electronics

Article

Remote Sensing Image Road Extraction Network Based on MSPFE-Net

Zhiheng Wei [1,2] and Zhenyu Zhang [1,2,*]

[1] School of Information Science and Engineering, Xinjiang University, Urumqi 830017, China; wei577245796@163.com
[2] Xinjiang Key Laboratory of Multilingual Information Technology, Xinjiang University, Urumqi 830017, China
* Correspondence: zhangzhenyu@xju.edu.cn

Abstract: Road extraction is a hot task in the field of remote sensing, and it has been widely concerned and applied by researchers, especially using deep learning methods. However, many models using convolutional neural networks ignore the attributes of roads, and the shape of the road is banded and discrete. In addition, the continuity and accuracy of road extraction are also affected by narrow roads and roads blocked by trees. This paper designs a network (MSPFE-Net) based on multi-level strip pooling and feature enhancement. The overall architecture of MSPFE-Net is encoder-decoder, and this network has two main modules. One is a multi-level strip pooling module, which aggregates long-range dependencies of different levels to ensure the connectivity of the road. The other module is the feature enhancement module, which is used to enhance the clarity and local details of the road. We perform a series of experiments on the dataset, Massachusetts Roads Dataset, a public dataset. The experimental data showed that the model in this paper was better than the comparison models.

Keywords: road extraction; convolutional neural network; remote sensing images; strip pooling

check for updates

Citation: Wei, Z.; Zhang, Z. Remote Sensing Image Road Extraction Network Based on MSPFE-Net. *Electronics* **2023**, *12*, 1713. https://doi.org/10.3390/electronics12071713

Academic Editors: Yue Wu, Kai Qin, Maoguo Gong and Qiguang Miao

Received: 11 March 2023
Revised: 27 March 2023
Accepted: 31 March 2023
Published: 4 April 2023

1. Introduction

The continuous progress of remote sensing and artificial intelligence has laid a solid theoretical and technical foundation for road extraction. As an important ground object, roads play a crucial role in intelligent transportation [1,2], urban planning [3,4], and emergency tasks [5,6]. Traditional remote sensing road extraction uses manual interpretation, which consumes a lot of workforce and time. The automatic extraction of roads is helpful to reduce the manual workload, and it also accelerates the speed of road extraction, so road extraction has excellent research value. Due to the different shapes of roads, the influence of background factors, and being blocked by trees or shadows, the road extraction task is difficult and challenging [7–9].

Deep learning promotes the progress of computer vision, especially in object detection, semantic segmentation, image classification, and other aspects, and it has a good effect. Scholars also began to use deep learning technology to complete remote sensing image road extraction [10–13]. Although the model based on deep learning has achieved good results in extracting road tasks, and many road extraction algorithms have problems, such as road breaking caused by occlusion, difficult extraction of the narrow road, and incorrect identification of roads and background. In order to solve these problems, road extraction models need strong long-distance dependencies or global context information, and the road extraction algorithm usually uses attention mechanism or atrous convolution technology to obtain long-distance dependencies or global context information. The attention mechanism is a method to improve the ability of global context modeling. However, it consumes a lot of memory. Other methods include atrous convolution and spatial pyramid pooling, which can expand the receptive field of convolutional neural network, but the strip target features extracted by the square window may be mixed with irrelevant target information.

Inspired by the idea of strip pooling proposed by Hou, Feng et al. [14], this paper introduces and improves upon it. The strip pooling has several distinct characteristics. Firstly, the strip pooling has a long and banded shape at a dimension, so it can capture long-range relationships of isolated regions. Then, strip pooling maintains a narrow shape along a spatial dimension, which helps to capture the local feature of targets and can reduce the interference of irrelevant target information. The network combined with strip pooling has the ability to obtain multiple types of contexts. It is fundamentally different from traditional spatial pooling. The idea of strip pooling is well applied in remote sensing image road extraction scenes. Therefore, a multi-level strip pooling and feature enhancement network (MSPFE-Net) called MSPFE-Net is designed in this paper. In MSPFE-Net, the multi-level strip pooling module is responsible for fully extracting long-range context information. The feature enhancement module is used to enhance the clarity and local details of the road.

The main content of this article consists of the following:

1. A multi-level strip pooling module (MSPM) was designed to extract global context information to ensure the connectivity and integrity of road extraction.
2. A feature enhancement module (FEM) was proposed, which mainly enhanced the clarity and local details of the road
3. MSPFE-Net is designed and implemented for road extraction tasks. The effectiveness of MSPFE-Net was verified on the Massachusetts Roads Dataset.

The chapter structure of this article is arranged as follows: Section 2 introduces the related work. Section 3 shows the structure of MSPFE-Net and explains the rationale for each module. Section 4 shows experimental contents, containing experimental datasets, evaluation methods, experimental settings, and experimental results. Sections 5 and 6 introduce this paper's discussion and conclusion, respectively.

2. Related Works

In recent years, a series of algorithms for road extraction have been proposed. According to the characteristics of various algorithms, there are mainly two types: traditional type and type of deep learning.

2.1. Traditional Type

Traditional types have the following methods: template matching method, knowledge-driven method, and object-oriented method [15]. Template matching is a method that applies geometric, topological, and radial features of road images. According to the template type, it can be divided into rule templates and variable templates. The advantages of the rule template are less computation, good stability, and simplicity, while the disadvantages are affected by the transformation of radiation characteristics. The advantage of a variable template is that it can be applied to images with irregular road edges and irregular road radiation information. The disadvantage is that it requires a large amount of computation. The primary process is to design the template according to the rules, obtain the regional extreme value through the template using the measure function and update the road location. Haverkamp [16] used the comparative analysis of multiple rectangular templates to rotate at certain angular intervals to form a group of discrete rectangular templates. In the knowledge-driven method, knowledge can be divided into geometric knowledge, contextual knowledge, and auxiliary knowledge. Wenfeng Wang et al. [17] put forward a straight-line detection algorithm using the property of parallel edges of roads, recognized parallel features using principal component analysis, and direction consistency criteria. The object-oriented method is to obtain the output results by segmentation, classification, and post-processing of the input image. The segmentation methods include threshold segmentation, graph segmentation, and edge segmentation. Classification methods include geometric feature classification and SVM; post-processing includes tensor voting and mathematical morphology. Maboudi et al. [18] used guided filtering to eliminate the inconsistency of pavement image texture and then used a method including color and shape data for road extraction.

