads from remote sensing images. Wang et al. [13] propose an automatic road extraction method for vague aerial images with an improved Canny edge detection operator and Hough line transform algorithm. Herumurti et al. [14] propose a road extraction based on zebra crossings detection. Song et al. [15] develop an approach for road extraction utilizing pixel spectral information for classification and image segmentation-derived object features. However, these methods based on hand-designed road features are generally inefficient and unsuitable for large-scale remote sensing data.

### 2.2. CNN-Based Methods

*(1) Segmentation of Roads:* Semantic segmentation is a basic and essential research domain in computer vision. With the great success of deep learning in the field of semantic segmentation, some studies [3,16,17] consider the road extraction as a binary semantic segmentation problem using CNN-based approaches. Mnih et al. [18] firstly present a neural network-based approach with restricted Boltzmann machine (RBM) for detecting roads in high-resolution aerial images. Some deep learning models with encoder–decoder structures such as UNet [19] and LinkNet [20] have been proven to be efficient in the field of semantic segmentation, and their variants have also been widely proposed to segment roads [8,10,21]. A semantic segmentation neural network, which combines the strengths of residual learning and U-Net, is proposed for road area extraction by Zhang et al. [22]. Zhou et al. [23] follow the LinkNet architecture and employ dilated convolution layers with both cascade mode and parallel mode to enlarge the receptive field. Zhou et al. [24] propose an HsgNet, which inserts a Middle Block based on bilinear pooling into the middle of LinkNet between the encoder and decoder. These methods usually perform better compared with traditional methods, but they cannot guarantee the connectivity of roads and thus produce fragmented road segments.

*(2) Connectivity of Roads:* Recently, several works are proposed to obtain segmentation results with better road connectivity. Li et al. [25] put forward a road extraction method based on a LinkNet deep learning model, and at the pre-processing step, an auxiliary constraint task is designed to solve the connectivity problem caused by occlusions. Zhou et al. [21] propose a fusion network (FuNet) with the fusion of remote sensing imagery and location data, and a universal iteration reinforcement (IteR) module is added to enhance the ability of road connectivity reasoning. Meanwhile, Zhang et al. [26] introduce a deep learning-based multistage framework to extract the road surface and road centerline simultaneously. They initially segment roads with an FCN-based model, after which an iterative search strategy is applied to track consecutive and complete road networks. However, the iterative steps are time-consuming. The authors of [6] employ a novel

linearity index for the discrimination of elongated road segments from other objects and customized tensor voting, which is utilized to fill missing parts of the road network. In these approaches, pre-processing or post-processing is added to maintain the connectivity of roads. However, they are time consuming and not suitable for some areas with high road density and occlusions.

In conclusion, although the CNN-based methods have greatly improved the accuracy of road extraction, most of them are simple extensions of the widely used CNN architectures and do not consider the structural features of roads. Therefore, there are still margins to improve the accuracy of road extraction in terms of topological connectivity.

## 3. Materials and Methods

### *3.1. Dataset*

A public road extraction dataset DeepGlobe [27] is applied for evaluating the performance of the proposed method. The dataset provides images with a pixel size of 1024 × 1024 and a pixel resolution of 50 cm/pixel which includes multiple scenes such as cities, villages, wild suburbs, seashores, tropical rainforests, etc. The dataset contains 6226 images and corresponding annotated ground truth labels. In this paper, we divide it into 4696 images for training and 1530 for testing following [4]. Some samples are shown in Figure 1.

**Figure 1.** Examples of the original images and coressponding ground truth. The white represents the road pixels and the black represents the non-road pixels.

Data enhancement is a common and useful strategy, which can enhance the generalizability of deep learning models. In this paper, data enhancement is applied by flipping and shifting the images randomly with a probability of 0.5. The visualization of data enhancement is shown in Figure 2.

(**a**) (**b**) (**c**)

**Figure 2.** Samples of data augmentation adopted in this paper. (**a**) is the original image; (**b**) is the result after shifting; (**c**) is the result after flipping.

### *3.2. Evaluation Metrics*

In this paper, to evaluate the performance of our proposed method and other methods, we use several metrics of overall accuracy (OA), Intersection over Union (IoU), Mean Intersection over Union (MIoU), and F1-score. These are the most widely used measurements in both road extraction and other segmentation tasks [28]. They are defined by:

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{2}$$

$$F1\text{-}score = \frac{2 \times precision \times recall}{precision + recall} \tag{3}$$

in which,

$$precision = \frac{TP}{TP + FP}, recall = \frac{TP}{TP + FN} \tag{4}$$

where $TP$, $FP$, $TN$ and $FN$ are the true positive, false positive, true negative and false negative, respectively.

### 3.3. Network Structure

We propose a multi-scale strip pooling network (MSPNet) for road extraction from remote sensing images as illustrated in Figure 3. Encoder–decoder networks are applied to many computer vision tasks, and their superiority has also been validated [19,29]. Therefore, we apply one as the overall architecture of our proposed MSPNet. The ResNet [30] is widely used in image recognition because of its outstanding performance for feature learning. The ResNet contains a series of residual neural network models, which have different numbers of layers. The CNN-based model with more layers generally improves the performance, while it requires a higher computational cost and training time. We will provide analysis of the performance and parameter cost for a ResNet series network in the next subsection. Here, we employ ResNet-34 pre-trained on ImageNet [31] as the encoder.

