applied sciences

MDPI

Article

HRU-Net: High-Resolution Remote Sensing Image Road Extraction Based on Multi-Scale Fusion

Anchao Yin †, Chao Ren *,†, Zhiheng Yan, Xiaoqin Xue, Weiting Yue, Zhenkui Wei, Jieyu Liang, Xudong Zhang and Xiaoqi Lin

College of Geomatics and Geoinformation, Guilin University of Technology, Guilin 541004, China; yinanchao@glut.edu.cn (A.Y.); yanzhiheng@glut.edu.cn (Z.Y.); 2120222038@glut.edu.cn (X.X.); yueweiting@glut.edu.cn (W.Y.); weizhenkui@glut.edu.cn (Z.W.); liangjieyu@glut.edu.cn (J.L.); 1020211828@glut.edu.cn (X.Z.); linxiaoqi@glut.edu.cn (X.L.)

* Correspondence: renchao@glut.edu.cn

† These authors contributed equally to this work.

Abstract: Road extraction from high-resolution satellite images has become a significant focus in the field of remote sensing image analysis. However, factors such as shadow occlusion and spectral confusion hinder the accuracy and consistency of road extraction in satellite images. To overcome these challenges, this paper presents a multi-scale fusion-based road extraction framework, HRU-Net, which exploits the various scales and resolutions of image features generated during the encoding and decoding processes. First, during the encoding phase, we develop a multi-scale feature fusion module with upsampling capabilities (UMR module) to capture fine details, enhancing shadowed areas and road boundaries. Next, in the decoding phase, we design a multi-feature fusion module (MPF module) to obtain multi-scale spatial information, enabling better differentiation between roads and objects with similar spectral characteristics. The network simultaneously integrates multi-scale feature information during the downsampling process, producing high-resolution feature maps through progressive cross-layer connections, thereby enabling more effective high-resolution prediction tasks. We conduct comparative experiments and quantitative evaluations of the proposed HRU-Net framework against existing algorithms (U-Net, ResNet, DeepLabV3, ResUnet, HRNet) using the Massachusetts Road Dataset. On this basis, this paper selects three network models (U-Net, HRNet, and HRU-Net) to conduct comparative experiments and quantitative evaluations on the DeepGlobe Road Dataset. The experimental results demonstrate that the HRU-Net framework outperforms its counterparts in terms of accuracy and mean intersection over union. In summary, the HRU-Net model proposed in this paper skillfully exploits information from different resolution feature maps, effectively addressing the challenges of discontinuous road extraction and reduced accuracy caused by shadow occlusion and spectral confusion factors. In complex satellite image scenarios, the model accurately extracts comprehensive road regions.

Keywords: high-resolution remote sensing images; road extraction; shadow occlusion; spectral confusion; multi-scale fusion

check for updates

Citation: Yin, A.; Ren, C.; Yan, Z.; Xue, X.; Yue, W.; Wei, Z.; Liang, J.; Zhang, X.; Lin, X. HRU-Net: High-Resolution Remote Sensing Image Road Extraction Based on Multi-Scale Fusion. *Appl. Sci.* **2023**, *13*, 8237. https://doi.org/10.3390/app13148237

Academic Editors: Qizhi Xu, Jin Zheng and Feng Gao

Received: 15 June 2023
Revised: 9 July 2023
Accepted: 14 July 2023
Published: 15 July 2023

1. Introduction

As an essential component of transportation, roadways exert a substantial impact on various aspects of contemporary life, encompassing urban and rural development, traffic management, and autonomous vehicle navigation [1]. In tandem with the evolution of satellite remote sensing technology, high-resolution satellite imagery has arisen as a crucial asset for digital image processing in the modern era [2]. As a result, the pursuit of high-precision road extraction from high-resolution satellite imagery has attracted considerable scholarly attention in recent years. However, complications from factors such as illumination, shadow occlusion [3], and noise yield divergent features among identical road targets, while spectral ambiguity resulting from the influence of neighboring materials (non-road targets displaying road-like properties) intensifies the difficulty of accurately and comprehensively extracting roads from high-resolution satellite imagery [4].

Lately, in conjunction with the swift advancement of deep learning, an increasing number of researchers have employed deep learning methods within the domains of image categorization and semantic partitioning [5]. Convolutional neural networks have demonstrated their efficacy in road extraction tasks due to their remarkable feature extraction capabilities [6]. In comparison to conventional approaches requiring the manual extraction of superficial features, deep learning methods not only enhance the precision of road extraction but also display resilient performance in large-scale extraction efforts [7].

Mnih et al. [8] pioneered the implementation of deep learning techniques in road extraction, proposing a method employing restricted Boltzmann machines for identifying roads in high-resolution satellite imagery. Subsequently, Long et al. [9] presented fully convolutional networks (FCNs), effectively addressing semantic segmentation in images. By replacing fully connected layers with standard convolutions, FCNs enabled the shift from elementary classification to pixel-level classification. Unfortunately, the continuous downsampling operations in FCNs lead to the loss of numerous fine-grained details on small feature maps, complicating the restoration of the original resolution and resulting in segmentation challenges, such as blurred boundaries, indistinct edges, and insufficient feature representation [10].

To counter these limitations, Ronneberger et al. introduced the U-Net, integrating supplementary skip connections into an FCN to bolster the network's ability to manage intricate details. U-Net pioneered the encoder–decoder architecture, in which the encoder reduces spatial dimensions and captures spatial detail information, while the decoder reuses low-level features from the encoding stage to progressively restore input size and recover spatial position information [11]. This approach adeptly retrieves a wealth of spatial detail information and optimally exploits road texture information. However, as the depth of the U-Net increases, issues such as gradient vanishing and explosion may occur. Consequently, existing research focuses on enhancing network performance and stability by leveraging the U-Net's unique encoding and decoding architecture, while simultaneously refining the network model [12]. For example, Zhang et al. [13] combined the advantages of U-Net and ResNet to create the ResUNet, promoting information propagation through abundant contract connections while fully capitalizing on the superior gradient propagation stability of the residual connection structure. Oktay et al. [14] incorporated an attention mechanism into U-Net, yielding the attention U-Net model. This model highlights segmented targets by suppressing feature responses in irrelevant background areas, thereby improving segmentation accuracy and robustness. Furthermore, Zhou et al. [15] developed the D-LinkNet model, which integrates dilated convolution modules into the U-Net model to expand the receptive field and enhance the recognition capacity of large-scale objects in the image by scaling and cropping the input image to various dimensions [16]. Nonetheless, the capacity to discern multi-scale objects in the image remains insufficient, potentially leading to the loss of road details, false positives, and false negatives, which negatively impact the road segmentation outcome. To address these constraints, several empirical studies [2,17–21] have demonstrated that road extraction algorithms can be improved through the fusion of multi-scale spatial features.

For fixed-scale input images, the information acquired by feature extraction operators and classifiers remains constant. Insufficient information can result in improper classification, while excessive information can impede target identification. Consequently, Chen et al. [22] proposed the DeepLab series of networks, utilizing dilated convolutions instead of upsampling and combining dilated convolutions with atrous spatial pyramid pooling modules to augment the receptive field and obtain multi-scale features, thus enhancing the model's multi-scale prediction capabilities. However, the feature extraction method based on dilated convolutions can easily lose information related to small-scale targets, as the convolution kernels solely extract features from restricted regions. Wang et al. [23] introduced the HRNet, which employs a multi-branch structure to simultaneously maintain high- and low-resolution features and repeatedly performs multi-scale fusion to generate rich high-resolution representations. Nevertheless, in occluded images, some resolutions may exhibit information deficits due to partial regions being obscured, ultimately affecting the effectiveness of the resulting feature maps.

To effectively address the aforementioned challenges and further enhance road extraction performance, this paper presents a deep learning network model, HRU-Net, specifically tailored for road features in remote sensing imagery, drawing inspiration from U-Net and HRNet. The HRU-Net model inherits and incorporates the encoder–decoder architecture of the U-Net, establishing an information propagation pathway for replicating low-level features to their corresponding high-level representations, thus enriching high-level semantic features with the finer details of low-level features. The model employs a parallel structure within the sub-network to simultaneously preserve high- and low-resolution semantic information. Concurrently, multiple UMR and MPF modules are designed between the encoding and decoding components to optimally exploit multi-scale information. Both modules progressively combine feature maps of different resolutions, integrating the global contextual information of low-resolution feature maps with the robust detail information of high-resolution feature maps to generate high-resolution feature map representations. Furthermore, the different resolution feature maps acquired after fusion by the two modules undergo frequent information exchange via a parallel structure, enhancing the utilization of advantageous information, such as regions and boundaries, and mitigating the impact of extraneous information, such as shadow occlusion. This process yields a high-resolution feature map through consistent information interaction, ensuring the final prediction results closely approximate pixel-level accuracy and achieve a more precise local feature discrimination capacity.