2.2. Type of Deep Learning

Using a convolutional neural network (CNN) to obtain road features from many image sample data. Mnih and Hinton [19] combined deep learning with road extraction for the first time, using a restricted Boltzmann machine to detect road areas from images. P. Li et al. [20] designed a network combining CNN and linear integral convolution to extract roads. Wei et al. [21] applied an improved cross entropy loss function to the CNN, which can help improve the topological information of the road. The fully convolutional network (FCN) was proposed in 2014 [22]. The deconvolution of FCN can make the final feature map have the same size as the input image after up-sampling and predict the category of each pixel. Although FCN plays a pioneering role in semantic segmentation, its accuracy is low. Since the U-Net model was applied in the task of medical image segmentation, it achieved good results. The U-Net network is an improved fully convolutional network [23]. It includes the skip connection. In the process of up-sampling, the feature map in the process of down-sampling is fused in concatenate. Many subsequent image segmentation models have adopted this idea and improved on it. In the field of road extraction, there is no exception. Singh et al. [24] proposed their improved U-Net model to realize the function of road extraction. He et al. [25] designed the deep residual network to make the number of network layers deeper. Zhang et al. [26] realized road extraction by introducing ResNet into U-Net and combining the advantages of both. In order to accomplish road extraction at different scales, Gao et al. [27] introduced a feature pyramid and proposed a multi-feature pyramid network (MFPN). Cheng et al. [28] proposed a cascading end-to-end network (CasNet), which completes the road detection task and the road centerline extraction task simultaneously through two cascading networks. However, these methods have an insufficient receptive field to capture effective and rich long-distance context information, which is crucial for road extraction. The lack of long-distance context information will directly lead to the discontinuity of road extraction results or even the phenomenon that roads cannot be extracted completely. In order to connect discontinuous broken roads, many researchers have considered various schemes to capture long-distance context information to model the topological relationship between broken roads [29]. The main method is to use atrous convolution [30]. Atrous convolution can effectively expand the receptive field without increasing the amount of computation. Taking into account the natural connectivity and large span of the road, Zhou et al. [31] added the concatenation mode and parallel mode of different atrous convolution to form D-LinkNet for road extraction. In order to extract road features at different scales, He et al. [32] introduced the atrous spatial pyramid pooling (ASPP) module. According to the above analysis, compared with the traditional method, the deep learning method can greatly improve the accuracy and automation of road extraction, but there are still problems of road breaking caused by occlusion in the road extraction results [33]. Although many researchers have offered some solutions, there is still a lack of high-performance, end-to-end road extraction networks that can solve this problem. Tao et al. [34] proposed to integrate a well designed spatial information inference structure into the deeplabv3+ network to maintain the continuity of road extraction by realizing multi-direction transmission of information between pixels. Zhou et al. [35] proposed a boundary and topologically-aware road extraction network (BT-RoadNet) in order to improve the quality of road boundaries and solve the problem of road discontinuity. The network is divided into thick and thin prediction modules to obtain detailed boundary information, and the spatial context module is designed to solve the problem of discontinuous road results. Lu et al. [36] proposed GAMSNet, which uses multi-scale residual learning to extract multi-scale features, and then it uses global perception operations to capture long-distance relationships. Tan et al. [37] proposed a new end-to-end encoder-decoder architecture network to solve the problem of road location information loss due to reduced spatial resolution. This network obtains different levels of features by encoders, and the decoder is composed of a scale fusion module and a scale sensitive module, respectively, achieving the task of fusing features and assigning weights. Zhu et al. [38] designed a global context-aware batch processing independent network

(GCB-Net), which effectively integrates global context features by using the improved non-local module as a global context-aware module. Wang et al. [39] designed a model combining global attention, and it can enhance the performance of road segmentation.

In conclusion, Although the method based on deep learning effectively extracts roads, it is still difficult to extract roads especially in complex scenes. Therefore, it is necessary to fully consider the structural characteristics of roads and improve the accuracy of road extraction in complex scenes.

3. Methods

The details of the proposed network model are introduced in this section. Section 3.1 shows the architecture of the MSPFE-Net; Section 3.2 shows the multi-level strip pooling module. Section 3.3 introduces the feature enhancement module in detail. Section 3.4 introduces the loss function.

3.1. MSPFE-Net Model

The MSPFE-Net is shown in Figure 1, it is mainly composed of the encoder, multi-level strip pooling module (MSPM), feature enhancement module (FEM), and decoder. The encoder uses the Resnet50 network. The output results of the encoder in the first four layers are used as inputs to MSPM, which is used to strengthen the long-range dependencies of the model. The output results of the encoder in the fifth layer are used as inputs to the feature enhancement module, and it focuses on collecting various types of contexts through different pooling operations to make the road feature representation more discriminating. The output feature of MSPM will be added with the up-sampled feature map in the decoder.

3.2. Multi-Level Strip Pooling Module

Figure 2 explains the theory of the strip pooling. The strip pooling window performs pooling at horizontal or vertical dimensions, and the input feature is a two-dimensional tensor $x \in R^{H \times W}$. Different from two-dimensional average pooling, the method of strip pooling is to sum the value of a row or column and divide it by the number of rows or columns, respectively. Therefore, the horizontal strip pooling output $y^h \in R^H$ can be expressed as:

$$y_i^h = \frac{1}{W}\sum_{j=0}^{W-1} x_{i,j} \tag{1}$$

Similarly, the vertical strip pooling outputs $y^v \in R^W$ can be expressed as:

$$y_i^v = \frac{1}{H}\sum_{i=0}^{H-1} x_{i,j} \tag{2}$$

The strip pooling in Figure 2 is similar to the traditional pooling method, and the pixel values are averaged over the locations on the feature maps corresponding to the pooling kernels. A feature map is an input, here actually C × H × W. For ease of description, only one channel is drawn. The feature map input processing principle of C channels is the same as that of one channel operation shown here. After horizontal and vertical strip pooling, the input feature map becomes H × 1 and 1 × W. The element values within the pooling window are averaged, and the value is used as the pooling output value. Subsequently, one-dimensional convolution is conducted on the output values, and the two output feature maps are expanded along the horizontal and vertical directions. Then, the two feature maps had the same size, and the expanded feature maps corresponding to the same position were summed to obtain the H × W feature map.

One of the difficulties of road extraction is maintaining the road's continuity. In order to reduce the impact of this problem, MSPM is proposed in this paper to fully obtain long-range dependencies to keep the connectivity and integrity of road topology. The core idea of MSPM is to extract different features by strip pooling at different levels and fuse these features. MSPM is added to the skip connection section of MSPFE-Net to extract long-range context information at different levels.

The framework of MSPM is shown in Figure 3, which contains three strip pooling sub-blocks of different sizes, L1, L2, and L3, respectively. Input feature x is processed by the L1, L2, and L3 sub-blocks, respectively, then the model obtains three output features y^{L1}, y^{L2}, and y^{L3} and add the corresponding position of y^{L1}, y^{L2}, and y^{L3}. The last operation is to multiply the summed result with the input x. Finally, there is the output result of MSPM, y^{out} can be expressed as:

