remote sensing

MDPI

Article

DPSDA-Net: Dual-Path Convolutional Neural Network with Strip Dilated Attention Module for Road Extraction from High-Resolution Remote Sensing Images

Like Zhao [1], Linfeng Ye [1], Mi Zhang [2,*], Huawei Jiang [1], Zhen Yang [1] and Mingwang Yang [1]

[1] College of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, China; like_zhao@haut.edu.cn (L.Z.); 2021930850@stu.haut.edu.cn (L.Y.)
[2] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430072, China
* Correspondence: mizhang@whu.edu.cn

Abstract: Roads extracted from high-resolution remote sensing images are widely used in many fields, such as autonomous driving, road planning, disaster relief, etc. However, road extraction from high-resolution remote sensing images has certain deficiencies in connectivity and completeness due to obstruction by surrounding ground objects, the influence of similar targets, and the slender structure of roads themselves. To address this issue, we propose a novel dual-path convolutional neural network with a strip dilated attention module, named DPSDA-Net, which adopts a U-shaped encoder–decoder structure, combining the powerful advantages of attention mechanism, dilated convolution, and strip convolution. The encoder utilizes ResNet50 as its basic architecture. A strip position attention mechanism is added between each residual block to strengthen the coherent semantic information of a road. A long-distance shortcut connection operation is introduced to preserve the spatial information characteristics of the original image during the downsampling process. At the same time, a pyramid dilated module with a strip convolution and attention mechanism is constructed between the encoder and decoder to enhance the network feature extraction ability and multi-scale extraction of road feature information, expand the model's receptive field, and pay more attention to the global spatial semantic and connectivity information. To verify the reliability of the proposed model, road extraction was carried out on the Massachusetts dataset and the LRSNY dataset. The experimental results show that, compared with other typical road extraction methods, the proposed model achieved a higher F1 score and IOU. The DPSDA-Net model can comprehensively characterize the structural features of roads, extract roads more accurately, retain road details, and improve the connectivity and integrity of road extraction in remote sensing images.

Keywords: road extraction; occlusion of objects; strip convolution; position attention; dilated convolution

Citation: Zhao, L.; Ye, L.; Zhang, M.; Jiang, H.; Yang, Z.; Yang, M. DPSDA-Net: Dual-Path Convolutional Neural Network with Strip Dilated Attention Module for Road Extraction from High-Resolution Remote Sensing Images. *Remote Sens.* **2023**, *15*, 3741. https://doi.org/10.3390/rs15153741

Academic Editors: Pedram Ghamisi and Yonghao Xu

Received: 29 April 2023
Revised: 12 July 2023
Accepted: 25 July 2023
Published: 27 July 2023

1. Introduction

Urban informatization construction requires the rapid acquisition of a large number of basic geographic information data. Extraction of ground objects from remote sensing images, which has the advantages of wide spatial coverage, strong timeliness, low cost, and high resolution, is an important means of establishing and updating geospatial databases [1]. As the foundation of transportation, roads not only play an irreplaceable role in people's lives but also play an important role in the economy and military. Road extraction from remote sensing images has become a hot topic and a difficult issue in the literature. Grayscale characteristics, geometric structure, and texture information presented by different ground objects in high-resolution remote sensing images are different, which provides a basis for obtaining various ground object information. Conducting research on road extraction from high-resolution remote sensing images can greatly improve the

efficiency and accuracy of road data collection, help save labor and resources, and provide technical support for road traffic and real-time geospatial information data updates.