The proposed HRU-Net offers the following contributions: the design of UMR and MPF modules within the network. The UMR module is a multi-scale fusion module with upsampling functionality, merging features of varying resolutions post-upsampling to generate larger resolution features for subsequent input into the subnet. In the MPF module, features of identical resolution are combined and decoded synchronously via upsampling. Simultaneously, the overarching network employs a parallel structure to maintain the parallelism of low-resolution and high-resolution feature maps, continuously executing multi-scale fusion operations to acquire multi-scale information and perpetually exchanging information between different resolutions. Through incessant information exchange, the model more effectively integrates multi-scale semantic information, considering the semantic information of high and low features, thereby augmenting the expressive capacity of the network model. In comparison to existing models such as U-Net, ResNet, DeepLabV3, ResUnet, and HRNet, the HRU-Net model's unique parallel connection structure and continuous multi-scale feature integration and information exchange not only preserve detailed information but also capture global features, achieving more accurate road recognition.

The content arrangement of this paper in subsequent sections is as follows: Section 2 provides detailed information regarding the road detection network proposed in this paper; Section 3 presents the experimental results, encompassing data introduction, experimental settings, and result analysis; Section 4 constitutes the discussion segment; Section 5 offers concluding remarks.

2. Methods

2.1. U-Net

The U-Net derives its name from its distinctive "U" shape structure, which was conceived and adapted from the FCN network architecture [24]. In contrast to the FCN network, the salient feature of the U-Net is its ability to generate a higher-precision image segmentation model with a reduced quantity of training images, attributable to its unique "U" shape structure. The left structure of the network constitutes the encoder, employing convolution and pooling operations to downsample the input image, obtaining contextual information, and generating feature maps to extract rudimentary information. The right structure represents the decoder, executing upsampling operations on the feature maps to acquire more profound features. The central structure is the skip connection component, fusing deep and shallow feature maps procured at the same stage to form a feature-enhanced network. The network framework is implemented using convolution, pooling, activation, and normalization techniques [25].

(1) Convolution: Convolution is the process of filtering the input image or feature map to extract its feature information [26]. In the U-Net, convolution operations usually use a 3 × 3 convolution kernel to convolve the input feature map and obtain the output feature map. The convolution operation can be expressed by Equation (1):

$$y = f\left(\sum_{i=1}^{n} X_i * W_i + b\right) \tag{1}$$

where X_i represents the input feature map, W_i represents the convolution kernel, b represents the bias term, $*$ represents the convolution operation, n represents the number of convolution kernels, and f represents the activation function.

(2) Pooling: Pooling operations involve the downsampling of input feature maps, decreasing the resolution and dimensions to enhance the computational efficiency and diminish the complexity [27]. Pooling techniques can be categorized into two primary types: max pooling and average pooling. Max pooling slides a fixed-size window across the input feature map, selecting the maximum value within the window for output. This operation accentuates prominent features while effectively reducing spatial resolution. Conversely, average pooling utilizes the average value within the window as output, smoothing the input feature map's information and efficiently lowering the spatial resolution. In the U-Net, 2 × 2 max pooling is typically employed, extracting the maximum value from four adjacent pixels in the input feature map.

(3) Activation Function: The activation function applies a nonlinear transformation to the outcomes of convolution and pooling processes, bolstering their expressive capabilities. Within the HRU-Net, the rectified linear unit (ReLU) function is employed, offering rapid convergence and enhanced generalization capacity [28]. The ReLU function can be expressed by Equation (2):

$$f(x) = max(0, x) \tag{2}$$

(4) Normalization Layer: The normalization layer is a prevalent structure in neural networks, responsible for normalizing the input data to stabilize and regulate the distribution. In this study, batch normalization was employed, normalizing each layer's input data within the network [29]. This process results in more stable data distribution, facilitating model training acceleration and enhancing generalization capabilities.

(5) Upsampling: Upsampling is the process of enlarging low-resolution feature maps into high-resolution ones [30]. Typically, it is used alongside downsampling as a conventional feature extraction technique. Within the realm of deep learning, two principal techniques are utilized for the operation of upsampling, namely transposed convolution (also referred to as deconvolution) and bilinear interpolation. In the current investigation, our choice fell upon bilinear interpolation to serve as the upsampling module. This preference is rooted in several key factors: Firstly, bilinear interpolation is more computationally efficient and faster compared to transposed convolution. Secondly, bilinear interpolation is devoid of any parameters that require learning, thereby simplifying the model and reducing the potential for overfitting. Lastly, it avoids the so-called "checkerboard effect" that can result from transposed convolution, thereby ensuring a smoother output image [31].

Thus, through employing bilinear interpolation for image enlargement in our research, we have managed to preserve, to a certain extent, image smoothness and detail information.

This has had the effect of enhancing the network's resolution, thus augmenting the model's accuracy and stability.

2.2. HRNet

A high-resolution network (HRNet) is a convolutional neural network that was introduced in 2019 for high-resolution image classification and segmentation. It enhances the quality of feature representation through multi-resolution feature extraction and fusion while bolstering computational efficiency via a cross-branch network structure [32]. HRNet employs a parallel branch configuration, directing input images to multiple sub-networks for feature extraction. Subsequently, it interacts with and fuses different resolution features, performing classification or segmentation through global pooling and fully connected layers [33]. HRNet's exceptional performance stems from the application of techniques, such as multi-resolution feature extraction and fusion, and cross-branch network structures, which effectively elevate the quality of feature representation and computational efficiency.

2.3. HRU-Net

2.3.1. Multi-Feature Module Construction

In delivering greater road details, high-resolution remote sensing images also expose distinctions among roads at varying levels, such as road materials, grades, and surrounding environments. To further augment the network's capacity for detail expression and feature extraction, we initially contemplated utilizing a "U"-shaped network structure in constructing our method. This structure performs four downsampling operations in the encoder, broadening the receptive field range. It enables features to propagate between low-level and high-level features more straightforwardly, fosters backpropagation in network model training, and supplements additional detail information to high-level semantic features. Nevertheless, while expanding the receptive field through downsampling, some road detail information is lost, proving challenging to recover or even vanishing during forward propagation in the network. Consequently, we designed the UMR and MRF modules to concurrently consider multiple levels of semantic features, capture multiscale information, and contemplate local road detail information.

(a) UMR Module

In this study, as the extraction level deepened, the semantic features obtained gradually deepened and became more abstract. We designed four multi-feature fusion modules (UMRs) in the encoding phase, which fused feature maps from different levels of the network into a high-resolution feature map through upsampling techniques. This design helps to capture and utilize abstract features at each level, thereby enhancing the model's feature expression capabilities.

As shown in Figure 1, the UMR module first received feature maps from the previous and the same layer of the network as inputs. Subsequently, we performed bilinear interpolation on the small-sized feature map from the next layer of the network, aligning its size with the other inputs. Afterward, we performed a Concat operation on all adjusted feature maps to merge and further extract features. During the forward propagation process of the model, we used parallel downsampling strategies to process the high-resolution feature map fused after four UMR modules, thereby reducing the size of the output feature map. At the same time, we used the output of this module as the input for path 1 of the next UMR module in order to repeat the Concat connection and downsampling operations, gradually enhancing the resolution of the feature map.

(b) MRF Module

In the decoder part of our study, we set up four multi-resolution feature (MRF) modules to cascade feature maps at different levels for feature extraction. As depicted in Figure 2, the MRF module receives the output feature map of the upsampling multi-feature fusion (UMR) module at the same level in the encoder stage as the input through path 1 and receives the feature map upsampled in the decoder stage through path 2.