$$y^{out} = x \otimes \left(y^{L1} + y^{L2} + y^{L3}\right) \tag{3}$$

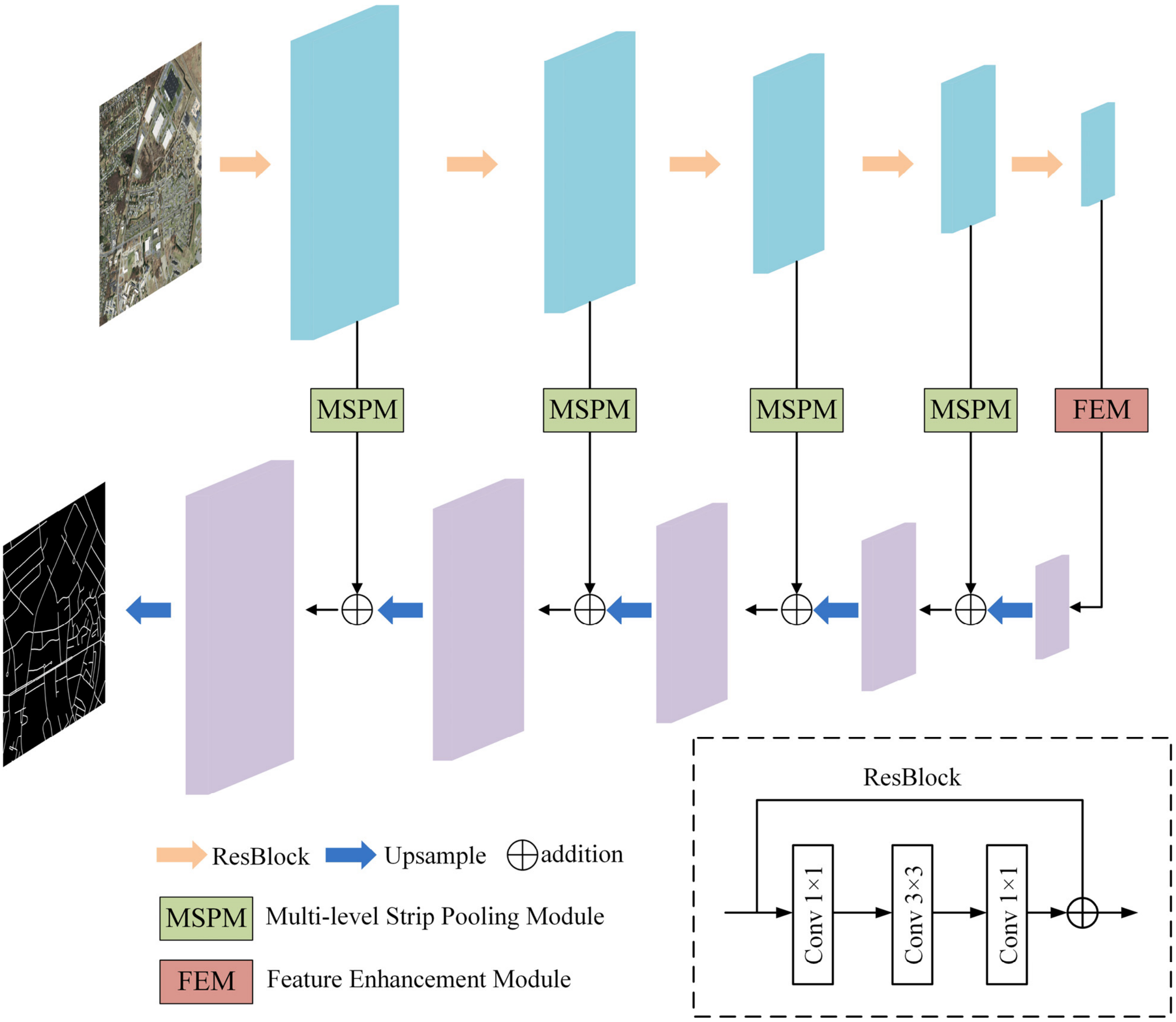


Figure 1. The architecture of MSPFE-Net.

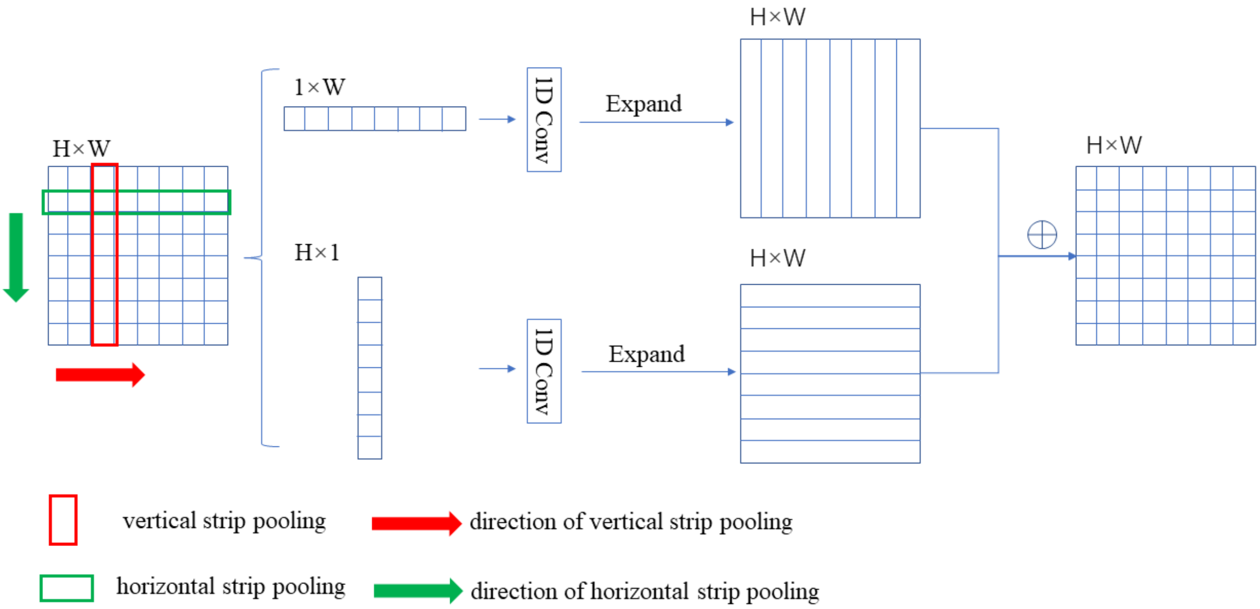


Figure 2. Illustration of strip pooling.

3.3. Feature Enhancement Module

The FEM, shown in Figure 4, focuses on collecting various types of contexts through different pooling operations to make feature representation more discriminating. The advantage is that it can be used continuously to extend long-range dependencies and reinforce local details.

The feature enhancement module is composed of four sub-modules, namely, 3 × 3 convolution, 3 × 3 maximum pooling, horizontal strip pooling, and vertical strip pooling. These four sub-modules are represented as f_1, f_2, f_3, and f_4. The input of FEM is represented as x, and the output of FEM is represented as y. Firstly, 3 × 3 convolution and 3 × 3 maximum pooling were carried out to add the above result features to obtain the feature map, namely, y_1. Similarly, horizontal strip pooling and vertical strip pooling were carried out, and the results were added to obtain the feature map, namely, y_2, which was splicing y_1 and y_2. After 1 × 1 conv, the final result feature y was obtained. They capture both long-range and local dependencies information, and it is essential for remote sensing image road extraction scene resolution networks. FEM can be expressed as:

$$y_1 = f_1(x) + f_2(x) \tag{4}$$

$$y_2 = f_3(x) + f_4(x) \tag{5}$$

$$y = C_{1\times 1}(CONCAT(y_1, y_2)) \tag{6}$$

where $C_{1\times 1}$ is 1 × 1 convolution, and CONCAT is a concatenate operation.

For long-range dependencies, unlike previous work using a global averaging pooling layer, we capture context information by using both horizontal and vertical strip pooling operations. At the same time, strip pooling makes it possible to connect discrete areas and code areas with strip structures throughout the road scene. However, in the case of tight distribution of semantic regions, capturing road local context information also requires spatial pooling. Considering this, convolution operation and pooling layer are used to obtain short-range dependencies.