At present, road extraction methods for remote sensing images can be roughly divided into two categories: traditional methods and deep learning methods. In traditional road extraction technology, road extraction is usually based on pixel, region, structural feature, and texture information. Pixels are the smallest units in remote sensing images, and their size determines the resolution and amount of information contained in the image, which can accurately reflect the image features and carry the spatial and spectral characteristics of the image. For example, Cao et al. [2] used the integration of pixel-level features and multi-scale features to carry out the coarse extraction of a road network and reduced the influence of "salt noise" in the process of road extraction. A region-based approach divides the image into roads and non-roads by classifying similar parts of the image. For instance, Li et al. [3] extracted roads by acquiring road regions of interest from VHR (very high-resolution remote sensing) images using a hierarchical representation of binomial trees. Lei et al. [4] utilized a region growing method for selected road seed points, and combined edge information and morphological methods to extract road networks, but its performance may decrease for road surfaces with significant interference and inconsistent grayscale. Road extraction results can be achieved according to the structural information of roads, including linear, near-circular, and intersecting structures. For example, Netanyahu et al. [5] used local operators to detect road area pixels and extracted roads by finding sets of approximately linear or circular road pixels. Pan et al. [6], on the basis of mathematical morphology, utilized different structural elements to extract road edges according to edge shapes in the image. However, the extraction effect for circular roads needs to be improved since circular structures on a computer are represented as approximations. In addition, roads in remote sensing images usually have unique texture features, such as brightness changes, texture orientation, and grayscale distribution, which can be used to extract roads from remote sensing images. For example, Zang et al. [7] adopted joint-enhanced filtering to suppress high-contrast texture interference, followed by adaptive smoothing to suppress noise and smooth heavy texture, while retaining potential road surface structures. Zhang et al. [8] used an angular texture feature model to predict road centerlines and achieve semi-automatic extraction of roads, but there was still noise interference from similar ground objects, and the extraction results still required manual post processing. Therefore, traditional road extraction methods are more suitable for remote sensing images with clear, semantically distinct information and relatively simple road structures. With further improvement in the resolution of remote sensing images and an increase in the diversity of road backgrounds, traditional methods face challenges and have certain limitations in universality and road feature representation.

With the development of deep learning, the convolutional neural network (CNN) has received widespread attention from researchers, and it has been proven to be more efficient than traditional methods in semantic segmentation. For example, the fully convolutional neural network, based on the CNN method proposed by Buslaev et al. [9], utilizes deconvolution for upsampling operations and applies end-to-end convolutional neural networks to perform semantic segmentation. The convolutional neural network U-Net [10], which effectively solves the image segmentation problem, has a network structure based on encoder–decoder architecture with a convolutional deepening of the decoder part to recover the image features. Zhang et al. [11] proposed an image segmentation model, deep residual U-Net, in order to overcome the degradation problem caused by increasing depth. Residual convolution is used to replace general convolution, which can better extract and restore image features, ensure more stable and reliable performance of the model, and achieve good results in image segmentation tasks. In LinkNet [12], proposed by Chaurasia et al., the encoder part adopts continuous convolutional layers for feature extraction and shortcut connections to add the shallow and deep feature maps, while in the decoding stage, deconvolutional layers and skip connections are utilized to recover image resolution and feature information, and the network model improves the speed while ensuring model accuracy. Zhou et al. [13] designed D-LinkNet with the addition of dilated convolution,

which has a larger receptive field and can retain more feature information to improve the connectivity problem when compared with LinkNet. Wang et al. [14] proposed NL-LinkNet using, for the first time, neural non-local operations for road extraction; in addition, the model utilized global features to improve the extraction of obscured roads. Xu et al. [15] designed L-UNet, which uses the mobile flipped bottleneck convolution module to extract features and greatly improves the segmentation effect while reducing the number of parameters by expanding the convolution and compressing the excitation module. However, for a part of the road covered by thick clouds, the extraction results were still insufficient in terms of continuity. Xie et al. [16] proposed a global perceptual network, HsgNet, based on bilinear pooling sorting spatial information, using LinkNet as the basic architecture, and embedding a middle block between the encoder and decoder to learn and save the semantic dependencies of images, which greatly reduces the running time, but it is still lacking in the connectivity of the road topology. Chen et al. [17] proposed a biased U-shaped network that adds decoding branches during upsampling and utilizes convolutional filters of different sizes in the branches to obtain multiple semantic data. Ren et al. [18] fused multi-scale semantic features of high-resolution images by constructing a capsule U-Net architecture and utilized contextual attributes to generate class-specific feature encoding. Ding L et al. [19] designed DiResNet, a direction-aware residual network, which uses deconvolution and structural supervision to strengthen the learning of road topology, a direction-supervised module to enhance linear features, and finally, combines topology and linear features to extract roads. Filin et al. [20] adopted a combination of neural networks and post-processing algorithms to obtain both road masks and road vectors, and the acquired road vectors can be used to process higher-level tasks. Wang et al. [21] proposed an inner convolutional network integrated encoder–decoder that sliced feature maps by columns or rows, transferred spatial information in the slices, and designed a conditional random field with road direction as the energy item for a post-processing method. The network effectively improved the connectivity of the extraction results of roads affected by occlusion. Sun et al. [22] proposed a stacked, U-shaped network, which connected two convolutional neural networks, to extract roads. The output of the first network and each part of the decoder is used as the input of the second network to improve the accuracy of road extraction, and the post-processing method of road map vectorization and hierarchical threshold shortest path search effectively improves the recall of road extraction.