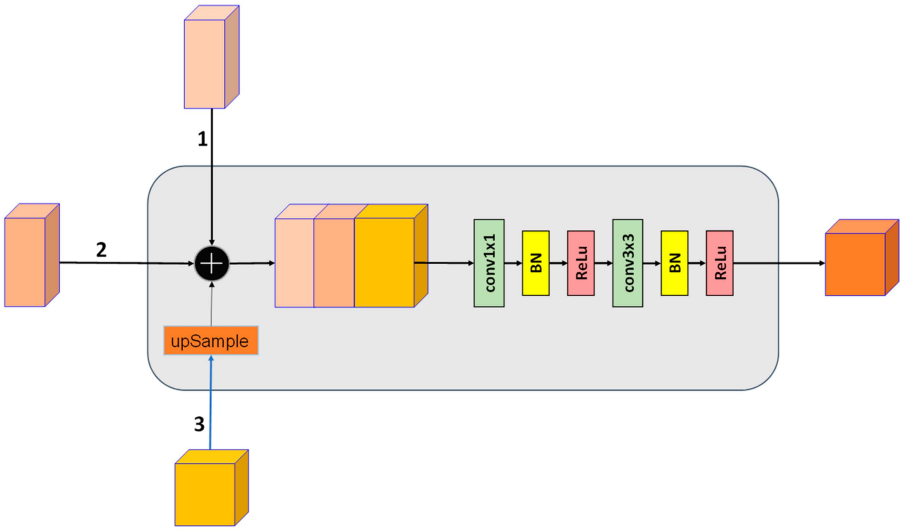

Figure 1. This is the UMR module structure diagram. Different colored cubes represent feature maps of different sizes. The numbers 1, 2, and 3 respectively signify feature map input connections coming from the previous layer, the same layer, and the next layer in the network. Squares represent feature maps.

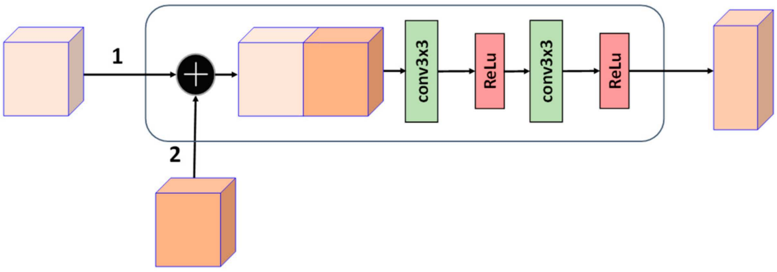

Figure 2. This is the MRF module structure diagram. Different colored cubes represent feature maps of different sizes. The numbers 1 and 2 respectively signify the feature map input connections coming from the same layer and the next layer in the network. Squares represent feature maps.

Initially, two feature maps of the same size were input for a Concat operation, and feature extraction was performed through a combination of two 3 × 3 convolutions and ReLU activation functions. Subsequently, the output of this module was connected with the output of the UMR module at the next level and used as the input for the next MRF module for concatenation and feature extraction operations.

In this process, we chose the Concat operation over summation because concatenation can retain the original information of the feature maps at each level, avoiding mutual interference of features. Our choice of the combination of 3 × 3 convolution and the ReLU activation function aimed to achieve a good feature extraction effect while keeping the number of parameters under control.

This design allowed the decoder to effectively utilize multi-scale feature maps generated by the encoder, and the MRF modules iteratively fused and upsampled the feature maps. By concatenating the feature maps at different levels, the network can better capture both high-level semantic information and low-level detail information. This approach enhanced the network's ability to handle complex road scenarios in high-resolution remote sensing images, leading to more accurate road extraction and segmentation results.

2.3.2. Realization Principle

The dual multi-tiered feature integration frameworks (UMR module and MRF module) delineated in this study employ convolution, pooling, activation, and normalization techniques to diminish the primary attributes of the image while utilizing upsampling to augment resolution and intricate specifics. These frameworks assimilate data from diverse resolutions, merging the granular high-level semantic elements with delicate minutiae to generate high-resolution feature mappings that concurrently contemplate both categories of attributes.

2.3.3. Network Structure

The multi-feature fusion module was integrated into the standard U-Net architecture to create a multi-scale fusion network—the HRU-Net. The HRU-Net structure is depicted in Figure 3.

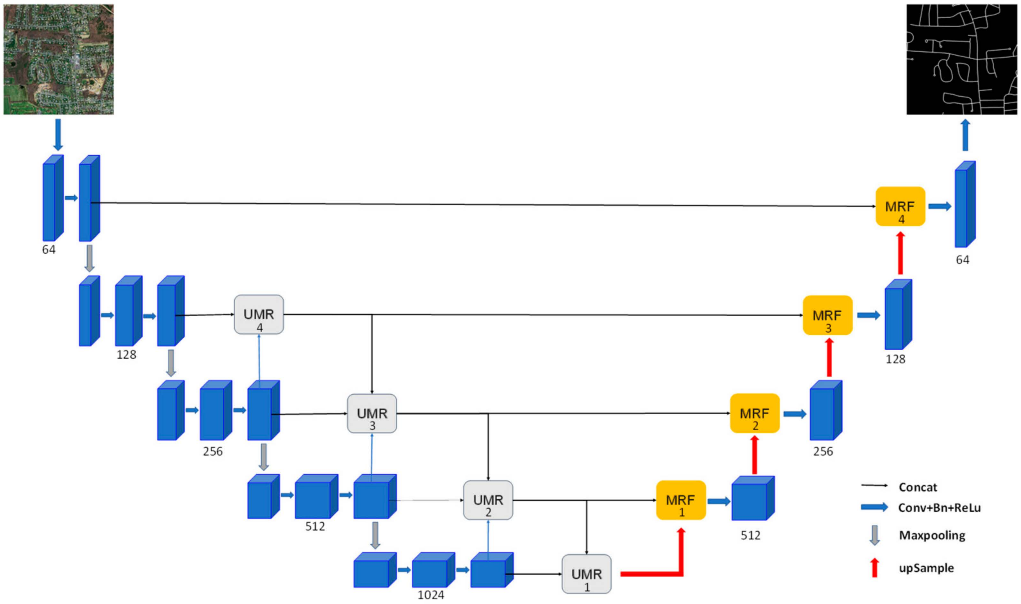

Figure 3. HRU-Net structure. Given a remote sensing image, abstract features are extracted from the input to obtain the final pixel-wise road map prediction. Each blue box represents a multi-channel feature map, with the number of channels annotated beneath the box. Gray boxes symbolize operations from UMR modules, while orange-yellow boxes correspond to operations from MRF modules. These are arranged in ascending order from the bottom of the model upwards, numbered 1, 2, 3, 4. Arrows indicate distinct operations.

3. Experiments and Results

3.1. Dataset Descriptions

3.1.1. Massachusetts Road Dataset

The Massachusetts Road Dataset, globally recognized as the most extensive road dataset accessible to the public, was established by MiHn and Hinton [34]. This dataset covers approximately 500 square kilometers, includes a diverse array of geographic regions from urban to rural, and encompasses an assortment of road types such as highways, rural dirt roads, and asphalt roads. Notably, it also includes confounding factors, such as rivers and railways, which can mimic the appearance of roads. The dataset comprises 1171 images, each with a per-pixel resolution of 1.2 m and dimensions of 1500 × 1500 pixels. The dataset is categorized into 1108 training images, 14 validation images, and 49 test images. The training and validation sets feature binary image labels, where road pixels are denoted as 1 and background pixels as 0.

In light of hardware constraints and the necessity of optimizing model performance, the utilization of large-sized images for training was deemed inefficient. Therefore, in this study, the 1108 training images and their corresponding labels were segmented into 27,700 sub-images, each with dimensions of 256 × 256 pixels. The model was subjected to a training–validation split at a 9:1 ratio. Consequently, the finalized training set encapsulated 20,776 images, each with dimensions of 256 × 256 pixels, whereas the validation set contained 6924 images of identical dimensions. To prevent overfitting and to guarantee a comprehensive performance evaluation, a test set was compiled, incorporating 63 images from the original dataset, including the 14 validation images and the 49 test images.

3.1.2. DeepGlobe Road Dataset

This dataset comprises images collected from Thailand, Indonesia, and India, cumulatively covering an area of 2220 square kilometers, which includes urban and suburban locales. It consists of 6226 training images, each with a resolution of 1024 × 1024 pixels and a spatial resolution of 0.5 m per pixel. In this dataset, roads are annotated as the foreground, while other objects are marked as the background. The data are partitioned into training, validation, and testing sets at an 8:1:1 ratio, adhering to standard academic practices.