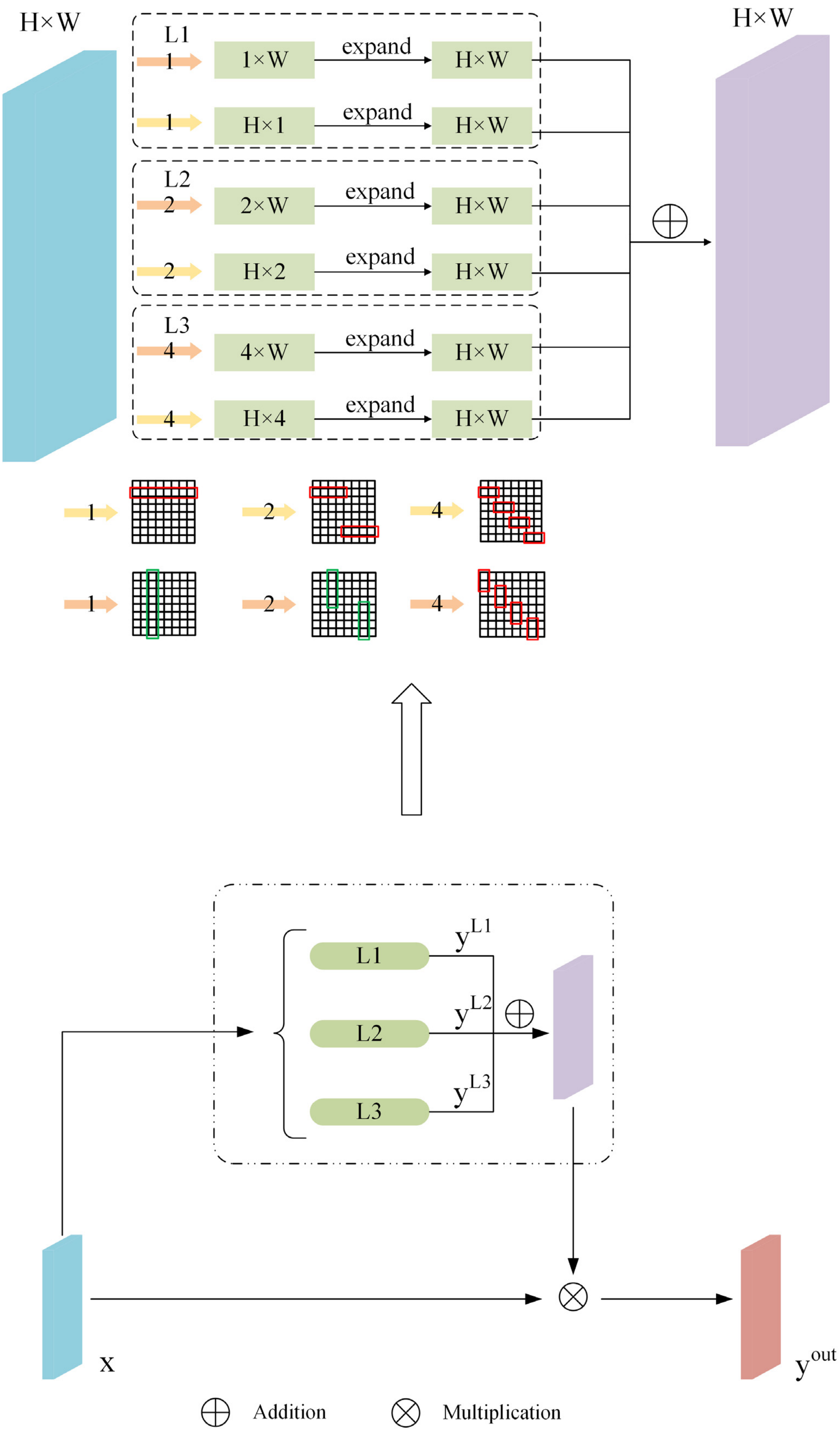


Figure 3. The structure of MSPM.

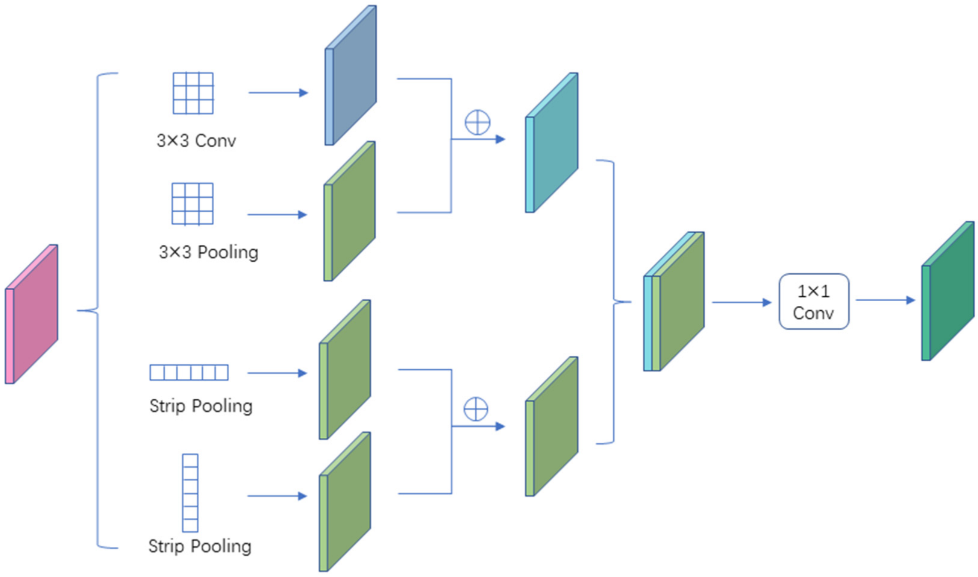


Figure 4. Illustration of feature enhancement module.

3.4. Loss Function

The binary cross entropy loss function is applied to most pixel-level segmentation tasks. However, when the number of pixels on the target is much smaller than the number of pixels in the background, that is, the samples are highly unbalanced, and the loss function has the disadvantage of misleading the model to seriously bias the background. In this paper, it is necessary to judge whether the pixels predicted by the model are roads or backgrounds. The road area is small, and the background area is too large. If the binary cross entropy loss function is used, this will make the model deviate from the optimal direction during the training process. To reduce the impact of this problem, the dice coefficient loss function and the focal loss function are used together as the loss function.

The dice coefficient loss function is calculated as follows:

$$L_d = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \tag{7}$$

In the formula, X is the generated prediction map, Y is the label, $|X \cap Y|$ is the intersection of label and prediction, $|X|$ is the number of elements of the label, and $|Y|$ represents the number of predicted elements.

The focal loss function is based on the binary cross entropy loss. It is a dynamically scaled cross entropy loss. Through a dynamic scaling factor, the weight of easily distinguishable samples can be dynamically reduced in the training process to focus on those indistinguishable samples quickly. The focal loss function is as follows:

$$L_f = FL(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{8}$$

Among them, $-\log(p_t)$ is the initial cross entropy loss function, α is the weight parameter between categories, $(1 - p_t)^{\gamma}$ is the easy/hard sample adjustment factor, and γ is the focusing parameter.

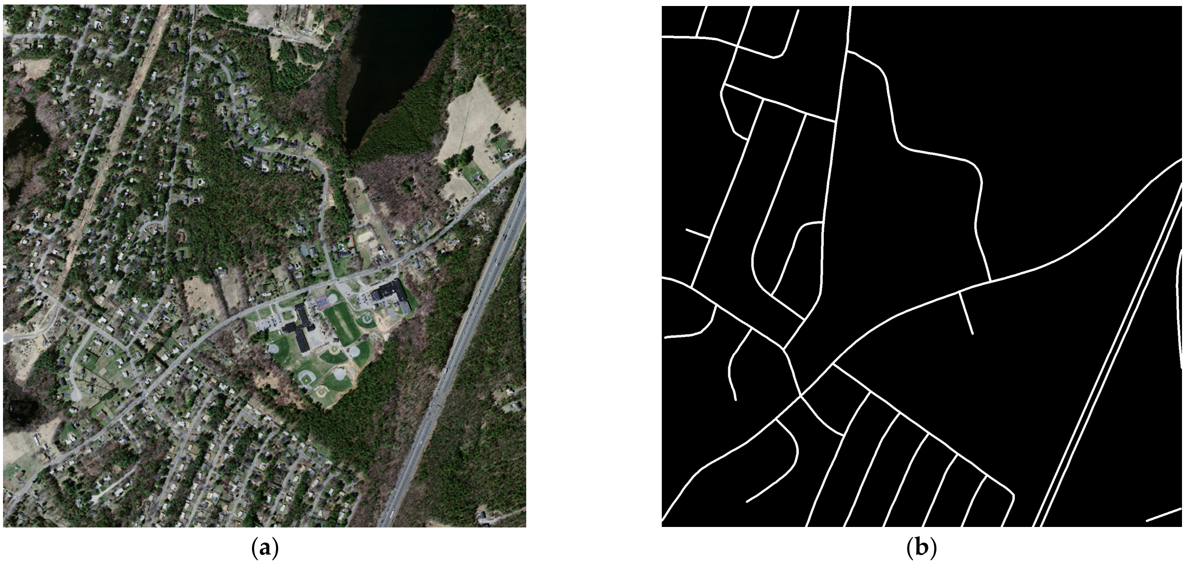
The final loss function is the sum of the dice coefficient loss function and the focal loss function, namely:

$$L_{loss} = L_f + L_d \tag{9}$$

4. Results

4.1. Dataset

Massachusetts Roads Dataset [40] is used in the experimental dataset as shown in Figure 5. The pixel size of the Massachusetts Roads Dataset is 1500 × 1500, and there are 1171 pairs of images and labels. In this experiment, the number of training images, test images, and validation images is 1108, 49, and 14, respectively.