At the same time, the attention mechanism designed from the study of people's vision allows the model to focus on the most concerned regions to improve the accuracy and robustness of the model. Hu et al. [23] proposed a network structure for semantic segmentation, SENet, with an SE (squeeze-and-excitation) module designed to adaptively adjust the importance of each channel and help the model automatically focus on the important features and suppress the unimportant ones. Fu et al. [24] proposed a dual-attention network, DANet, which uses a position attention module to aggregate and weigh the features at each position, and a channel attention module to emphasize the interdependent channel mapping. Further, the outputs of the two modules are combined to obtain more accurate segmentation results by improved feature representation. Woo S et al. [25] proposed a lightweight attention module (CBAM) to infer attention from channel and spatial dimensions, which can be effectively embedded in different network models and greatly improve the model's extraction capability. Li et al. [26] added a global attention module and a core attention module to DenseUNet, proposing a new network model, CADUNet, which reduced the impact of tree canopy occlusion on roads to a certain extent; however, the problem persisted that road edges obtained were not smooth. Wang et al. [27] constructed DA-RoadNet, a road extraction network with semantic inference capability, which enhanced the semantic features of images by integrating dual attention modules and achieved good results in road extraction, but there is a great deal of room for improvement in terms of integrity. Wang et al. [28] adopted a double-decoding network structure to obtain more detailed features and introduced a dilated convolutional attention module (DCAM) between the encoder and decoder, combining cascaded dilated convolution and

a convolutional attention module to extract the multi-scale features and attention-aware features to improve the network's ability to extract roads in complex environments.

There are still many challenges in road extraction from remote sensing images due to similar objects, complex scenes, lighting changes, and shooting conditions. Compared with traditional methods, deep convolutional neural networks achieve better results in road extraction and can extract detailed semantic information from the combination of low-level and high-level feature maps. The fact that road extraction results are broken due to object occlusion and its own complex structure and the great difference in the width and length of different roads remains a problem. Some of the above methods utilized, such as expanding the receptive field and position attention, resolve this problem. However, less consideration is granted to the global position, semantic information, and the narrow and long structural characteristics of roads, which leads to a lack of connectivity and integrity of extracted roads. How to extract semantic information of global position and capture pixel-remote dependence to improve the connectivity and integrity of roads is the focus of this paper. Combining the semantic information of global positioning and the narrow and long structure of roads, a novel road extraction model, DPSDA-Net, was designed by adopting strip convolution as the entry point and incorporating a position attention module and pyramidal dilated convolution.

The contributions of this paper are summarized as follows:

(1) A dual-path strip dilated convolutional network, DPSDA-Net, was proposed, which adopts a dual-path downsampling method to extract increased semantic feature information at local and global positions by introducing strip positional attention and pyramid dilated convolutions. The network can learn long-range semantic information relationships between distance pixels, capture long-distance dependencies, and perform correlation modeling for different spatial position information features, enhancing the segmentation ability of the model and improving the connectivity and integrity of the road.

(2) According to the characteristics of slenderness, long span, and connectivity of the road, a strip position attention module (STPA) was constructed to enhance the network's learning features. A combination of strip convolution and traditional convolution is used to extract the vertical, horizontal, left, and right diagonals of the road. Introducing more contextual information can effectively improve the learning efficiency of the global location of semantic information and enhance the ability to mine road features.