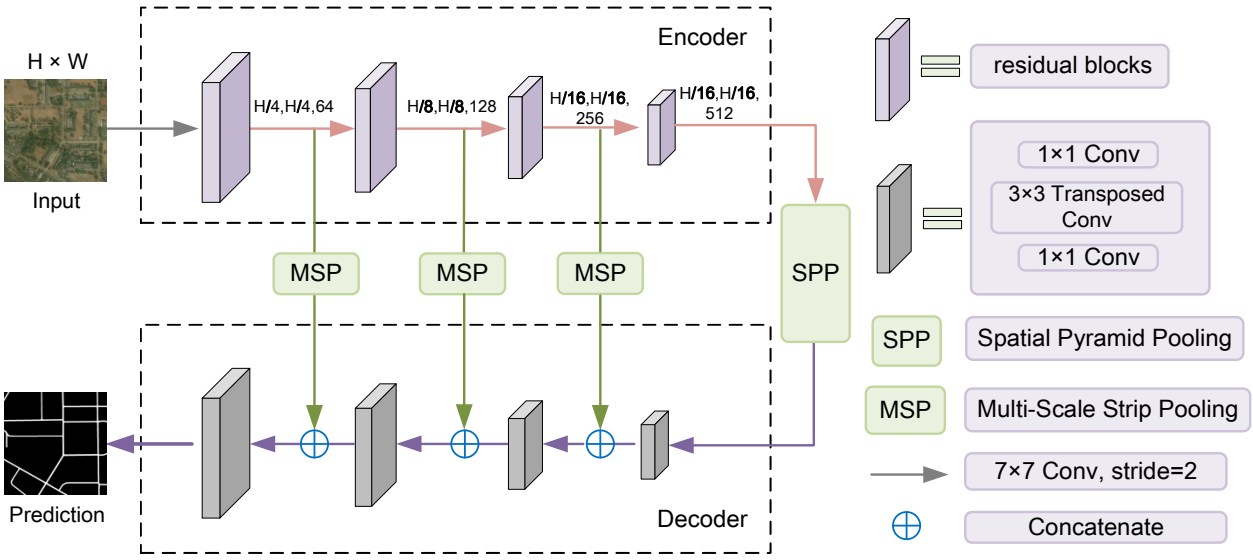


**Figure 3.** Overall architecture of the proposed multi-scale strip pooling network (MSPNet). The encoder module contains four residual blocks, the MSP module is applied to extracting linear features of the roads, and the PPM module is used to learn complex features from different scales. The decoder module is applied for gradually upsampling the resolution of the output feature map to get the final per-pixel prediction.

The roads in remote sensing images are narrow, long span, and straight distribution that often produce fragmented road segments. To improve the connectivity of roads, our proposed multi-scale strip pooling (MSP) module is sequentially added to the first three skip-connection paths to extract linear features of roads at multiple scales. Different from the roads, most background objects have bulk shapes. Considering the advantages of traditional pooling, we adopt the Pyramid Pooling Module (PPM) [32] to effectively learn complex features that are added to the last skip-connection path. Due to the complex distribution of roads and other objects, the use of PPM will enhance the learning ability

of complex features and thus improve the performance. In the decoder module, we apply transpose convolution for upsampling the feature maps to an appropriate size and 1 × 1 convolution for adjusting the channels of feature maps; then, we concatenated with the corresponding output feature map of the MSP module in the skip-connection path.

### *3.4. Multi-Scale Strip Pooling*

Most natural objects often have bulk shapes. Accordingly, the traditional kernel shape of the pooling layer in most CNN architectures is designed to be square for feature learning, which is suitable for most computer vision tasks. However, the roads in remote sensing images are narrow, long-span and can be described as elongated areas. Traditional pooling layers with square kernel shapes neglect the modeling of linear features of roads. By contrast, the strip pooling is more in line with the shape of the roads, which utilizes a long but narrow kernel to capture long-range dependencies in road regions and thus enhance the embedding of linear features within CNN models. The strengthened learning ability of linear features is helpful for retaining the connectivity and making the segmentation results more complete.

Motivated by the above fact and the advantages of strip pooling, we develop a multi-scale strip pooling module (MSP) to help our MSPNet generate road networks with better connectivity. As shown in Figure 4, MSP uses multiple strip pooling layers with long but narrow kernel shapes to capture multi-scale long-range context from horizontal and vertical directions. In addition, the two directions we choose are also aligned with the distribution of most roads in remote sensing images.