3.2. Experimental Settings

3.2.1. Hyperparameter Settings

The experimental parameters for this study's model were determined based on numerous preliminary experiments and subsequently refined through extensive experimentation. The training process employed cross-validation, wherein both the training and validation sets were input into the model. After each training iteration on the training set, a batch size of data was selected from the validation set to calculate the model's loss and accuracy, optimizing the model's training. Due to GPU computing power constraints, the input image size was adjusted to 256 × 256 pixels, and the computationally efficient Adam optimizer was used to update network parameters. Each batch processed eight images. The learning rate was initially set to 0.0001, with a minimum learning rate of 0.000001, and the number of training iterations was set to 100.

The model in this study was a binary classification problem that performed pixel-wise classification to determine whether each pixel is a road or background. Since roads constitute approximately 10% of the total area in the remote sensing image, this study considered the commonly used cross-entropy loss function for road extraction, while also taking into account the Dice coefficient loss function, typically used in medical image segmentation, to mitigate the issue of imbalanced samples [35].

In the case of binary classification, the calculation formula for the cross-entropy loss function is shown in Equation (3):

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(\hat{y_i}\right) + (1 - y_i)\log\left(1 - \hat{y_i}\right)\right] \tag{3}$$

where y represents the true pixel label value, $\hat{y_i}$ represents the predicted label pixel value by the model, and N represents the number of pixels.

The Dice loss function is a measure of set similarity, typically used to calculate the similarity between two samples, with its values ranging from 0 to 1. In image segmentation tasks, the Dice coefficient is often used to measure the consistency between the predicted segmentation area and the actual segmentation area. The formula for the Dice coefficient as a loss function is shown in Equation (4):

$$L_{dice} = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \tag{4}$$

In the formula, X is the predicted image generated by the model, Y is the true label of the input image, $|X|$ represents the number of pixels in the predicted image, $|Y|$ represents the number of pixels in the true label, and $|X \cap Y|$ represents the intersection between the predicted image and the true label.

The Dice loss function is robust to imbalanced sample problems and can balance the issue of sample imbalance. This is particularly relevant for road extraction tasks where there is a significant imbalance between positive and negative samples. In such cases, if only pixel-level cross-entropy loss is used, the model may be biased towards predicting the class with more samples. The Dice loss function can mitigate this problem to a certain extent.

Finally, in this experiment, the cross-entropy loss and the Dice coefficient loss were combined by adding the two losses together as the loss function of this experiment, which is shown in Equation (5):

$$L_{loss} = L_{BCE} + L_{dice} \tag{5}$$

The binary cross-entropy (BCE) loss and Dice loss are complementary to each other to a certain extent. In the context of road extraction tasks, implementing this novel loss function considers optimizing the substantial amount of background pixels via the BCE loss. Simultaneously, it acknowledges the spatial relationship between positive and negative samples as indicated by the Dice loss, thus effectively handling the issue of imbalanced samples. In the early stages of prediction, when the discrepancy between the predicted result and the true label is significant, the BCE loss can provide a considerable gradient to help the model converge rapidly. As the prediction approaches the actual label, the gradient of BCE loss diminishes. In contrast, the Dice loss can deliver a consistent gradient at this stage, allowing for better optimization of the model. This complementary nature aids in optimizing the model both globally and locally.

3.2.2. Training Environment Description

To evaluate the effectiveness of the proposed HRU-Net model for road extraction in remote sensing images, training, validation, and testing were conducted using the Massachusetts Road Dataset. The experiment was designed within the Pytorch (version 1.11.0) framework and developed using JetBrains PyCharm 2021. Python served as the programming language, while the hardware configuration included an Intel(R) Core(TM) i7-@2.50 GHz processor and an Nvidia GeForce RTX 3060 graphics card for acceleration. The memory size was 12 GB, and the operating system employed was Windows 10.3.2.3.

3.2.3. Evaluation Metrics

To quantify the road extraction performance of our model, we employed a confusion matrix to assess the model's performance in binary classification problems. As shown in Table 1, the labels were categorized into positive and negative samples, while the prediction results were divided into true positives and true negatives (TPs and TNs), as well as false positives and false negatives (FPs and FNs) to represent accurate and inaccurate predictions, respectively. To thoroughly evaluate the model, this study utilized three evaluation metrics: precision (P), recall (R), and intersection over union (IOU). These metrics served to assess the road extraction performance of the network model.

Table 1. Differences between positive and negative samples.

Real Category	**Predictive Results**	
	Road	**Non-Road**
Road	True positive (TP)	False negative (FN)
Non -road	False positive (FP)	True negative (TN)

Precision: This is the proportion of true positive instances amongst all instances predicted by the model as the positive class. In other words, the precision reflects the proportion of actual roads in all instances predicted as roads by the model.

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

Recall: This is the proportion of actual positive samples that were correctly identified by the model, indicating the proportion of correctly labeled roads by the model.

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

Intersection over Union (IoU): IoU is a commonly used evaluation metric in semantic segmentation and object detection. It measures the accuracy of detecting ground objects from a dataset and quantifies the degree of fit between the extracted results and the ground truth labels. The higher the IoU value, the bigger the overlapping region between the results and the ground truth.

$$IOU = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{8}$$

In this experiment, we evaluated the road extraction performance of the HRU-Net model by calculating the precision, recall, and intersection over union (IOU), and compared the experimental results with other advanced remote sensing image road extraction methods.

3.3. Results and Analysis

To assess the road extraction capability of the HRU-Net model, we carried out a comparative analysis between HRU-Net and other cutting-edge segmentation algorithms, such as U-Net, ResNet, DeepLabV3, ResUnet, and HRNet. Throughout the experiment, each compared network, along with our proposed network, utilized the same virtual environment and parameter settings during the training, validation, and prediction process.

3.3.1. Test on Massachusetts Road Dataset

The performance indicators for the six methodologies are presented in Table 2. From the performance comparison table, it is evident that the relatively simple U-Net and HRNet structures exhibited slightly inferior feature extraction performance compared to the other more complex models in the table. The unique residual structure of the ResNet and its deep layer hierarchy provided better semantic expression capabilities. The ResUnet, which combined the residual network and U-Net, offered better network depth than U-Net and superior high-to-low level information transmission ability compared to ResNet, demonstrating enhanced performance in the average accuracy, recall, and average intersection over union metrics. The HRU-Net structure proposed in this paper outperformed all other models in every indicator.

Table 2. Quantitative evaluation of six methods conducted on the Massachusetts Dataset. The closer the index value is to 1, the better the effect. IoU: Intersection over union.

Scheme	Network	Precision (%)	Recall (%)	IoU (%)
One	U-Net	78.82	83.76	77.67
Two	ResNet	78.93	83.25	77.90
Three	Deeplabv3	79.56	83.90	75.45
Four	ResUnet	79.29	83.85	77.97
Five	HRNet	77.96	83.90	77.54
Six	HRU-Net (ours)	80.09	84.85	78.62

To more intuitively compare the road extraction effects of various network models, this paper selected five 512 × 512 images from the test data for display and evaluation. To emphasize the road details, we set the roads in the predicted images to be displayed in red.

Figure 4 displays the road extraction results from different networks. From top to bottom, it presents the original remote sensing image, the original ground truth label, and the predicted results of U-net (Scheme 1), ResNet (Scheme 2), DeepLabV3 (Scheme 3), ResUnet

(Scheme 4), HRNet (Scheme 5), and HRU-Net (Scheme 6). The road extraction analysis primarily consisted of the main road extraction analysis and the detail extraction analysis of small roads. Shadow occlusion and spectral confusion caused by surrounding materials similar to roads often impacted road extraction completeness. In Scheme 1, numerous instances of inadequate edge detail extraction and disconnection occurred during road extraction. Schemes 2–4 employed the residual network structure and attention mechanism to concentrate on more road detail information, which improved the disconnection situation in road extraction, but some roads remained unextracted. Scheme 5 utilized a multi-scale parallel network structure to address incomplete extraction and enhance road connectivity. Scheme 6, proposed in this paper, ensured the completeness of road extraction and resolved the disconnection caused by shadow occlusion in positions 3 and 5, guaranteeing the completeness of the extracted road at the intersection and significantly improving road extraction efficiency and accuracy.