(3) A long-distance shortcut connection (LDSC) operation was designed to downsample the input image with the same resolution as the feature map of each layer. It performs a connection operation so that low-level features can be transferred to a high-level, while relatively preserving the original image spatial information.

(4) A pyramid dilated strip attention convolution module (PDSA), in which strip convolutions are included, was designed. The strip convolution branches establish rich structural feature relationships between road elements in adjacent and non-adjacent elements, and the pyramidal dilated convolution can effectively increase the receptive field of the convolution layer to obtain a larger range of contextual information while performing multi-scale feature extraction to further enhance the feature representation capability of the model and improve the continuity of road extraction.

2. Methods

In this section, the structure of the DPSDA-Net model is first introduced. Then, the constructed strip positional convolutional attention module STPA, long-distance shortcut connection operation LDSC, and pyramid dilated strip convolution attention module PDSA are described in detail. Finally, the loss function of the model is briefly presented.

2.1. Workflow

The workflow of the proposed method is shown in Figure 1, which mainly includes the following steps: (1) crop the images from the training, validation, and test sets in the

original dataset. (2) The loss backpropagation calculated using the predicted image and the ground truth guide the network training. (3) Extract the road and evaluate the model performance.

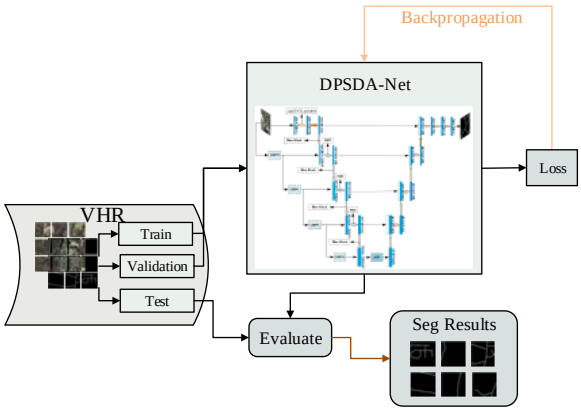

Figure 1. Workflow of the proposed road extraction method.

2.2. DPSDA-Net

DPSDA-Net adopts a dual-path encoding method to improve spatial perception and feature representation by introducing strip convolution, attention mechanism, and pyramid-dilated convolution. At the same time, features of the road structure are extracted in multi-scale, global semantic information is captured, and remote dependency modeling is carried out to reduce the influence of road occlusion and similar ground objects and improve the accuracy of the model's extraction. The model structure is shown in Figure 2. The backbone is based on U-Net and consists of a contraction path, bridge connection, and extension path. It is worthwhile to note that there are two paths for model contraction. The first path is the feature extraction path, which utilizes residual blocks for obtaining rich contextual semantic features of the image. The second path is a long-distance shortcut connection operation to retain, to a greater extent, the spatial information of the input image. The bridge connection part further excavates the deep semantic features and feeds the output feature map into the extended path. The expansion path upsamples the output of the bridge connection part and recovers the semantic features and image resolution level by level. Finally, a skip connection path is used to connect the different level feature maps of the contraction path and the expansion path to recover the spatially informative features lost in the contraction process.

Regarding the dual paths of contraction, one is the feature extraction part, with a convolution of step size 2 and a 7×7 convolution kernel as the initial block. Then, four consecutive residual blocks are obtained, and the last three residuals' convolution, with a step size of 2, is used in the block to achieve the purpose of shrinking the image size. Finally, the output feature maps from each residual block are input to the strip position attention module (STPA) to extract and collect long-distance semantic information features of the roads. The second path is the long-distance shortcut connection operation (LDSC). This operation directly affects the original image and connects it with the feature map at the corresponding level, preserving the original spatial feature information.

In the bridge connection part, to address the structural characteristics and connectivity of narrow and long roads, the method of expanding the receptive field and strip convolution is adopted and the pyramid dilated strip attention module (PDSA), with a strip convolution branch and attention mechanism, is introduced. In this module, the strip convolution branch can better capture the semantic features of narrowness and length of the road, while the attention mechanism can better focus on the road area and improve the accuracy of road extraction.