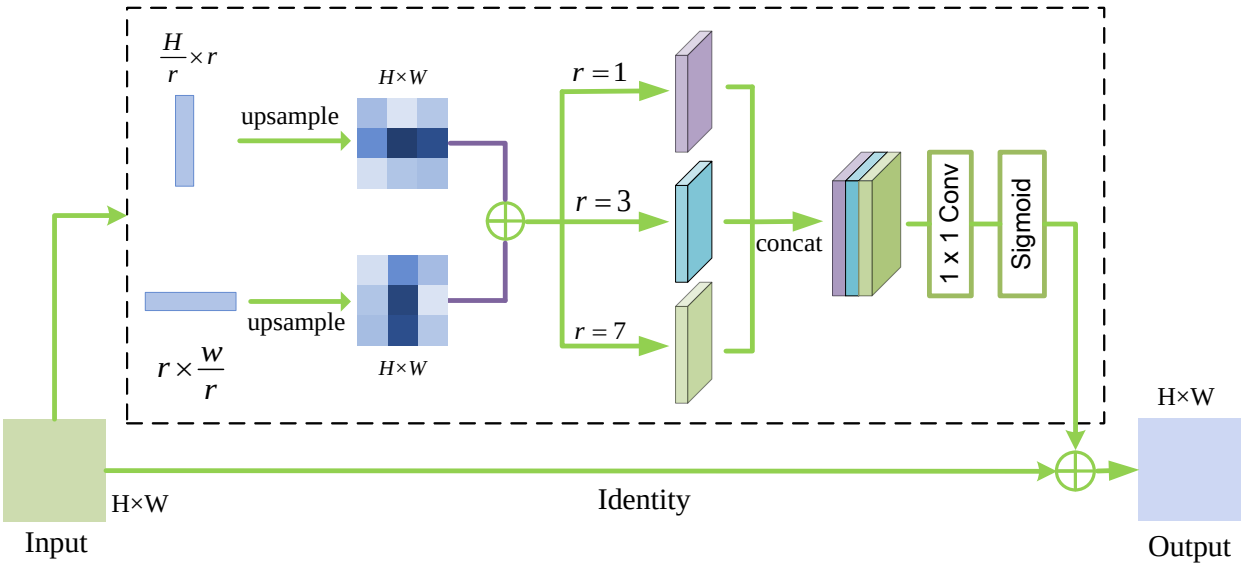


**Figure 4.** Multi-Scale Strip Pooling (MSP) Module.

Let $X \in \mathbb{R}^{H \times W}$ denote the input tensor for the MSP module, where $H$, $W$ represent the height and width. In the MSP module, $X$ is first fed into two pathways along either the horizontal or vertical spatial dimension, each of which contains a strip pooling layer with long but narrow kernel shapes of $\frac{H}{r} \times r$ or $r \times \frac{w}{r}$ to extract linear features of roads, where $r$ is the scaling factor for adjusting the kernel sizes. Let $y_r^h$ and $y_r^v$ be the output feature maps extracted by the two strip pooling layers along the horizontal or vertical direction. Then, we upsample the two feature maps to the same size of input tensor by using a bilinear upsampling layer. Afterwards, we combine the two feature maps of $y_r^h$ and $y_r^v$ to obtain $y_r$, which can be formulated as:

$$y_r = y_r^v + y_r^h \tag{5}$$

The feature maps of $y_r$ contain rich long-range contextual information of roads with different scales, which is related to the scaling factor $r$. The appropriate selection of the scaling factor $r$ is usually based on experience. In this paper, $r$ is set to 1, 3 and 7, respectively, to obtain three feature maps, containing information with three different scales. Like [32], different scales of features are concatenated as the final pooling global feature, which can be formulated as:

$$y = Concat(y_{r=1}, y_{r=3}, y_{r=7}) \tag{6}$$

Finally, the output of the MSP module can be written as:

$$Z = Scale(x, \alpha(f(y))) \tag{7}$$

where $Scale(\cdot, \cdot)$ represents element-wise multiplication, $\alpha$ is the sigmoid function and $f$ is a $1 \times 1$ convolution.

### 3.5. Loss Function

Road extraction can be formulated as a pixel-wise binary classification task in semantic segmentation. Cross-entropy is defined as a measure of the differences between two probability distributions for a given random variable. In the deep learning domain, the binary cross-entropy (BCE) loss function is used to optimize models in binary classification tasks. Assuming that the size of the input image is $H \times W$, then the BCE loss function is calculated as follows:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log p_i + (1 - y_i) \cdot \log(1 - p_i)] \tag{8}$$

where $y_i$ is the ground truth denoting road or background for a given pixel in position $i$, $p_i$ is the corresponding probability predicted by the model and $N = H \times W$. The BCE loss function separately evaluates the predicted classes of each pixel and then averages all pixels, so it can be considered that all pixels are learned equally. Thus, it is difficult to learn the features of road pixels when there is a great imbalance in which there are far fewer road pixels than background pixels. The dice-coefficient loss function is introduced to alleviate the above problem caused by sample imbalance. Compared with the BCE loss function, the dice-coefficient loss function directly supervises the similarity of prediction and ground truth [33], which can be calculated as follows:

$$L_{Dice} = 1 - \frac{1}{N}\sum_{i=1}^{i=N}\frac{2 \cdot y_i \cdot p_i}{y_i + p_i} \tag{9}$$

Due to the imbalance of road and non-road pixels, a simple combination of binary cross-entropy (BCE) and dice-coefficient (Dice) loss functions is used to train deep learning models to alleviate this problem in previous work [11,23], which can be defined as:

$$L_{loss} = L_{BCE} + L_{Dice} \tag{10}$$

This simple combination can be considered to have the same weight, which may lead to suboptimal training results in road extraction tasks. Therefore, the final loss function used in this paper is modified as follows:

$$L_{loss} = K \cdot L_{BCE} + (1 - K) \cdot L_{Dice} \tag{11}$$

where the adjustment factor $K$ is set to balance the loss contribution of the BCE and dice loss function.

## 4. Results

### *4.1. Implementation Details*

The proposed method is implemented on the PyTorch machine learning framework and is trained on two NVIDIA GeForce RTX 2080 Ti GPUs with 11 GB memory. The source code will be made available at: https://github.com/Shenming-Qu/MSPNet (accessed on 25 March 2022).

During the experiment, due to the limitation of GPU memory size, the batch size is set to 8. Following most previous works [34,35], we adopt stochastic gradient descent (SGD) with momentum as the optimizer, and the parameters for SGD are set as follows: weight decay is set to 0.0005, and momentum is set to 0.9. We adopt the "poly" learning rate policy ($base\ learning\ rate \times (1 - \frac{iter}{max\_iter}^{power})$ to gradually reduce the learning rate, where the *base learning rate* is set to 0.005 and *power* is set to 0.9. The number of training epoch is set to 200 by default. Finally, we save the trained model for testing the performance on the test set.

### *4.2. Ablation Experiment*

#### 4.2.1. Comparison of Backbone Networks

This subsection compares the performance and parameters of ResNet series networks as the encoder in the proposed model, with the purpose of selecting the appropriate ResNet model for subsequent research.