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

Figure 5. Massachusetts Road Dataset. (**a**) Image; (**b**) Label.

4.2. Evaluation Metrics

Selecting appropriate evaluation metrics is of great reference significance for evaluating the model's performance. This paper adopts Recall, Precision, *F1-Score* (F1), and intersection over union (*IoU*), and these evaluation metrics commonly used in semantic segmentation. *F1-Score* is calculated by two indicators. Intersection over union (*IoU*) refers to the ratio between the intersection of predicted road pixels and real labeled road pixels and their union. The specific calculation method is:

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{12}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{13}$$

In the formula, *TP* is the number of pixels correctly classified as roads, *TN* is the number of pixels correctly classified as non-roads, *FP* represents the number of pixels wrongly classified as roads, and *FN* represents the number of pixels wrongly classified as non-roads. The values of these evaluation metrics are in the range of [0,1]. The best effect is that the evaluation metrics value is equal to 1

4.3. Experimental Settings

In the process of model training, the batch size of each training input network sample is 4. The initial learning rate is 0.001, and the decay of the learning rate is adjusted by the cosine annealing algorithm. The maximum training epoch is 100. The optimizer uses the

Adam algorithm with momentum set to 0.9. The loss function combines the dice coefficient loss function and focal loss function.

4.4. Experimental Results and Analysis

In order to test and prove the feasibility and rationality of MSPFE-Net, the mainstream semantic segmentation network is applied to the task of road segmentation, and the MSPFE-Net is compared. Table 1 shows the comparison of each model in the road segmentation task. Analysis showed that: (1) the F1-Score of the proposed model was 12.03%, 6.58%, 3.35%, 2.17%, and 1.40% higher than that of Deeplabv3+, U-Net, HRNetV2, D-LinkNet, and RefineNet, respectively. (2) The IoU of the proposed model was 13.89%, 7.91%, 4.13%, 2.70%, and 1.74% higher than that of Deeplabv3+, U-Net, HRNetV2, D-LinkNet, and RefineNet, respectively. (3) From F1-Score and IoU, the MSPFE-Net is better than Deeplabv3+, U-Net, HRNetV2, D-LinkNet, and RefineNet.

Table 1. Numerical results of different networks.

Networks	Precision (%)	Recall (%)	F1 (%)	IoU (%)
Deeplabv3+	54.13%	73.25%	62.26%	45.20%
U-Net	77.41%	60.17%	67.71%	51.18%
HRNetV2	65.48%	77.38%	70.94%	54.96%
D-LinkNet	71.72%	72.51%	72.12%	56.39%
RefineNet	69.57%	76.54%	72.89%	57.35%
MSPFE-Net(ours)	73.11%	75.50%	74.29%	59.09%

Figure 6 shows the effect of each model. According to the road extraction result map based on Deeplabv3+, the connectivity and geometric topological relationship of the road remain relatively complete, but the details of the road edge are rough, and more areas are misjudged as roads in the background map. This phenomenon is mainly due to the fact that the Deeplabv3+ model focuses on extracting deep semantic information, while the overall shape of the road is thin, and road details will be lost through the Deeplabv3+ backbone network. The road extraction effect based on the U-Net model is relatively clear in the details of road edges, and adjacent roads can be accurately displayed and separated. However, because of the limitation of the traditional convolution receptive field, the long-range features cannot be captured effectively. Therefore, the overall connectivity of roads extracted by U-Net is poor, and roads have more fracture zones, especially thin and narrow roads that have long fracture zones. The road extraction effect based on the D-LinkNet model is generally good, but D-LinkNet does not obtain enough long-range context information, so some roads appear discontinuous, and edge details are not clear. The overall structure of the road extracted by HRNetV2 is relatively complete, but the edge of the road is rough, and adjacent roads cannot be distinguished. The road extracted by RefineNet has good continuity, but there are some misjudgments. By means of incorporating the MSPM and FEM, the presented model in this paper has proficiently conserved intricate details whilst capturing the long-range feature relationships pertaining to the road network. Evidently, based on the visual analysis of Figure 6, the extracted road has exhibited substantial preservation of overall framework and connectivity whilst manifesting enhanced clarity of edge details. Consequently, this breakthrough endeavor promotes the depiction of intricate details and effective feature recognition, thus paving the way for an improved comprehension of road and their networks.

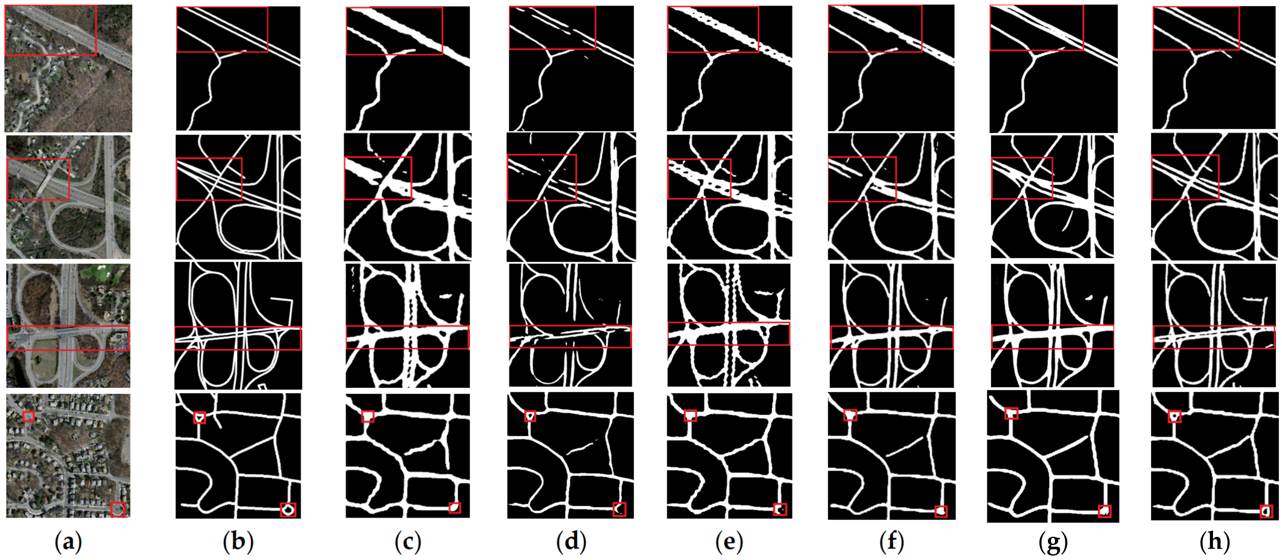


Figure 6. Results of MSPFE-Net and other methods. (**a**) Image; (**b**) Label; (**c**) DeepLabv3+; (**d**) U-Net; (**e**) HRNetV2; (**f**) D-LinkNet; (**g**) RefineNet (**h**) MSPFE-Net.(Red squares represent key areas.).

In Figure 7, it can be found that the road is blocked by trees, the MSFPE-Net accurately restored the road, while the road extracted by other comparison models is discontinuous or even not extracted.