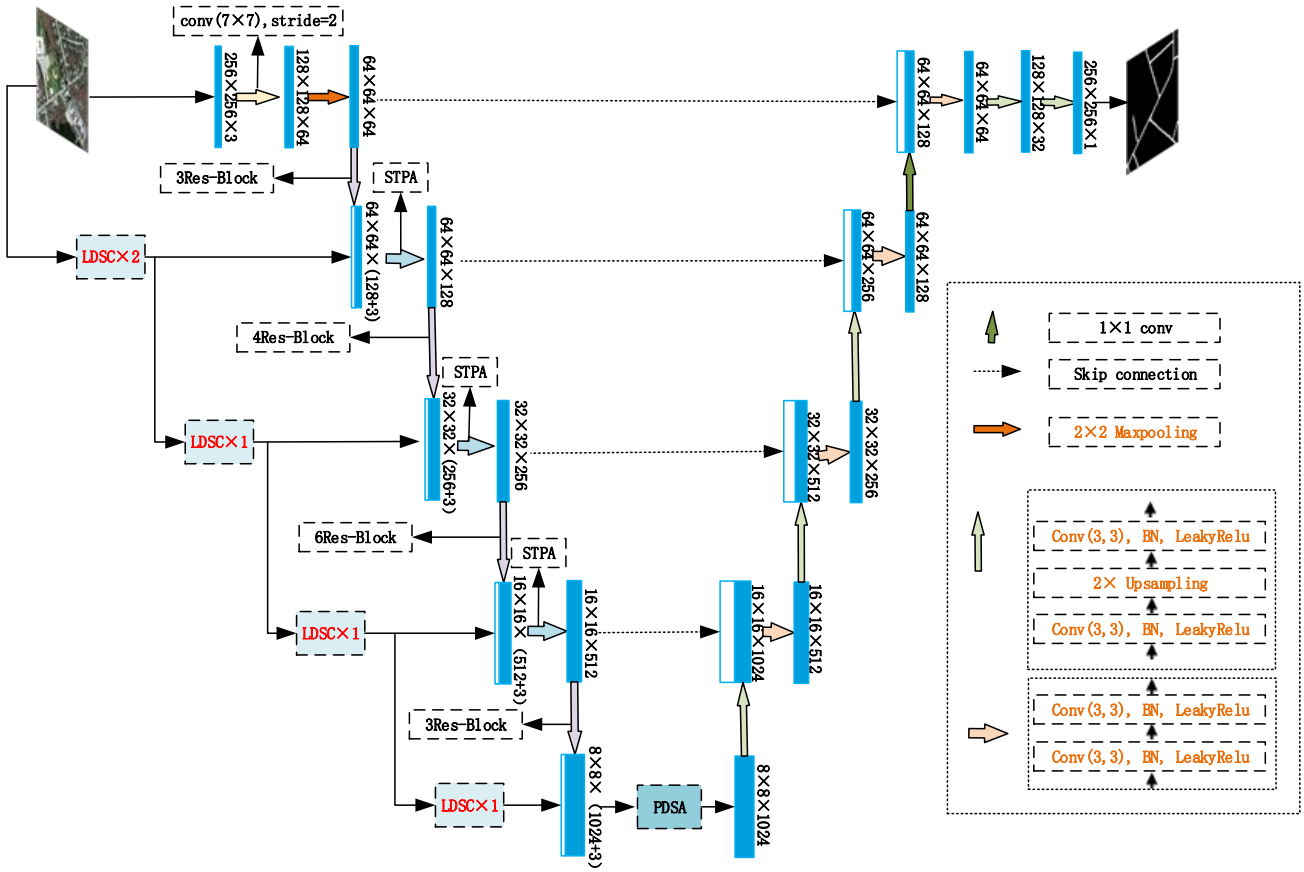

Figure 2. DPSDA-Net.