Figure 5 plots the progression of F1-score values when ResNet series models are used as the backbone of the encoder during the training process. Table 1 provides summary statistics for performance evaluated with the metrics of F1-score and the number of parameters. Through experiments, it is found that the performance of the network improves with the increase of parameters. The ResNet with 101 parameter layers performs best in terms of F1-score, 85.14% obtained, while the Resnet-18 performs the worst for its lowest number of parameter layers, which is 3.35% lower than the former. Sequential performance was achieved by ResNet-50 and ResNet-34, reaching F1-scores of 84.51% and 84.92%, respectively. The above four models have similar results except for ResNet-18. However, it cannot be ignored that ResNet-50 and ResNet-101 have several times the number of parameters compared with ResNet-34, while it gives trivial performance gain. Since road extraction is a simple binary segmentation problem that does not need to model complex background information, the parameters in ResNet-34 are powerful enough for road extraction tasks. Finally, we choose the ResNet34 model as the backbone of the encoder based on performance and parameter considerations.

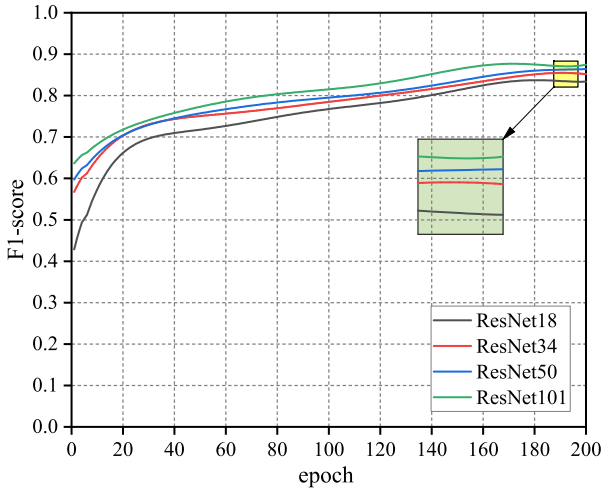


**Figure 5.** Progression of F1-score values for four ResNet series models as the backbone of the encoder during training. The ResNet models are with 18 layers (ResNet18), 34 layers (ResNet34), 50 layers (ResNet50) and 101 layers (ResNet101).

**Table 1.** Performance and parameters comparison of ResNet series models.

| Settings | #Params | F1-Score |
|---|---|---|
| Backbone (with ResNet18) | 80 M | 81.79% |
| Backbone (with ResNet34) | 118 M | 84.51% |
| Backbone (with ResNet50) | 830 M | 84.92% |
| Backbone (with ResNet101) | 902 M | 85.14% |

#### 4.2.2. Influence of Hyper-Parameter $K$

To alleviate the problem caused by imbalance between road and background pixels as described in Section 3.5, the loss function used to optimize our model contains a manually selected hyper-parameter $K$, which is used to balance the loss contribution of BCE and dice loss function. Therefore, choosing an appropriate hyper-parameter $K$ may be beneficial to reach the local optimum quickly and improve performance. We train our proposed model with different values of hyper-parameter $K$ in ascending orders, while other conditions are maintained the same to select an optimal $K$. Considering representativeness and experiment quantity, $K$ is set to 0, 0.2, 0.4, 0.6, 0.8 and 1, respectively, to observe the performances in this paper.

The experiment results with different configurations of hyper-parameter $K$ are shown in Figure 6. It can be seen that the F1-score and MIoU are fluctuating as $K$ increases. When $K = 2$, it achieves a best MIoU score of 86.74% and F1-score of 84.51%. It is noted that the performance is reduced to 82.28% in terms of F1-score with $K = 0$ and 79.41% with $K = 1$; the comparison results show that the separate use of BCE ($K = 0$) or dice ($K = 0$) loss function cannot obtain the optimal results for road extraction. As listed in Table 2, the performance with $K = 0.2$ is also better than the simple combination of "BCE + Dice", which obtains improvements of 1.48% on the MIoU score and 0.75% on the F1-score.

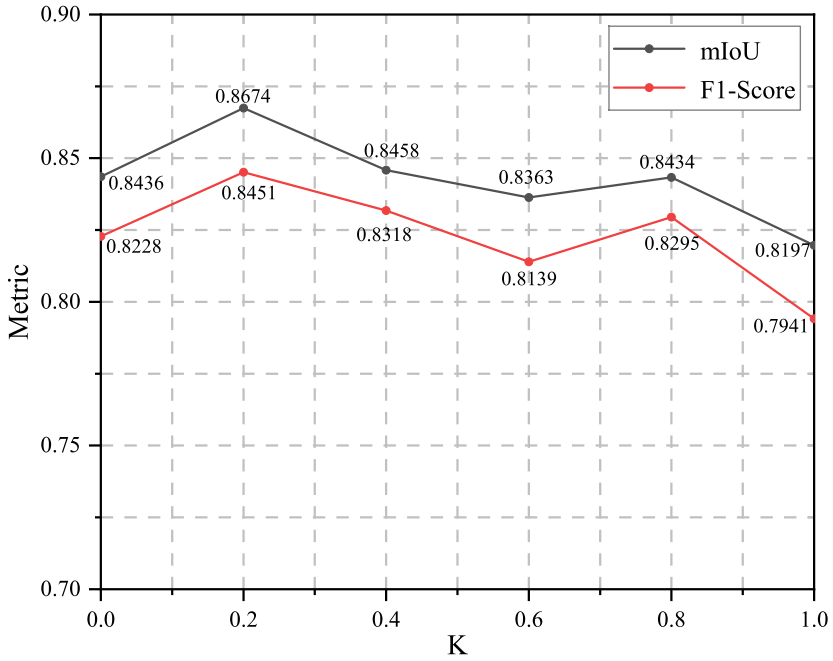


**Figure 6.** Comparison experiments of different values of the hyper-parameter $K$ of the proposed MSPNet, evaluated by F1-score and MIoU metrics.