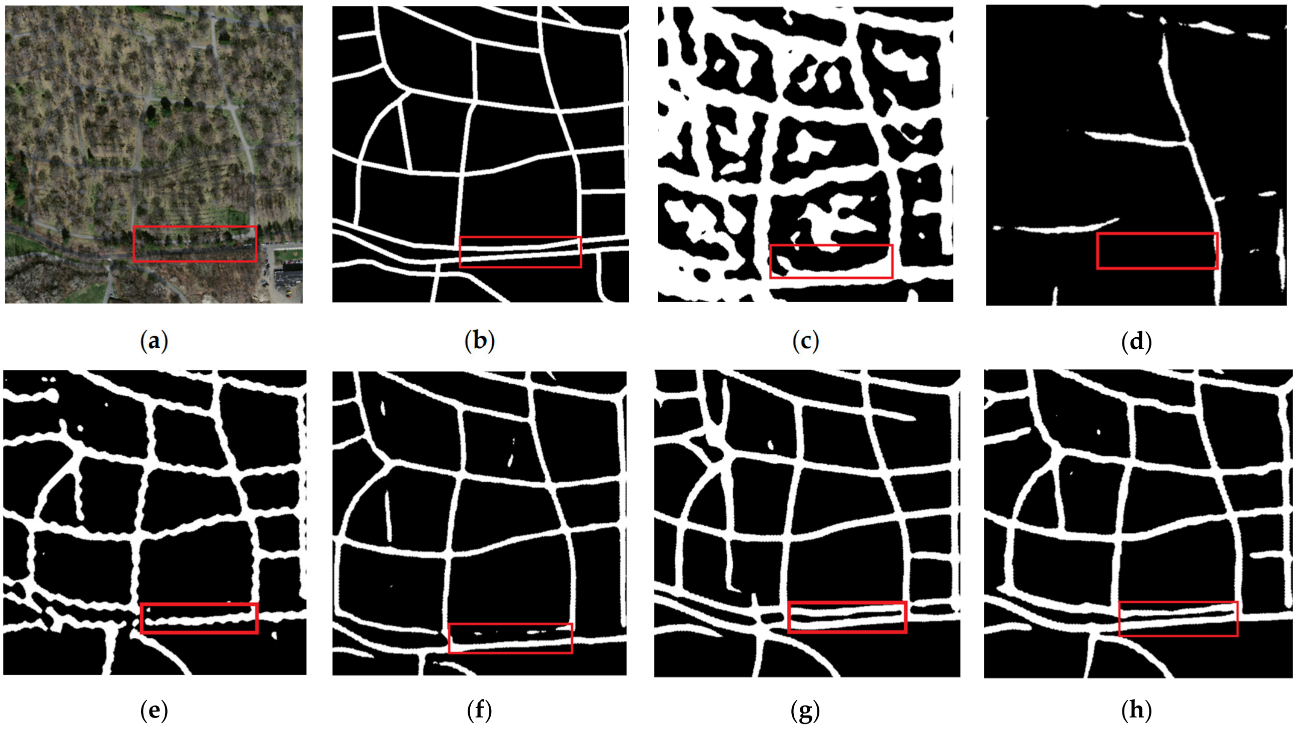


Figure 7. Results of road extraction in occlusion scene. (**a**) Image; (**b**) Label; (**c**) DeepLabv3+; (**d**) U-Net; (**e**) HRNetV2; (**f**) D-LinkNet; (**g**) RefineNet; and (**h**) MSPFE-Net.(Red squares represent key areas.).

In Figure 8, it can be found that the background of some areas is similar to the road, and the MSFPE-Net basically accurately extracts the road, while other comparison models mistakenly regard the road as the background.

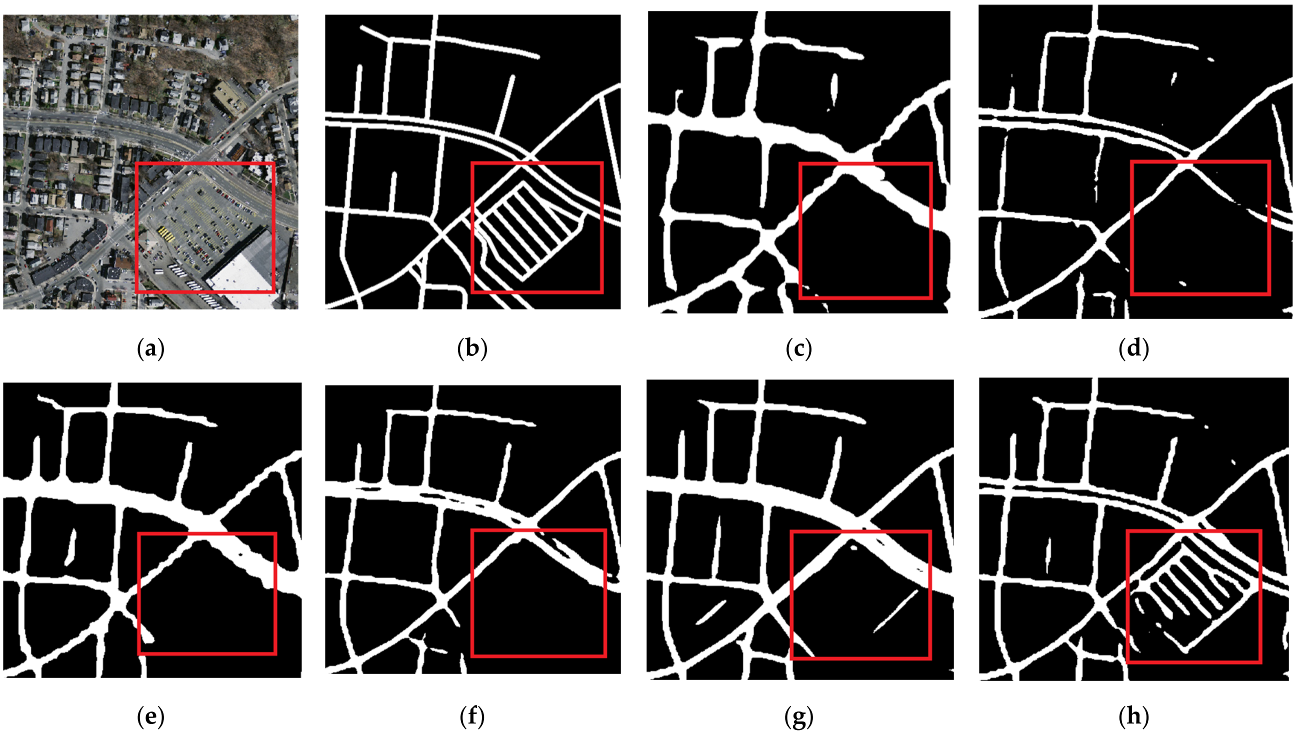


Figure 8. Results of road extraction (the road is similar to the background). (**a**) Image; (**b**) Label; (**c**) DeepLabv3+; (**d**) U-Net; (**e**) HRNetV2; (**f**) D-LinkNet; (**g**) RefineNet; and (**h**) MSPFE-Net.(Red squares represent key areas.).

This method uses multi-level strip pooling combined with a feature enhancement module to ensure road connectivity and road edge details. The goal of the multi-level strip pooling module is to obtain the global context information and long-range dependencies and connect the discretely distributed paths in the image. The feature enhancement module is used to obtain the road's local context information and improve the road edge's segmentation effect.

5. Discussion

5.1. Ablation Experiments

5.1.1. Influence of MSPM and FEM

We conducted a series of ablation experiments on MSPFE-NET using the Massachusetts Road Dataset. In order to prove the effectiveness of each module, the baseline is U-Net with ResNet50 as the backbone, and then each module is added separately. Table 2 shows the experimental data of all main modules.