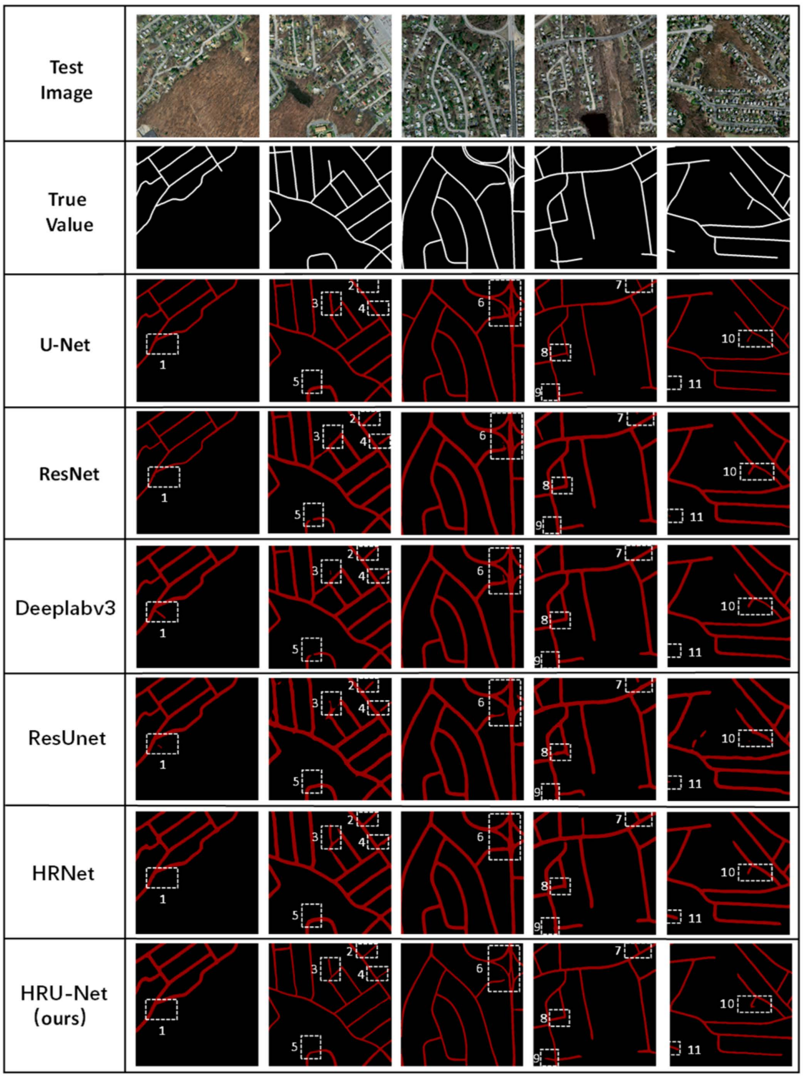

Figure 4. Comparison of road extraction effects of 6 schemes in 5 different test areas in Massachusetts Road Dataset.

To more clearly analyze the effects, we annotated 11 representative road areas in the images for analysis. Areas 1, 3, 7, and 8 were heavily affected by shadow occlusion; Areas 2, 4, and 5 had significant spectral confusion issues; and Areas 6, 9, 10, and 11 were difficult to extract due to the complexity of the road regions. In Scheme 1, the extraction of the roads at Areas 2, 4, 7, 9, 10, and 11 was hindered by the influence of other objects in the surrounding area with similar materials to the roads. Schemes 2–4 incorporated more road details, extracting roads to varying degrees at Areas 2, 4, 7, and 10. Schemes 5 and 6 frequently exchanged information among different scales, allowing them to fully extract the roads at Area 11. In this paper, Scheme 6, after introducing the UMR and MRF modules, was the only model capable of fully extracting the roads at Area 9 while also fully extracting the roads in other areas. The proposed algorithm effectively recovered disrupted roads and improved the completeness of road extraction, particularly in areas with severe obstruction and similar spectral phenomena. In Areas 1 and 8, where the road curvature changed significantly, Schemes 2–5 presented various degrees of noise and misidentification, incorrectly identifying other objects as roads. However, Schemes 1 and 6, which employed a "U" network structure, successfully extracted the roads, showcasing good generalization capabilities. In the most complex area (Area 6) in the tested remote sensing image, where two separate roads intersect at the center, Schemes 2–5 identified the two roads as a single road. Scheme 1 successfully identified the two roads but merged them with the lower left road. Scheme 6 not only separated the two roads better than Scheme 1 but also solved the road merging problem in Schemes 1 and 5. In summary, the proposed road extraction model HRU-Net, which uses a "U" network structure and incorporates multiple UMR and MRF modules, exhibited excellent extraction accuracy and completeness (including the extraction of roads with obstructions such as shadows, trees, narrow rural roads, and road connectivity).

3.3.2. Test on DeepGlobe Road Dataset

In this study, we proposed the HRU-Net, an enhanced model building upon the established foundations of the U-Net and HRNet. In order to further verify the effectiveness of the HRU-Net model, we put these three methodologies to the test using the DeepGlobe Road Extraction Dataset. The quantitative comparisons of the three methodologies, as conducted on the DeepGlobe Road Extraction Dataset, are displayed in Table 3. As demonstrated by the results, the HRU-Net model outperformed the others across all three metrics. It achieved an accuracy increase of 4.55% over HRNet and 2.63% over U-Net. In terms of recall, the HRU-Net model surpassed HRNet by 2.11% and U-Net by 1.75%. Regarding the IoU metric, the model exhibited improvements of 4.00% and 1.87% compared to HRNet and U-Net, respectively. These results underscore the superiority of our approach on the DeepGlobe Road Extraction Dataset.

Table 3. Quantitative evaluation of six methods conducted on the DeepGlobe Dataset. The closer the index value is to 1, the better the effect. IoU: Intersection over union.

Network	Precision (%)	Recall (%)	IoU (%)
HRNet	81.55	83.09	73.23
U-Net	83.43	84.45	75.36
HRU-Net (ours)	86.06	85.2	77.23

Figure 5 presents selected examples from the DeepGlobe Road Extraction Dataset, demonstrating the efficacy of different models. From top to bottom, the images show the original input, followed by the results from the HRNet, U-Net, and our proposed HRU-Net, respectively. In the first image, a blurred road scenario is depicted. Here, the HRNet and HRU-Net models showcase commendable extraction capabilities, with the U-Net model being the only one to exhibit over-extraction. The second image presents a case of severe shadow occlusion on the road, where both the HRNet and U-Net struggled to maintain road connectivity and the U-Net additionally failed to extract some road sections. In the third image, we have an example of a well-defined road where, intriguingly, only the HRU-Net model successfully avoided road disconnection when addressing the minor road on the left. Lastly, the fourth image provides an example of a complex urban road system.

In terms of road extraction integrity and continuity, the HRU-Net model demonstrated superior performance over both the HRNet and U-Net models.

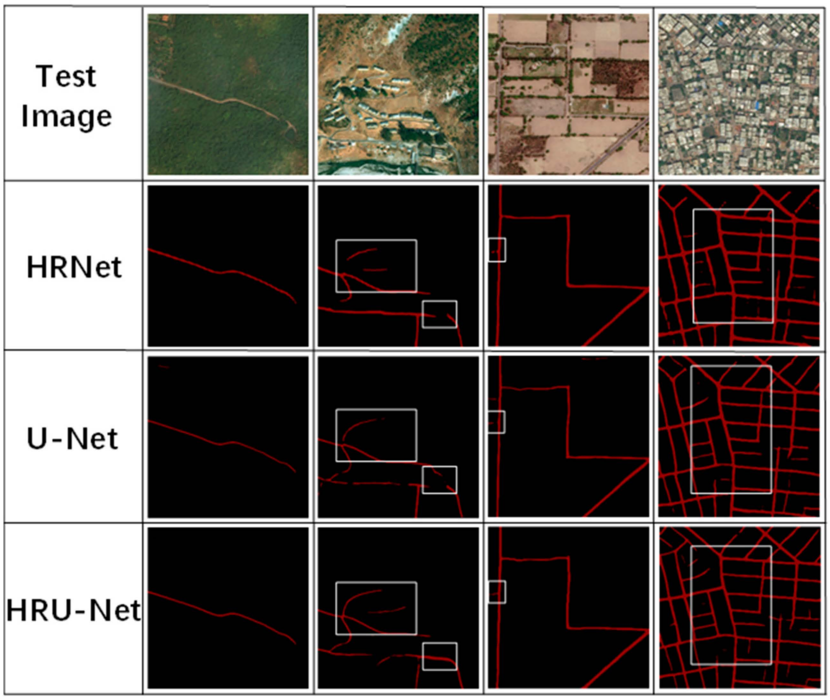

Figure 5. Comparison of road extraction effects of 3 schemes in 4 different test areas in DeepGlobe Road Dataset.