**Table 2.** Performance and parameters comparison of ResNet series models.

| Settings | MIoU | F1-Score |
|---|---|---|
| BCE + Dice | 85.26% | 83.76% |
| $(1 - K$ [a]$)$ BCE + $K$ [a] Dice | 86.74% | 84.51% |

[a] The hyper-parameter $K$ is set to 0.2 in this experiment.

To further illustrate the influences of different weight combinations between dice and BCE loss function, we plot the changes in the loss value during training with respect to the number of epochs, as shown in Figure 7. It can be seen that the loss function curve with $K = 0.2$ is smoother than another, which indicates that there is a reasonable allocation of loss weights. To better optimize our model, we adopt $K = 0.2$ in the method evaluations.

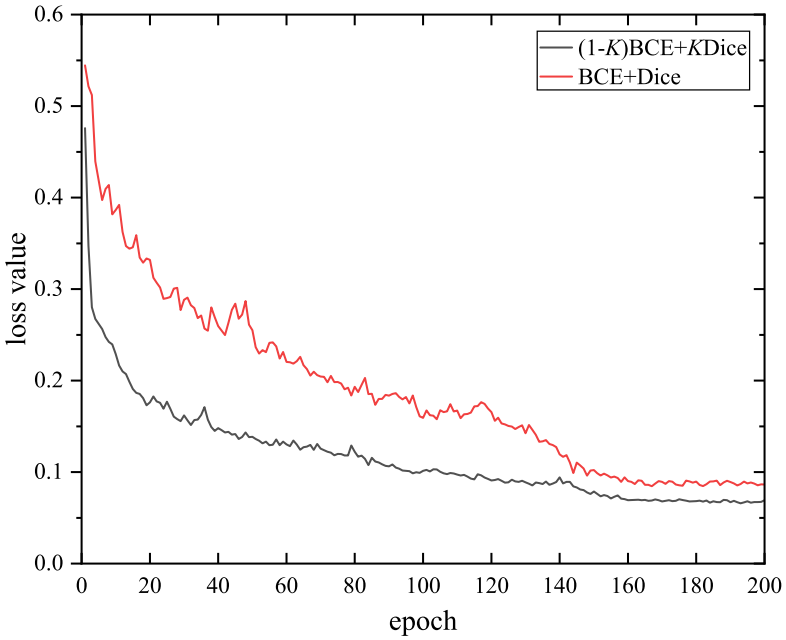


**Figure 7.** Progression of loss values in the training process. $K$ is set to 0.2 in this comparison experiment.

### *4.3. Comparison with State-of-the-Art Methods*

To evaluate the performance of the proposed model for road extraction from remote sensing images, we compared with several baseline and state-of-the-art methods: FCN [2], ResUNet [22], D-LinkNet [23], and SE-DeepLab [36]. ResUNet is built with residual learning and UNet; D-LinkNet is a variant of LinkNet architecture and added dilated convolution layers in the center part, which has achieved best performance in the CVPR DeepGlobe 2018 Road Extraction Challenge [27]; SE-Deeplab employs the structure of Deeplab v3 and incorporates a squeeze-and-excitation (SE) module. All these models are trained with the same learning rates and employ the same data processing to ensure fairness.

Table 3 reports the quantitative experiment results of the compared methods on the DeepGlobe dataset. The accuracy of FCN is lower than those of other methods, which is mainly due to the loss of spatial details. The D-LinkNet with additional dilated convolution layers outperforms ResUNet, obtaining an improvement of 1.5% on the IoU score and 0.67% on the F1-score. The design of an SE module in SE-DeepLab improves the IoU score by 7.62% and the F1-score by 4.01% compared with the D-LinkNet.

Compared with those methods, the proposed MSPNet achieves the best performance in OA, IoU, and F1-score. For example, MSPNet obtains an F1-score of 84.51% and an IoU score of 73.64%, which are better than SE-DeepLab by 1.78% and 2.27%, respectively. We attribute these significant performance gains to three factors. (1) Due to the symmetric network design with skip-connections, the proposed MSPNet is able to preserve low-level spatial details and thus extract roads with smoother boundaries from remote sensing images. (2) The strip pooling design in our proposed model enhances the embedding of linear features, which greatly improves the connectivity of roads; therefore, the proposed MSP module can further improve the segmentation accuracy. (3) The joint supervision of BCE and dice loss alleviates the class imbalance problem.

**Table 3.** Results of the comparative experiments.

| Methods | OA (%) | IoU (%) | F1 (%) |
|---|---|---|---|
| FCN | 96.52 | 60.51 | 74.85 |
| ResUNet | 97.45 | 62.74 | 77.89 |
| D-LinkNet | 97.61 | 64.24 | 78.56 |
| CADUNet [37] | / | 66.38 | 78.75 |
| DSE-LinkNet [3] | / | 69.57 | 76.73 |
| HsgNet [24] | / | / | 82.90 |
| SE-Deeplab | 98.12 | 71.86 | 82.57 |
| MSPNet (ours) | 98.71 | 73.64 | 84.51 |

The partially visualized segmentation results of our MSPNet and other methods are illustrated in Figure 8, which shows three examples from different scenes. The corresponding distributions of FP and FN are plotted in Figure 9. The results extracted by FCN are worse than those of other methods, which is mainly caused by the loss of spatial information after multiple downsampling operations in its early layers. The results of UNet and D-LinkNet are similar, and they also both contain many missing connections and FP pixels. The SE-Deeplab shows better road connectivity compared with other methods, and it obtains the second-best results. In comparison, our proposed method extracts roads with better connectivity and smoother road edges. The roads segmented by other methods may be interrupted especially in some regions where occlusions exist, while our MSPNet recovers the connectivity very well by effectively capturing long-range dependencies along road regions. For example, in the results of rural areas (the first row in Figure 8), there are several residential houses in the upper-right and lower-left corner of the image, which are planted many trees on both sides; the results extracted by other methods contain some broken road segments and fail to maintain the connectivity, but the result extracted by MSPNet is consistent with the ground truth. In addition, in the results of densely connected road areas (the second and third row in Figure 8), other methods recognize some branch roads as background, while the proposed methods are consistent with those of ground truth, and very few FP pixels exist. These visualized segmentation results verify the superiority of our MSPNet in the task of road extraction from remote sensing images.