Table 2. Ablation experiments of MSPM and FEM.

Networks	Precision (%)	Recall (%)	F1 (%)	IoU (%)
Baseline	75.92%	63.28%	69.02%	52.70%
Baseline + MSPM (L1)	72.54%	70.23%	71.37%	55.48%
Baseline + MSPM (L1 + L2)	68.87%	77.95%	73.13%	57.64%
Baseline + MSPM (L1 + L2 + L3)	69.87%	77.65%	73.55%	58.17%
Baseline + FEM	68.71%	71.76%	70.20%	54.08%

After adding MSPM (L1) to the baseline, Recall, F1, and IoU were enhanced by 6.95%, 2.35%, and 2.78%, respectively. After adding L2 to Baseline + MSPM (L1), the Recall of F1 and IoU of Baseline + MSPM (L1 + L2) increased by 7.72%, 1.76%, and 2.16% compared with Baseline + MSPM (L1), respectively. After adding L3 to Baseline + MSPM (L1 + L2), the Precision, F1, and IoU of Baseline + MSPM (L1 + L2) increased by 1.00%, 0.42%, and 0.53% over Baseline + MSPM (L1 + L2), respectively. According to all the results, as shown in Figure 9, with the addition of strip pooling of different levels (L1, L2, L3) into the model, the overall connection of the road is better, and the form of the slender road is higher, which fully verifies the effectiveness of MSPM. After adding FEM to the baseline, Recall, F1, and IoU enhanced by 8.48%, 1.18%, and 1.38%, respectively. This proves that FEM plays a certain role in improving road extraction capability.

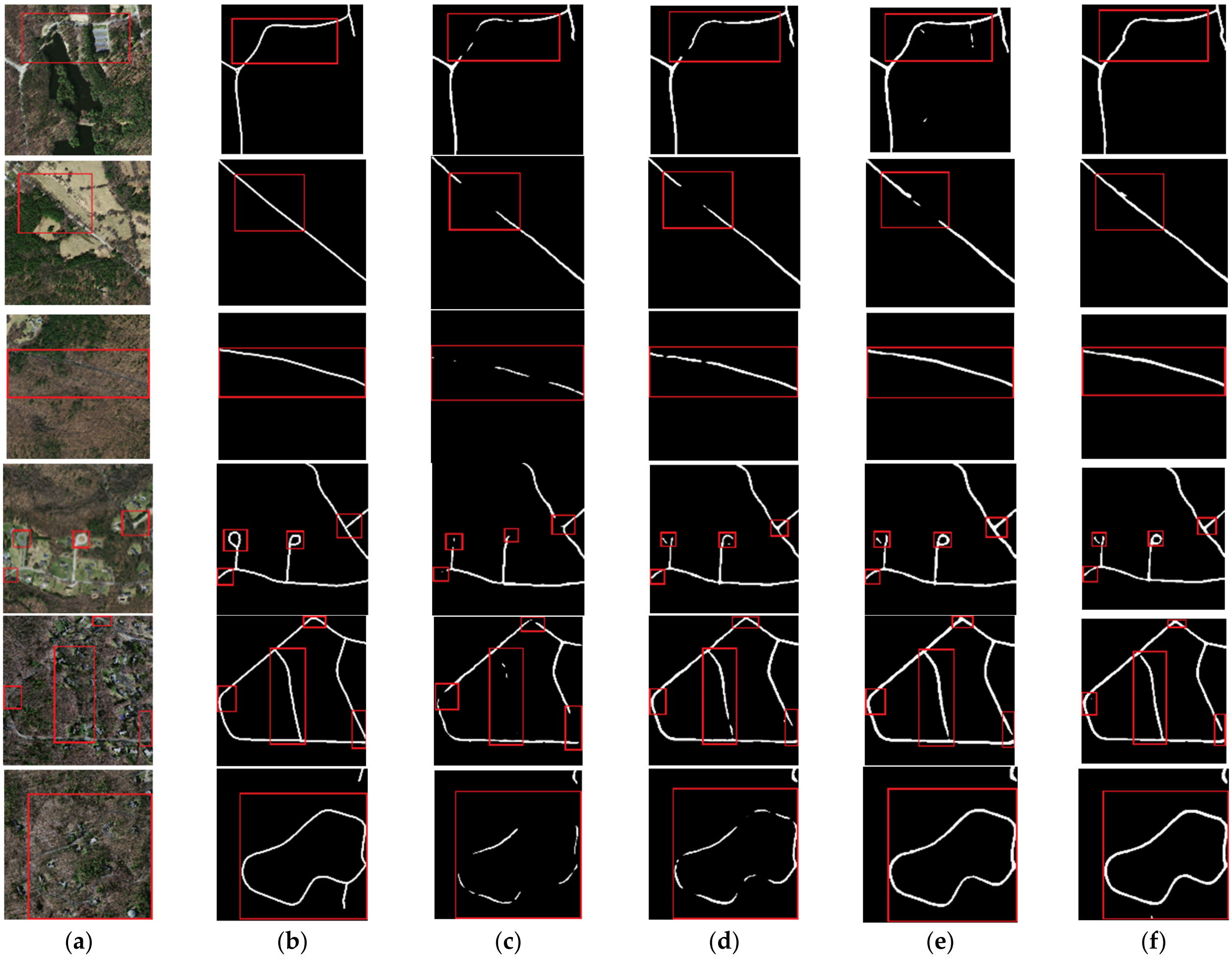


Figure 9. Comparison of adding MSPM and FEM. (**a**) Image; (**b**) Label; (**c**) Baseline; (**d**) Baseline + MSPM (L1); (**e**) Baseline + MSPM (L1 + L2); and (**f**) Baseline + MSPM (L1 + L2 + L3).(Red squares represent key areas.).

5.1.2. Comparison of Loss Function

In order to prove the effectiveness of the dice coefficient loss function and focal loss function, the baseline is MSPFE-Net. Table 3 shows the experimental results.

Table 3. Ablation experiments of MSPM and FEM.

Networks	Precision (%)	Recall (%)	F1 (%)	IoU (%)
Baseline + cross entropy loss function	79.62%	65.84%	72.08%	56.34%
Baseline + focal loss function	77.12%	68.04%	72.30%	56.61%
Baseline + dice coefficient loss function	73.33%	74.37%	73.84%	58.53%

According to Table 3, after the addition of the focal loss function, the evaluation metrics have not changed much. When the dice coefficient loss function was used as a loss function alone, the evaluation metrics increased significantly, with IoU increasing by 2.19%.

6. Conclusions

MSPFE-Net is designed and implemented to extract roads, which can extract narrow roads and also restore roads that are covered by trees or shadows. When the road is similar to the background, the MSPFE-Net basically accurately extracts the road. MSPFE-Net ensures the connectedness and accuracy of the road. MSPFE-Net utilizes a multi-level strip pooling module to collect context information for road extraction. This module incorporates both horizontal and vertical strip pooling operations to gather context information of different levels and long-range dependencies. Due to the full acquisition of context information, the continuity of the road is improved. In areas with dense roads, the MSPFE-Net uses a feature enhancement module to collect local context information and enhance the segmentation effect of the road edge. Experimental results show that MSPFE-Net is better than other comparative models in experiments on evaluation metrics and results from images. Although MSPFE-Net has basically completed the task of road segmentation, roads are similar to the background in some areas, and there are also a few discontinuous roads.

Author Contributions: Conceptualization, Z.W. and Z.Z.; methodology, Z.W.; software, Z.W.; validation, Z.W.; formal analysis, Z.W. and Z.Z.; investigation, Z.W. and Z.Z.; writing—original draft preparation, Z.W.; writing—review and editing, Z.W.; supervision, Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors appreciate the constructive comments and suggestions from the reviewers that helped improve the quality of this manuscript.