During the shrinking process, the spatial information of the image will be lost. In the expansion process, the method of connecting upsampling result and output of the STPA module is adopted to restore the spatial information. At the same time, bilinear upsampling with a factor of 2 is used to ensure the accuracy of feature map upsampling while reducing computational complexity. Finally, the sigmoid activation function, utilized to classify the output, and segmentation result of road area, is obtained.

2.3. Explanation of STPA in the Encoder

Combined with the structural characteristics of the road itself, a module named strip position attention (STPA) was designed, which was added to the encoding part of the model so that the network can extract spatial structure information that conforms to the characteristics of the road during the encoding stage.

The conventional position attention module captures the spatial dependence between any two positions in the feature map. For a specific feature, features at all positions are weighed and updated, and the weight is a feature that is common between the two corresponding positions. Any two positions with similar features can contribute to each other, modeling the contextual relationship on local features. However, roads usually appear in high-resolution remote sensing images in the form of strips with a large span, that are narrow and with continuous distribution. The standard square convolution is limited by its convolution kernel and cannot capture the linear features of roads well, because it is difficult to capture information that is unrelated to roads in the extraction process. Figure 3 illustrates the difference between the standard convolution and strip convolution. H, V, LD, and RD represent horizontal, vertical, left diagonal, and right diagonal strip convolutions, respectively. Strip convolution conforms better to the shape of the road, which utilizes a long kernel shape along the spatial direction to capture the long-range correlation of the road area, and extracts the strip semantic relationship that

conforms to the road characteristics. Consequently, the strip convolution is selected to further improve the position attention module. Our STPA is different from the conventional attention mechanism. It enables the strip convolution kernel to simultaneously focus on multiple parts of the road at different locations and thereby can extract rich contextual information about the road. Consequently, the strip position attention mechanism can more accurately capture the relevant features of the road and improve the performance of road recognition. In addition, the shape of the strip convolution kernel can also reduce interference from pixels in non-road areas. Since the strip convolution kernel is adaptable to the shape of the road, it tends to select pixels in the road area when calculating the attention weight. By construct, the standard convolution operator only captures the local regions and brings the irrelevant features.

Figure 3. Comparison of strip convolution and standard convolution. Red and yellow colors represent standard convolution and strip convolution, respectively.

As shown in Figure 4, the local feature map A is first input into a 1×1 convolution layer to generate new feature maps Q, K, and V; thus, $\{Q, K, V\} \in R^{C \times H \times W}$, where C, H, and W denote the number of channels, height, and width, respectively. The obtained feature maps are passed through the strip convolution (ST) module to obtain new features M, N, and O, and then reshape the feature map M after the feature map Q passes through the strip convolution module into $R^{N \times C}$, $N = H \times W$, and reshape the feature maps N and O after the feature maps K and V pass through the strip convolution module into $R^{C \times N}$, $N = H \times W$. Then, matrix multiplication is performed between the transposes of M and N and the softmax layer is applied to calculate the spatial attention map $S \in R^{N \times N}$, which can learn the close correlation between each position and other positions and represent the relationship between them with more similar features. Then, matrix multiplication is used between O and S and the result is reshaped as $R^{C \times H \times W}$. Finally, an element-by-element summation operation is performed with the feature map A to obtain the final output result $E \in R^{C \times H \times W}$. The feature map E is achieved by adding all position features, which can effectively aggregate the global context and enable the model to better perceive the semantic information of the entire image. At the same time, it can improve the robustness of the model and reduce the mis-segmentation caused by local noise or interference.

Using W_0 to represent strip convolution operation f^{ST} for the four directions of horizontal, vertical, left diagonal, and right diagonal, W_1 for the reshape and transpose operation, W_2 for the reshape operation, σ for softmax; and using the strip position attention, STPA module can be expressed as follows in Equation (1):

$$\begin{aligned} E &= reshape\{\sigma\{reshape\ \&\ transpose(f^{ST}(Q)) \times reshape(f^{ST}(K))\} \times reshape(f^{ST}(V))\} + X \\ &= W_2\{\sigma\{W_1(W_0(Q)) \times W_2(W_0(K))\} \times W_2(W_0(V))\} + X \end{aligned} \quad (1)$$