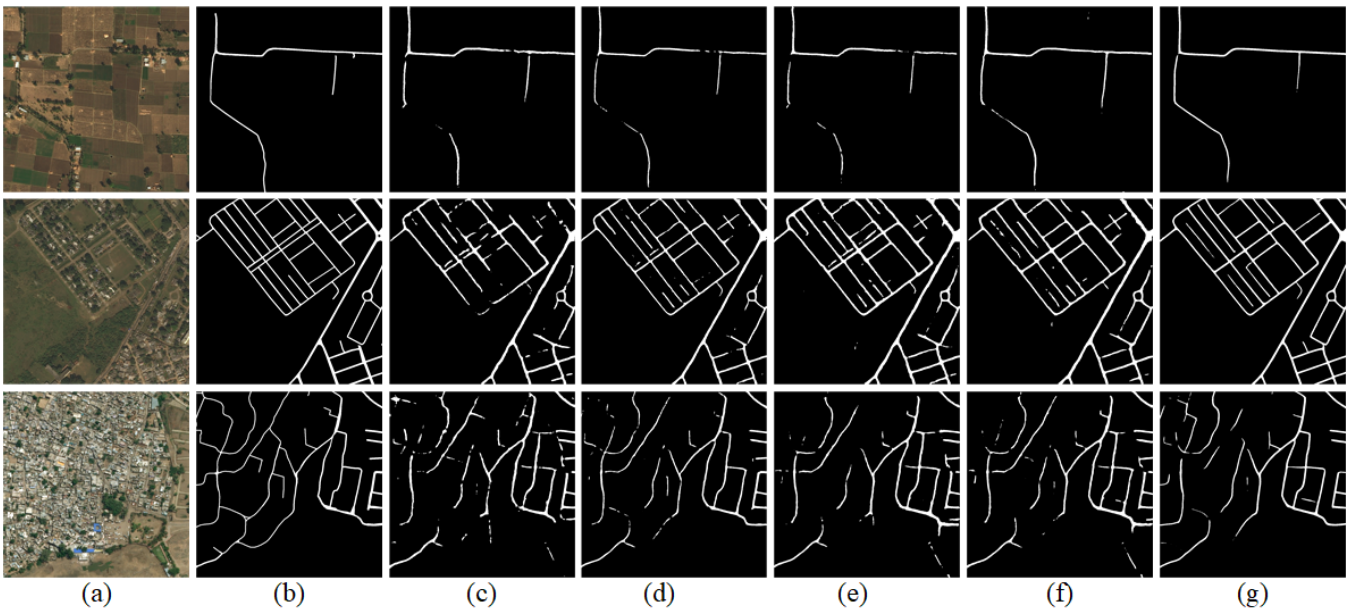


**Figure 8.** Visualization of segmentation results of proposed MSPNet and other methods. (**a**) Test image. (**b**) Ground Truth. (**c**–**g**) Road extraction results of FCN, ResUNet, D-LinkNet, SE-Deeplab and our MSPNet.

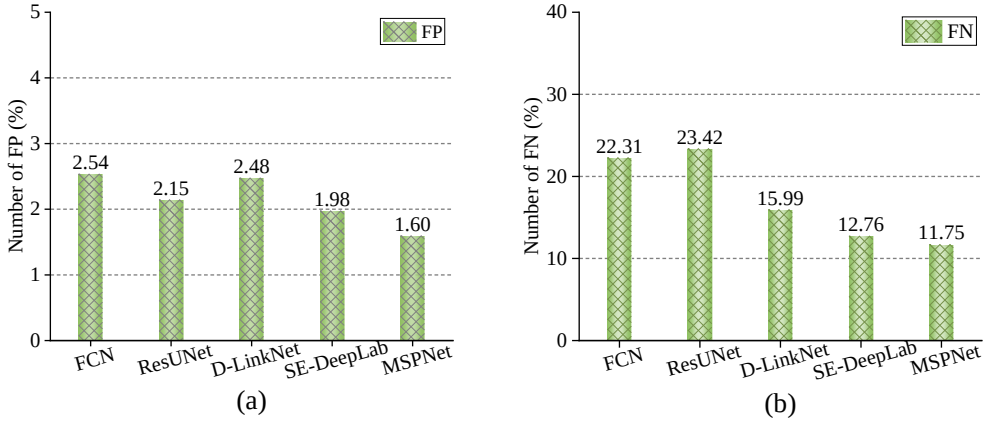


**Figure 9.** (**a**,**b**) respectively plot the comparison methods on misclassification percentages of false positive (FP) and false negative (FN) in the results of three samples in Figure 8.

## 5. Discussion

The above experimental results demonstrates that our proposed MSPNet obtains new competitive performance over other state-of-the-art methods. However, the road maps extracted by all the considered methods still have some interruptions. This situation exists in some special surface environments. We show two samples as illustrated in Figure 10. The road regions are severely occluded by a large number of trees that are difficult to distinguish in the upper-left corner of Figure 10 (the first row). Figure 10 (the second row) shows some areas of farmland, and the color of the roads is very similar to the surrounding environment. Our MSPNet fails to predicate complete road maps in these challenging regions. In the future studies, some other prior information will be introduced to segment the roads in these challenging regions, such as the direction information of the roads, which may be helpfull to generate more complete results in those occluded regions.