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

References

1. Xu, Y.; Xie, Z.; Feng, Y.; Chen, Z. Road Extraction from High-Resolution Remote Sensing Imagery Using Deep Learning. *Remote Sens.* **2018**, *10*, 1461. [CrossRef]
2. Li, Y.; Guo, L.; Rao, J.; Xu, L.; Jin, S. Road Segmentation Based on Hybrid Convolutional Network for High-Resolution Visible Remote Sensing Image. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 613–617. [CrossRef]
3. Bong, D.; Lai, K.C.; Joseph, A. Automatic Road Network Recognition and Extraction for Urban Planning. *Int. J. Appl. Sci. Eng. Technol.* **2009**, *5*, 209–215.
4. Hinz, S.; Baumgartner, A.; Ebner, H. Modeling Contextual Knowledge for Controlling Road Extraction in Urban Areas. In Proceedings of the IEEE/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas (Cat. No.01EX482), Rome, Italy, 8–9 November 2001; pp. 40–44.
5. Ma, H.; Lu, N.; Ge, L.; Li, Q.; You, X.; Li, X. Automatic Road Damage Detection Using High-Resolution Satellite Images and Road Maps. In Proceedings of the 2013 IEEE International Geoscience and Remote Sensing Symposium—IGARSS, Melbourne, VIC, Australia, 21–26 July 2013; pp. 3718–3721.
6. Li, Q.; Zhang, J.; Wang, N. Damaged Road Extraction from Post-Seismic Remote Sensing Images Based on Gis and Object-Oriented Method. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 4247–4250.
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.; Wang, M. Very High Resolution Remote Sensing Image Classification with SEEDS-CNN and Scale Effect Analysis for Superpixel CNN Classification. *Int. J. Remote Sens.* **2018**, *40*, 506–531. [CrossRef]
9. Cheng, G.; Zhu, F.; Xiang, S.; Pan, C. Accurate Urban Road Centerline Extraction from VHR Imagery via Multiscale Segmentatio-n and Tensor Voting. *Neurocomputing* **2016**, *205*, 407–420. [CrossRef]

10. Yang, M.; Yuan, Y.; Liu, G. SDU-Net: Road Extraction via Spatial Enhanced and Densely Connected U-Net. *Pattern Recognit.* **2022**, *126*, 108549. [CrossRef]
11. 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]
12. Zhang, X.; Ma, W.; Li, C.; Wu, J.; Tang, X.; Jiao, L. Fully Convolutional Network-Based Ensemble Method for Road Extraction From Aerial Images. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 1777–1781. [CrossRef]
13. Chen, Z.; Deng, L.; Luo, Y.; Li, D.; Marcato Junior, J.; Nunes Gonçalves, W.; Awal Md Nurunnabi, A.; Li, J.; Wang, C.; Li, D. Road Extraction in Remote Sensing Data: A Survey. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102833. [CrossRef]
14. Hou, Q.; Zhang, L.; Cheng, M.-M.; Feng, J. Strip Pooling: Rethinking Spatial Pooling for Scene Parsing. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
15. Raziq, A.; Xu, A.; Li, Y. Automatic Extraction of Urban Road Centerlines from High-Resolution Satellite Imagery Using Automa-tic Thresholding and Morphological Operation Method. *J. Geogr. Inf. Syst.* **2016**, *8*, 517–525. [CrossRef]
16. Haverkamp, D.S. Extracting Straight Road Structure in Urban Environments Using IKONOS Satellite Imagery. *Opt. Eng.* **2002**, *41*, 2107–2110. [CrossRef]
17. Wenfeng, W.; Shuhua, Z.; Yihao, F.; Weili, D. Parallel Edges Detection from Remote Sensing Image Using Local Orientation Co-ding. *Acta Opt. Sin.* **2012**, *32*, 0315001. [CrossRef]
18. 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]
19. Mnih, V.; Hinton, G.E. Learning to Detect Roads in High-Resolution Aerial Images. In Proceedings of the Computer Vision—ECCV 2010, Heraklion, Greece, 5–11 September 2010; Daniilidis, K., Maragos, P., Paragios, N., Eds.; Springer: Berlin, Heidelberg, 2010; pp. 210–223.
20. Li, P.; Zang, Y.; Wang, C.; Li, J.; Cheng, M.; Luo, L.; Yu, Y. Road Network Extraction via Deep Learning and Line Integral Convolution. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 1599–1602.
21. 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]
22. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651.
23. 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.
24. Singh, P.; Dash, R. A Two-Step Deep Convolution Neural Network for Road Extraction from Aerial Images. In Proceedings of the 2019 6th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 7–8 March 2019; pp. 660–664.
25. He, K.; Zhang, X.; Ren, S.; 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.
26. Zhang, Z.; Liu, Q.; Wang, Y. Road Extraction by Deep Residual U-Net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
27. Gao, X.; Sun, X.; Zhang, Y.; Yan, M.; Xu, G.; Sun, H.; Jiao, J.; Fu, K. An End-to-End Neural Network for Road Extraction From Remote Sensing Imagery by Multiple Feature Pyramid Network. *IEEE Access* **2018**, *6*, 39401–39414. [CrossRef]
28. 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]
29. Yingxiao, X.; Chen, H.; Du, C.; Li, J. MSACon: Mining Spatial Attention-Based Contextual Information for Road Extraction. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–17. [CrossRef]
30. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv* **2015**, arXiv:1511.07122.
31. Zhou, L.; 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), IEEE, Salt Lake City, UT, USA, 18–22 June 2018; pp. 192–1924.
32. He, H.; Yang, D.; Wang, S.; Wang, S.; Li, Y. Road Extraction by Using Atrous Spatial Pyramid Pooling Integrated Encoder-Decoder Network and Structural Similarity Loss. *Remote Sens.* **2019**, *11*, 1015. [CrossRef]
33. Chen, R.; Hu, Y.; Wu, T.; Peng, L. Spatial Attention Network for 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. 1841–1844.
34. Tao, C.; Qi, J.; Li, Y.; Wang, H.; Li, H. Spatial Information Inference Net: Road Extraction Using Road-Specific Contextual Information. *ISPRS J. Photogramm. Remote Sens.* **2019**, *158*, 155–166. [CrossRef]
35. Zhou, M.; Sui, H.; Chen, S.; Wang, J.; Chen, X. BT-RoadNet: A Boundary and Topologically-Aware Neural Network for Road Extraction from High-Resolution Remote Sensing Imagery. *ISPRS J. Photogramm. Remote Sens.* **2020**, *168*, 288–306. [CrossRef]
36. Lu, X.; Zhong, Y.; Zheng, Z.; Zhang, L. GAMSNet: Globally Aware Road Detection Network with Multi-Scale Residual Learning. *ISPRS J. Photogramm. Remote Sens.* **2021**, *175*, 340–352. [CrossRef]
37. Tan, X.; Xiao, Z.; Wan, Q.; Shao, W. Scale Sensitive Neural Network for Road Segmentation in High-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 533–537. [CrossRef]

38. Zhu, Q.; Zhang, Y.; Wang, L.; Zhong, Y.; Guan, Q.; Lu, X.; Zhang, L.; Li, D. A Global Context-Aware and Batch-Independent Network for Road Extraction from VHR Satellite Imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *175*, 353–365. [CrossRef]
39. Wang, S.; Yang, H.; Wu, Q.; Zheng, Z.; Wu, Y.; Li, J. An Improved Method for Road Extraction from High-Resolution Remote-Sensing Images That Enhances Boundary Information. *Sensors* **2020**, *20*, 2064. [CrossRef]
40. Mnih, V. Machine Learning for Aerial Image Labeling. Ph.D. Thesis, University of Toronto, Toronto, ON, Canada, 2013.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.