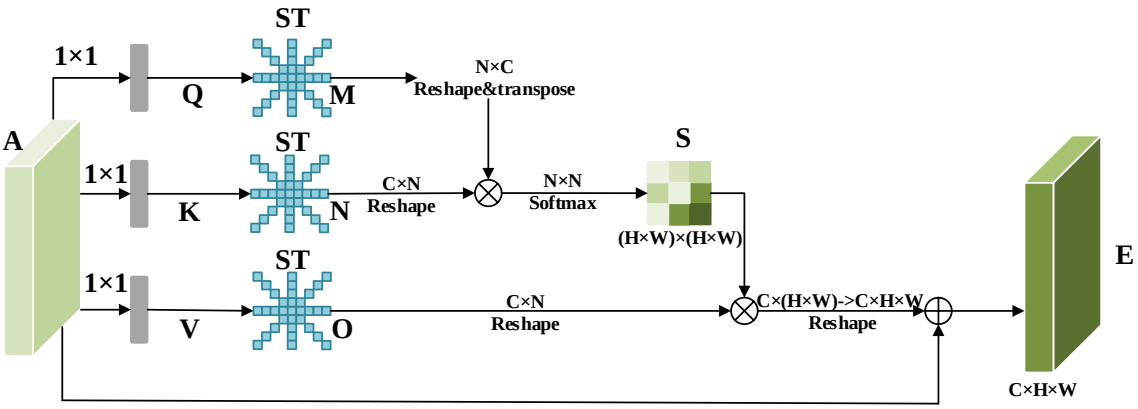

Figure 4. STPA Module.

The strip convolution ST module is shown in Figure 5. The convolution operation utilizes four strip convolutions of horizontal, vertical, left diagonal, and right diagonal to capture the long-distance contextual semantic information from four different directions. The input feature maps are denoted as $X \in R^{C \times H \times W}$, and X is the input tensor of strip convolution. After the input tensor undergoes a 1×1 convolution, the output results are subjected to strip convolution operations with the same size in different directions, and the output result of feature maps are stitched together, followed by interpolate BN (batch normalization), and ReLU. In addition, the number of output channels is changed through 1×1 convolution, BN, and ReLU, and then the next step of shape reshaping in the strip position attention is performed.

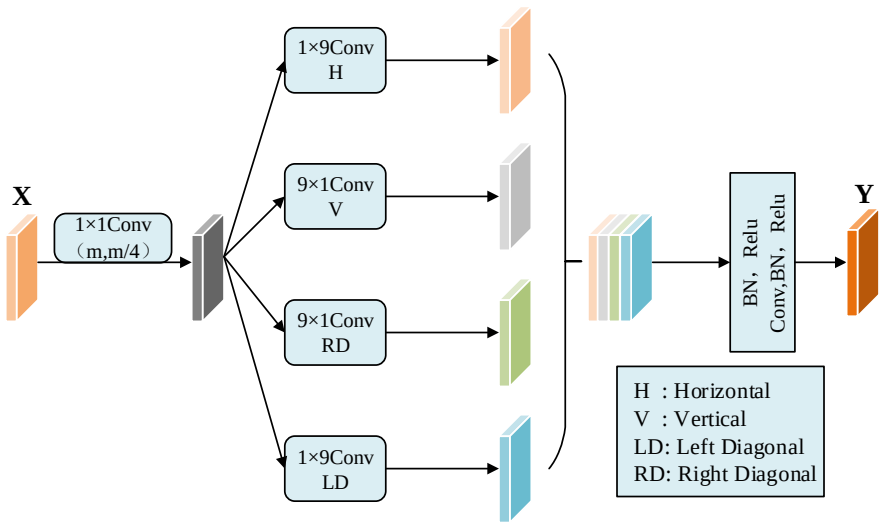

Figure 5. Strip convolution module.

Each input in the positional attention model is passed through the strip convolution module, and each element in the input feature map establishes a relationship with its surrounding elements. Clearly, such an operation can strengthen the feature connection between road pixels in different directions and enhance the connectivity of road extraction.