**Figure 10.** Two examples that our MSPNet fails. (**a**) Test image. (**b**) Ground Truth. (**c**) Segmentation results of our MSPNet.

To fairly evaluate different methods, as shown in Table 4, we list the floating point operations per second (FLOPS) and inference time of our MSPNet and several compared methods. The comparison experiments of all methods are run on a workstation with a NVIDIA RTX 2080Ti GPU. For fair comparison, the FLOPs and inference time are calculated based on an input size of 1024 × 1024. It is seen that the proposed MSPNet obtains

competitive inference time, while it has a reasonable computational cost compared with other methods. The comparison results illustrate that our proposed MSPNet is suitable for road extraction tasks from remote sensing images.

During the training process, we apply data augmentation for improving the generalization of the proposed model. We also perform comparative experiments to show the contribution of data augmentation, as shown in Table 5. We find that it gives an improvement of 0.38 in terms of IoU and 0.44 in F1-score. The results demonstrate that data augmentation is a useful strategy to improve performance.

**Table 4.** FLOPS and inference time of our proposed MSPNet and other methods. The inference times are calculated using 10 test images and then averaged.

| Methods | FLOPS (Gbps) | Interfence (s) |
|---|---|---|
| FCN | 97.47 | 0.097 |
| ResUNet | 191.36 | 0.145 |
| D-LinkNet | 84.51 | 0.123 |
| SE-Deeplab | 223.65 | 0.176 |
| MSPNet(ours) | 100.49 | 0.106 |

**Table 5.** Comparative study of data enhancement. ✗ and ✓ denote our use and non-use of data augmentation strategies, respectively.

| Data Augmentation | IoU (%) | F1 (%) |
|---|---|---|
| ✗ | 73.26 | 84.07 |
| ✓ | 73.64 | 84.51 |

## 6. Conclusions

In this paper, we propose an end-to-end road segmentation network for road extraction tasks from remote sensing images. Although the CNN-based methods have greatly improved the accuracy of road extraction over traditional approaches, the road connectivity should be further improved to generate more complete results. As one of the important geometric topological properties, road connectivity is necessary for autonomous driving, vehicle navigation, and route planning. However, the existing CNN-based methods often fail to predict road connectivity and thus produce fragmented road segments. As a comparison, our proposed MSPNet is able to generate the road segmentation results with better connectivity and therefore meet the requirements of large-scale remote sensing data analysis. Specifically, the proposed MSPNet strengthens the linear feature of roads by introducing strip pooling layers, where its pooling kernel shapes are more in line with the roads. Accordingly, a multi-scale strip (MSP) module is developed to learn multiple long-range contextual information. In this paper, the widely used design of a symmetric encoder–decoder network with skip-connections is adopted to the low-level features to recover the spatial details, which is beneficial for parsing high-resolution remote sensing images. What is more, to alleviate the problem caused by unbalanced road and background pixels, we have performed ablation experiments to adjust the loss contributions between cross entropy and dice-coefficient loss functions to suit the task of road extraction. We have also compared the performance and computational cost of ResNet series models as the backbone network to select an appropriate backbone. Experimental results on a popular benchmark DeepGlobe dataset show the superiority of our proposed MSPNet compared with several mainstream methods.

As discussed in the above section, some road types that have similar spectral and texture to background are not well identified, and there are still some discontinuities. Since road extraction can be viewed as a binary classification problem, it may be useful to suppress background noise to improve the generalization ability of the road segmentation model. This is left for our future studies.

**Author Contributions:** All authors made contributions to the manuscript. Conceptualization, methodology, software, validation and writing, S.L. and S.Q.; software and visualization, S.Q.; validation and investigation, B.Z.; writing and preparation, S.Q. and H.Z.; funding acquisition, S.Q. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was partially supported by the Science and Technology Development Plan Project of Henan Province, China (No. 212102210538,222102210101), Henan Key Science and Technology Project (No. 201300210400), Graduate Education Innovation and Quality Improvement Plan Project of Henan University (No. SYL20040121), National College Student Innovation and Entrepreneurship Training Program (No. 202110475135).

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

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data that support the findings of this study are available on request from the corresponding author.

**Acknowledgments:** The authors would like to thank all the anonymous reviewers for their helpful comments and suggestions to improve the manuscript.