3.4. Ablation Experiment

3.4.1. Exploring the Impact of Modules on the Network

To evaluate the roles of the UMR and MRF modules in the HRU-Net model, a series of experiments was conducted, and a detailed analysis was carried out on their performance in the task of road extraction.

Initially, both the experimental and control groups were designed for comparative assessment. The control group consisted of the full HRU-Net model, incorporating both the UMR and MRF modules, while the experimental group was modified by excluding either the UMR or MRF module. All other experimental parameters and datasets were held constant to ensure a fair comparison.

Throughout the experiment, a diverse set of high-resolution remote sensing images was used, and identical training and validation procedures were applied to train and evaluate the various models. Evaluation metrics, including precision (P), recall (R), and intersection over union (IoU), were employed to gauge the performance of road extraction, as detailed in the Table 4.

Table 4. UMR and MRF module performance accuracy comparison. The closer the index value is to 1, the better the effect. IoU: Intersection over union.

Scheme	Precision (%)	Recall (%)	IoU (%)
Remove UMR	78.92	83.14	77.91
Remove MRF	79.19	83.47	78.06
HRU-Net (ours)	80.09	84.85	78.62

In evaluating the UMR module, performance differences between the experimental group (excluding the UMR module) and the control group (HRU-Net) were compared. A similar comparison was carried out for the MRF module. The training was conducted using the same dataset and experimental parameters, with the performance results on the Massachusetts Road Dataset shown in Figure 6.

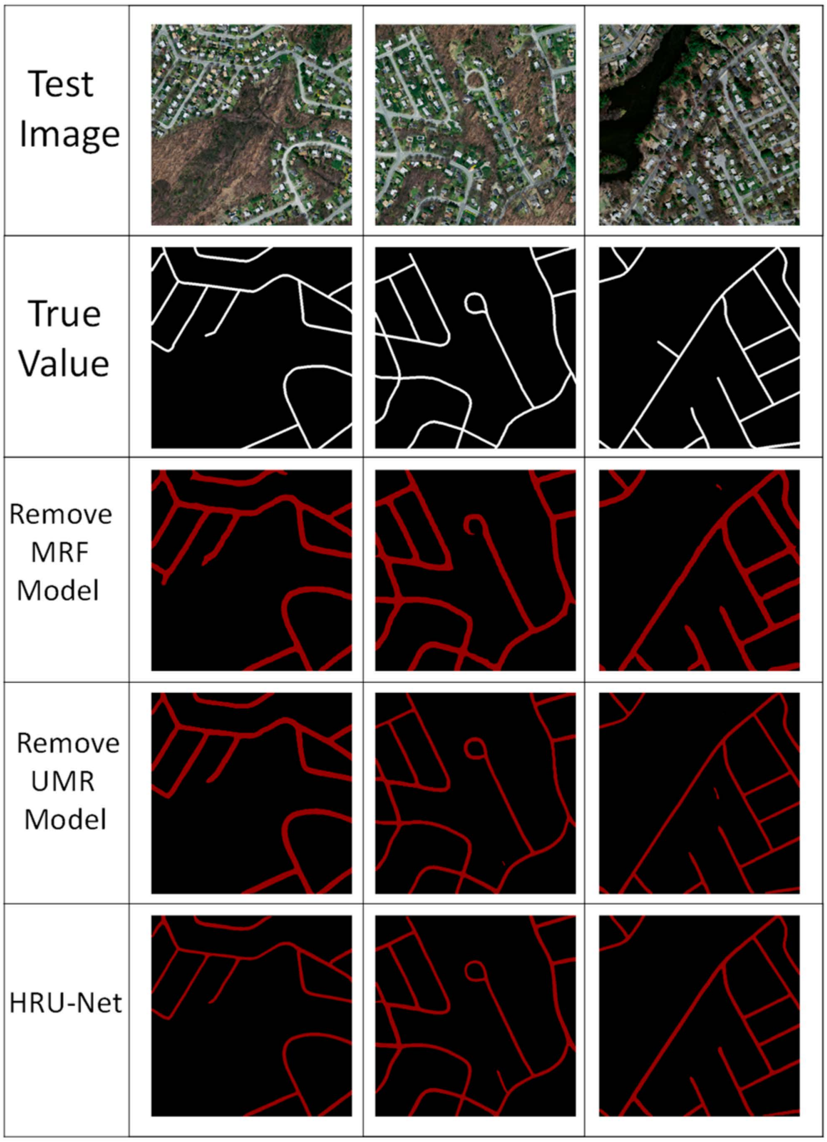

Figure 6. Performance experiment comparison of UMR and MRF modules.

Following the exclusion of the MRF module, the model's road edge extraction in some complex scenes was observed to be less precise, with blurred edges, breaks, or omissions. The MRF module played a key role in the HRU-Net model, effectively integrating multi-feature fusion and information transmission. By introducing the MRF module at the decoding stage, the model was better able to capture local road details and contextual relationships, thus enhancing its road extraction performance. Simultaneously, road discontinuities were observed when the UMR module was removed.

More specifically, in some complex scenes, the road continuity was disrupted, leading to interruptions or incomplete road segmentation. This might be due to a weakened ability.

In conclusion, the inclusion of the UMR and MRF modules in the HRU-Net model played a pivotal role in enhancing the performance of the road extraction task. Their design philosophies and functions were complementary, synergistically boosting the model's perception capabilities, feature expression ability, and accuracy. Consequently, they provided an effective solution for road extraction in high-resolution remote sensing images.

3.4.2. Exploring the Effect of the Number of Modules on the Network

In the first round of ablation experiments, the effectiveness of both the UMR module and the MRF module on the network was explored individually. To investigate the impact of module quantity on the network model, we selected the network model, which only retained the UMR1 module and MRF1 module, as the baseline network for the ablation study, as shown in dashed box 1 in Figure 7.

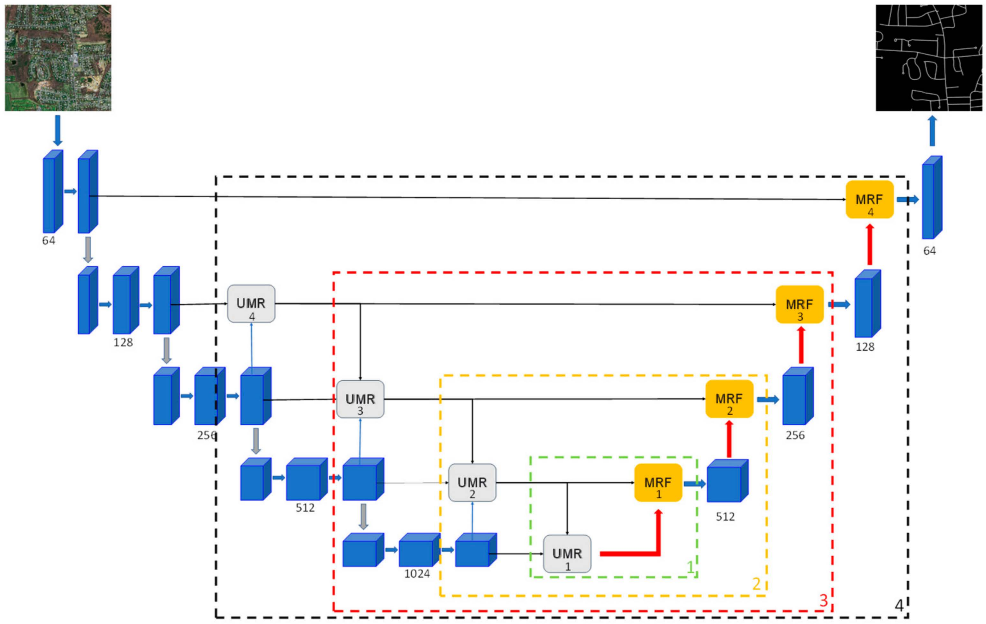

Figure 7. Schematic diagram of the scheme exploring the effect of the number of modules on the network. The dashed box only limits the number of UMR modules and MRF modules inside the box and does not limit other operations outside the box.