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

## References

1. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
2. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 640–651. [CrossRef]
3. Das, P.; Chand, S. Extracting road maps from high-resolution satellite imagery using refined DSE-LinkNet. *Connect. Sci.* **2021**, *33*, 278–295. [CrossRef]
4. Ding, L.; Bruzzone, L. DiResNet: Direction-Aware Residual Network for Road Extraction in VHR Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 10243–10254. [CrossRef]
5. Raziq, A.; Xu, A.; Yu, L. Automatic Extraction of Urban Road Centerlines from High-Resolution Satellite Imagery Using Automatic Thresholding and Morphological Operation Method. *J. Geogr. Inf. Syst.* **2016**, *8*, 517–525. [CrossRef]
6. Cheng, G.L.; Zhu, F.Y.; Xiang, S.M.; Wang, Y.; Pan, C.H. Accurate urban road centerline extraction from VHR imagery via multiscale segmentation and tensor voting. *Neurocomputing* **2016**, *205*, 407–420. [CrossRef]
7. Dai, J.; Wang, Y.; Du, Y.; Zhu, T.; Xie, S.; Li, C.; Fang, X. Development and prospect of road extraction method for optical remote sensing image. *J. Remote Sens.* **2020**, *24*, 804–823.
8. Xin, J.; Zhang, X.C.; Zhang, Z.Q.; Fang, W. Road Extraction of High-Resolution Remote Sensing Images Derived from DenseUNet. *Remote Sens.* **2019**, *11*, 2499. [CrossRef]
9. Han, Y.B.; Han, P.; Jia, M.L. Road extraction from high resolution remote sensing image via a deep residual and pyramid pooling network. *IET Image Process.* **2021**, *15*, 3080–3093. [CrossRef]
10. Ren, Y.F.; Yu, Y.T.; Guan, H.Y. DA-CapsUNet: A Dual-Attention Capsule U-Net for Road Extraction from Remote Sensing Imagery. *Remote Sens.* **2020**, *12*, 2866. [CrossRef]
11. Fan, K.L.; Li, Y.X.; Yuan, L.; Si, Y.; Tong, L. New Network Based on D-Linknet and Resnext for High Resolution Satellite Imagery Road Extraction. In Proceedings of the IGARSS 2020—2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October 2020; pp. 2599–2602. [CrossRef]
12. Fu, G.; Zhao, H.R.; Li, C.; Shi, L.M. Road Detection from Optical Remote Sensing Imagery Using Circular Projection Matching and Tracking Strategy. *J. Indian Soc. Remote Sens.* **2013**, *41*, 819–831. [CrossRef]
13. Ma, R.G.; Wang, W.X.; Liu, S. Extracting roads based on Retinex and improved Canny operator with shape criteria in vague and unevenly illuminated aerial images. *J. Appl. Remote Sens.* **2012**, *6*, 063610. [CrossRef]
14. Herumurti, D.; Uchimura, K.; Koutaki, G.; Uemura, T. Urban Road Network Extraction Based on Zebra Crossing Detection From a Very High Resolution RGB Aerial Image and DSM Data. In Proceedings of the 2013 International Conference on Signal-Image Technology and Internet-Based Systems (Sitis), Kyoto, Japan, 2–5 December 2013; pp. 79–84. [CrossRef]
15. Song, M.J.; Civco, D. Road extraction using SVM and image segmentation. *Photogramm. Eng. Remote Sens.* **2004**, *70*, 1365–1371. [CrossRef]
16. Mei, J.; Li, R.J.; Gao, W.; Cheng, M.M. CoANet: Connectivity Attention Network for Road Extraction from Satellite Imagery. *IEEE Trans. Image Process.* **2021**, *30*, 8540–8552. [CrossRef] [PubMed]
17. Ding, C.; Weng, L.G.; Xia, M.; Lin, H.F. Non-Local Feature Search Network for Building and Road Segmentation of Remote Sensing Image. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 245. [CrossRef]
18. Mnih, V.; Hinton, G.E. Learning to Detect Roads in High-Resolution Aerial Images. In Proceedings of the Computer Vision—ECCV 2010—11th European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010; Part VI. [CrossRef]

19. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015; pp. 234–241.
20. Chaurasia, A.; Culurciello, E. LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation. In Proceedings of the IEEE Visual Communications and Image Processing (VCIP), St. Petersburg, FL, USA, 10–13 December 2017; IEEE: New York, NY, USA, 2017.
21. Zhou, K.; Xie, Y.; Gao, Z.; Miao, F.; Zhang, L. FuNet: A Novel Road Extraction Network with Fusion of Location Data and Remote Sensing Imagery. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 39. [CrossRef]
22. 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]
23. 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. 192–196. [CrossRef]
24. Xie, Y.; Miao, F.; Zhou, K.; Peng, J. HsgNet: A Road Extraction Network Based on Global Perception of High-Order Spatial Information. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 571. [CrossRef]
25. Li, X.G.; Zhang, Z.; Lv, S.S.; Pan, M.; Ma, Q.; Yu, H.B. Road Extraction from High Spatial Resolution Remote Sensing Image Based on Multi-Task Key Point Constraints. *IEEE Access* **2021**, *9*, 95896–95910. [CrossRef]
26. Wei, Y.; Zhang, K.; Ji, S.P. Simultaneous Road Surface and Centerline Extraction From Large-Scale Remote Sensing Images Using CNN-Based Segmentation and Tracing. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8919–8931. [CrossRef]
27. Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; Raskar, R. Deepglobe 2018: A challenge to parse the earth through satellite images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 172–181.
28. Henry, C.; Azimi, S.M.; Merkle, N. Road Segmentation in SAR Satellite Images With Deep Fully Convolutional Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1867–1871. [CrossRef]
29. 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]
30. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
31. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
32. Zhao, H.S.; Shi, J.P.; Qi, X.J.; Wang, X.G.; Jia, J.Y. Pyramid Scene Parsing Network. In Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: New York, NY, USA, 2017; pp. 6230–6239. [CrossRef]
33. Tian, T.; Chu, Z.; Hu, Q.; Ma, L. Class-Wise Fully Convolutional Network for Semantic Segmentation of Remote Sensing Images. *Remote Sens.* **2021**, *13*, 3211. [CrossRef]
34. Hou, Q.; Zhang, L.; Cheng, M.M.; Feng, J. Strip Pooling: Rethinking Spatial Pooling for Scene Parsing. In Proceedings of the CVPR, Seattle, WA, USA, 14–19 June 2020.
35. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
36. Lin, Y.E.; Xu, D.Y.; Wang, N.; Shi, Z.; Chen, Q.X. Road Extraction from Very-High-Resolution Remote Sensing Images via a Nested SE-Deeplab Model. *Remote Sens.* **2020**, *12*, 2985; Erratum in *Remote Sens.* **2021**, *13*, 783. [CrossRef]
37. Li, J.; Liu, Y.; Zhang, Y.N.; Zhang, Y. Cascaded Attention DenseUNet (CADUNet) for Road Extraction from Very-High-Resolution Images. *Isprs Int. J. Geo-Inf.* **2021**, *10*, 329. [CrossRef]