The experimental design was as follows:

Plan One: On the basis of the baseline model, the UMR2 module and the MRF2 module were added, corresponding to dashed box 2 in the figure.

Plan Two: Building on Plan One, the UMR3 module and MRF3 module were added, corresponding to dashed box 3.

Choosing dashed box 4 refers to selecting the full HRU-Net model.

The quantitative evaluation of the DeepGlobe Road Dataset is presented in Table 5.

Table 5. Accuracy representation of different schemes on the DeepGlobe Road Dataset. The closer the index value is to 1, the better the effect. IoU: Intersection over union.

Network	Precision (%)	Recall (%)	IoU (%)
Baseline Model	83.91	84.76	75.83
Plan 1	84.45	85.03	76.32
Plan 2	85.76	85.12	76.71
HRU-Net (ours)	86.06	85.2	77.23

The experimental results show an improvement in model performance, including the accuracy, recall, and IoU, as the number of UMR and MPF modules increased. This indicates that the augmentation of module quantity aided the model in learning more complex features and enhanced the network's representational capacity. Moreover, the UMR and MPF modules facilitated the capture of multi-scale features, enabling the model to understand data at different scales, which was crucial for the task at hand.

The test results on the DeepGlobe Road Extraction Dataset are shown in Figure 8.

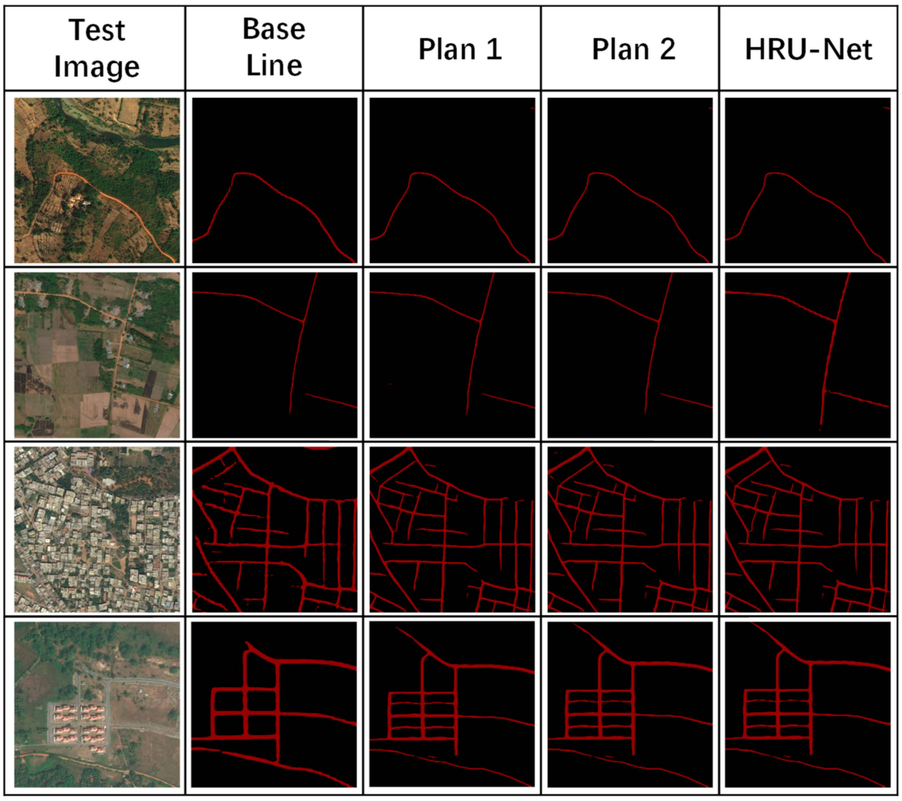

Figure 8. The extraction results of different experimental schemes on the DeepGlobe Road Extraction Dataset.

3.5. Computational Efficiency

In the field of deep learning, the number of parameters and floating point operations per second (FLOPs) serve as crucial measures for evaluating the complexity and computational demands of a model. A higher number of parameters might indicate a model's ability to learn more intricate patterns, but it could also lead to overfitting and extended training time. A comparison of the computational efficiency of different methods is shown in Table 6. Our model, the HRU-Net, demonstrated a balanced attribute with respect to both parameter quantity and FLOPs—neither being the simplest nor the most complex. Therefore, while significantly improving segmentation performance, the HRU-Net did not introduce an excessive number of parameters or noticeably increase the training time. This reflects that our model not only prioritizes performance but also focuses on computational efficiency and model generalizability.

Table 6. Comparison of computational efficiency of different methods.

Network	Parameters (M)	FLOPS (GLOPS)
U-Net	29.95	5.64
ResNet	25.56	5.40
Deeplabv3	5.87	6.61
ResUnet	38.52	8.64
HRNet	28.53	4.66
HRU-Net (ours)	**32.78**	**6.06**

4. Discussion

In this investigation, we assessed the efficacy of the HRU-Net model for road extraction from high-resolution remote sensing imagery. The experimental findings lend credence to our supposition that the HRU-Net model can proficiently delineate road information from these types of images. When appraised on the metrics of precision, recall, and intersection over union (IoU), the HRU-Net model manifested a superior performance in comparison to other cutting-edge methodologies employed for road extraction from remote sensing imagery, such as U-Net, ResNet, DeepLabV3, ResUnet, and HRNet [36–38]. These comparative findings portray the HRU-Net model as a promising candidate for implementing road extraction from high-resolution remote sensing images. These results bear significant practical implications. The competency to precisely delineate road information from high-resolution imagery can influence a gamut of domains including urban planning, traffic management, and disaster response [39]. The adeptness of the HRU-Net model could potentially transform these domains by providing higher precision and detail in road information.

Notwithstanding, our study was constrained by certain limitations. The precision of the HRU-Net model was compromised in certain intricate scenarios, as evidenced by blurred edges, breaks, or omissions when the MRF module was omitted. Likewise, the continuity of the road was disrupted, leading to interruptions or incomplete road segmentation when the UMR module was excluded. These constraints imply that while the HRU-Net model demonstrates promise, it requires further refinement to enhance its performance in intricate scenarios.

Interestingly, our observations indicate that the quantity of network modules had a consequential effect on the model's performance. The incorporation of the UMR and MRF modules into the HRU-Net model played an instrumental role in boosting the performance of the road extraction endeavor. Their design principles and functionalities acted in unison, thereby augmenting the model's perceptual capabilities, feature expression competency, and precision. Nevertheless, the escalation in the number of modules also culminated in an increase in computational complexity, which could potentially impact the model's efficiency.

For future investigations, we advocate for a deeper exploration into the UMR and MRF modules. Our findings denote that these modules contribute significantly to the enhancement of the road extraction task's performance. Subsequent research could focus on optimizing these modules to further ameliorate the model's performance in intricate scenarios, while also keeping a check on the equilibrium between performance augmentation and computational efficiency.

To summarize, our study provides substantial evidence that the HRU-Net model is a potent tool for road extraction from high-resolution remote sensing images. Despite the presence of certain limitations, the model demonstrated superior performance in comparison to other advanced methodologies in our experiments. The UMR and MRF modules, in particular, were critical in augmenting the model's performance. These findings not only contribute to the discipline of remote sensing image analysis but also lay the groundwork for future investigations.

5. Conclusions

In summary, this article proposes a road extraction model called the HRU-Net, which is based on the deep learning convolutional neural network model and adopts a "U" network architecture. The model combines the upsampling multi-scale feature fusion module (UMR module) and the multi-feature fusion module (MRF module) to enhance its ability to extract roads from high-resolution remote sensing images. Also, we implemented a new loss function to decrease the problem of class imbalance in our datasets and improved the result of road segmentation.

The multi-scale fusion module in the HRU-Net performs upsampling on inputs with different resolutions in the encoding stage and fuses feature maps of various scales for convolution in the decoding stage. By maintaining multi-scale information in parallel and preserving the parallelism of low-resolution and high-resolution feature maps, the network repeatedly exchanges information among different resolutions throughout its operation. This process enables the better fusion of multi-scale semantic information, the extraction of global multi-scale features, and the improvement of the network model's expression ability.

The proposed HRU-Net model demonstrated significant performance improvements when tested on the Massachusetts Road Dataset and the DeepGlobe Road Dataset. Its ability to effectively extract roads from high-resolution remote sensing images makes it a valuable tool for various road extraction tasks and applications.

Author Contributions: Conceptualization, Z.Y.; Methodology, C.R. and X.X.; Validation, W.Y.; Formal analysis, X.Z.; Investigation, X.L.; Writing—original draft, A.Y.; Visualization, Z.W.; Project administration, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Grant No. 42064003). The author is located at the Guilin University of Technology.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Wang, W.; Yang, N.; Zhang, Y.; Wang, F.; Cao, T.; Eklund, P. A review of road extraction from remote sensing images. *J. Traffic Transp. Eng.* **2016**, *3*, 271–282. [CrossRef]
2. Huang, X.; Zhang, L. Road centreline extraction from high-resolution imagery based on multiscale structural features and support vector machines. *Int. J. Remote Sens.* **2009**, *30*, 1977–1987. [CrossRef]
3. Bicego, M.; Dalfini, S.; Vernazza, G.; Murino, V. Automatic road extraction from aerial images by probabilistic contour tracking. In Proceedings of the 2003 International Conference on Image Processing (Cat. No.03CH37429), Barcelona, Spain, 14–17 September 2003; Volume 3, p. III-585. [CrossRef]
4. Baumgartner, A.; Steger, C.; Mayer, H.; Eckstein, W.; Ebner, H. Automatic road extraction based on multi-scale, grouping, and context. *Photogramm. Eng. Remote Sens.* **1999**, *65*, 777–786.
5. 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]
6. Abdollahi, A.; Pradhan, B.; Shukla, N.; Chakraborty, S.; Alamri, A. Deep Learning Approaches Applied to Remote Sensing Datasets for Road Extraction: A State-Of-The-Art Review. *Remote Sens.* **2020**, *12*, 1444. [CrossRef]
7. Choi, S.; Do, M. Development of the Road Pavement Deterioration Model Based on the Deep Learning Method. *Electronics* **2020**, *9*, 3. [CrossRef]
8. Mnih, V.; Hinton, G.E. Learning to Detect Roads in High-Resolution Aerial Images. In *Computer Vision—ECCV 2010*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 210–223. [CrossRef]
9. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440. Available online: https://openaccess.thecvf.com/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html (accessed on 6 March 2023).
10. Buslaev, A.; Seferbekov, S.; Iglovikov, V.; Shvets, A. Fully Convolutional Network for Automatic Road Extraction From Satellite Imagery. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 207–210. Available online: https://openaccess.thecvf.com/content_cvpr_2018_workshops/w4/html/Buslaev_Fully_Convolutional_Network_CVPR_2018_paper.htm (accessed on 6 March 2023).
11. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Proceedings of the 18th International Conference, Munich, Germany, 5–9 October 2015*; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241. [CrossRef]
12. Hou, Y.; Liu, Z.; Zhang, T.; Li, Y. C-UNet: Complement UNet for Remote Sensing Road Extraction. *Sensors* **2021**, *6*, 2153. [CrossRef]
13. Zhang, Z.; Liu, Q.; Wang, Y. Road extraction by deep residual u-net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
14. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning Where to Look for the Pancreas. *arXiv* **2018**, arXiv:1804.03999.
15. 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 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 182–186. Available online: https://openaccess.thecvf.com/content_cvpr_2018_workshops/w4/html/Zhou_D-LinkNet_LinkNet_With_CVPR_2018_paper.html (accessed on 6 March 2023).
16. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv* **2016**, arXiv:1511.07122.
17. Zhu, Q.; Li, Z.; Zhang, Y.; Guan, Q. Building extraction from high spatial resolution remote sensing images via multiscale-aware and segmentation-prior conditional random fields. *Remote Sens.* **2020**, *12*, 3983. [CrossRef]

18. Cheng, G.; Zhu, F.; Xiang, S.; Wang, Y.; Pan, C. Accurate urban road centerline extraction from VHR imagery via multiscale segmentation and tensor voting. *Neurocomputing* **2016**, *205*, 407–420. [CrossRef]
19. Du, S.; Du, S.; Liu, B.; Zhang, X. Context-enabled extraction of large-scale urban functional zones from very-high-resolution images: A multiscale segmentation approach. *Remote Sens.* **2019**, *11*, 1902. [CrossRef]
20. Salembier, P.; Serra, J.C. Morphological multiscale image segmentation. In Proceedings of the Visual Communications and Image Processing'92, Boston, MA, USA, 16 November 1992; pp. 620–631. [CrossRef]
21. Wu, Y.; Xia, Y.; Song, Y.; Zhang, Y.; Cai, W. Multiscale network followed network model for retinal vessel segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018, proceedings of the 21st International Conference, Granada, Spain, 16–20 September 2018, Proceedings, Part II 11*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 119–126.
22. 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.* **2018**, *40*, 834–848. [CrossRef]
23. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5693–5703. Available online: https://openaccess.thecvf.com/content_CVPR_2019/html/SunDeep_High-Resolution_Representation_Learning_for_Human_Pose_Estimation_CVPR_2019_paper.html (accessed on 6 March 2023).
24. Xiao, D.; Yin, L.; Fu, Y. Open-Pit Mine Road Extraction From High-Resolution Remote Sensing Images Using RATT-UNet. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
25. Abdollahi, A.; Pradhan, B.; Alamri, A. VNet: An End-to-End Fully Convolutional Neural Network for Road Extraction From High-Resolution Remote Sensing Data. *IEEE Access* **2020**, *8*, 179424–179436. [CrossRef]
26. Albawi, S.; Mohammed, T.A.; Al-Zawi, S. Understanding of a convolutional neural network. In Proceedings of the 2017 international conference on engineering and technology (ICET), Antalya, Turkey, 21–23 August 2017; pp. 1–6. [CrossRef]
27. Santos, C.D.; Tan, M.; Xiang, B.; Zhou, B. Attentive Pooling Networks. *arXiv* **2016**, arXiv:1602.03609.
28. Yarotsky, D. Error bounds for approximations with deep ReLU networks. *Neural Netw.* **2017**, *94*, 103–114. [CrossRef]
29. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer Normalization. *arXiv* **2016**, arXiv:1607.06450.
30. Joint Bilateral Upsampling | ACM Transactions on Graphics. Available online: https://dl.acm.org/doi/abs/10.1145/1276377.1276497 (accessed on 7 March 2023).
31. Bilinear Interpolation of Digital Images—ScienceDirect. Available online: https://www.sciencedirect.com/science/article/abs/pii/0304399181900619 (accessed on 8 July 2023).
32. Chen, D.; Zhong, Y.; Zheng, Z.; Ma, A.; Lu, X. Urban road mapping based on an end-to-end road vectorization mapping network framework. *ISPRS J. Photogramm. Remote Sens.* **2021**, *178*, 345–365. [CrossRef]
33. Jiang, X.; Li, Y.; Jiang, T.; Xie, J.; Wu, Y.; Cai, Q.; Jiang, J.; Xu, J.; Zhang, H. RoadFormer: Pyramidal deformable vision transformers for road network extraction with remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *113*, 102987. [CrossRef]
34. Zhong, Z.; Li, J.; Cui, W.; Jiang, H. Fully convolutional networks for building and road extraction: Preliminary results. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 1591–1594. [CrossRef]
35. Mena, J.B. State of the art on automatic road extraction for GIS update: A novel classification. *Pattern Recognit. Lett.* **2003**, *24*, 3037–3058. [CrossRef]
36. Tan, J.; Gao, M.; Yang, K.; Duan, T. Remote sensing road extraction by road segmentation network. *Appl. Sci.* **2021**, *11*, 5050. [CrossRef]
37. Minaee, S.; Boykov, Y.; Porikli, F.; Plaza, A.; Kehtarnavaz, N.; Terzopoulos, D. Image segmentation using deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3523–3542. [CrossRef]
38. Lian, R.; Wang, W.; Mustafa, N.; Huang, L. Road extraction methods in high-resolution remote sensing images: A comprehensive review. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 5489–5507. [CrossRef]
39. 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]

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.