2.4. Long Distance Shortcut Connection Operation

In order to effectively retain the spatial information lost during the downsampling process of the input image, a long-distance shortcut connection operation is designed and added between the input image and the downsampling unit. The operation can be divided into the following three parts: (1) average pooling with a convolution kernel size of 3 and

a step size of 2, which downsamples the image so that the image size is the same as the output feature map of the downsampling unit; (2) standard 3 × 3 convolution, which processes the average pooled feature map to learn the spatial information of the image; and (3) pointwise convolution, which performs feature extraction on a single point based on the spatial representation. Compared with conventional downsampling methods, the long-distance shortcut operation uses average pooling, a combination of standard convolution, and pointwise convolution, to extract input image features, which can effectively avoid the risk of losing spatial information relationships due to multiple downsamplings and convolution. The module of LDSC is shown in Figure 6.

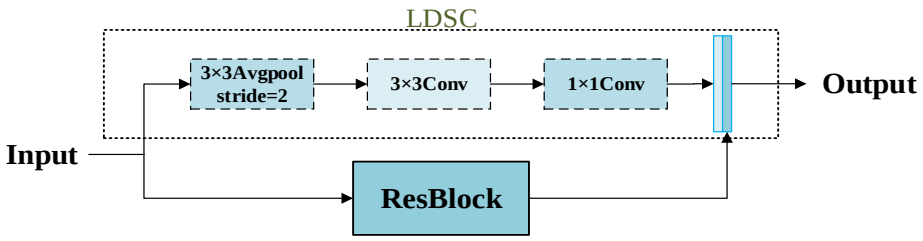

Figure 6. LDSC module.

2.5. PDSA Module

In convolutional neural networks, deep-level feature maps usually contain high-level semantic information, and the PDSA module is designed as the structural characteristics of the narrow and long road as the bridge connection part of the network to further excavation and utilization of this information, which can better extract road feature information and improve the connectivity of road extraction. Usually, the dilated pyramid convolution module is mostly used for the bridge connection part, which can effectively increase the receptive field, but the dilated convolution brings the problem of discontinuous pixel areas in the receptive field. In response to this problem, it is proposed to introduce a strip convolution branch into the pyramid dilated convolution module, extracting features synchronously with the dilated convolution branch to increase pixel continuity. The convolutional attention module (CBAM) [25] is embedded into the two branches of strip convolution and dilated convolution, divides the feature map into different feature channels through the channel attention mechanism, and establishes the importance of each channel degree model. The spatial attention mechanism is adopted to increase the spatial mapping of the feature map, which can help the network learn feature weights independently and obtain attention perception features of different receptive fields by training the channels and spatial dimensions of the feature map.

The PDSA module adopts the structure of parallel branches, which are three branches of global average pooling, strip convolution, and dilated convolution. In the first branch, the global average pooling operation is used to capture the global average feature. In the second branch, the strip convolution operation is adopted to obtain the global structural information. In the third branch, the dilated convolution can increase the receptive field and improve the integrity of the road. Theoretically, the size of the expansion rate of dilated convolution has a certain influence on the extraction results of the road. It is very beneficial to set a large expansion rate for long and narrow roads. On the contrary, for shorter roads, it is necessary to adopt a smaller expansion rate. In order to achieve a better extraction effect, different sizes of expansion rates are used to increase the receptive field and learn the roads' relevant features of different sizes.

The module structure is shown in Figure 7: $\mathbf{X}$ was used as the input feature map and passed into the three parallel branches, respectively. The first two branches perform global average pooling and strip convolution extraction operations, respectively. In the third branch, dilated point convolution is used to learn the linear representation, and the expansion rate and padding are, respectively, 2, 4, and 8; the convolution with a convolution kernel size of 3 expands the receptive field to obtain road features at different scales. The

output results of the dilated convolution and the strip convolution branch, respectively, are input into the convolution attention module to strengthen the road detail features. Additionally, the output results of all operations are connected with the input feature map to retain the original input image feature information. Finally, the pointwise convolution operation is adopted to extract single-point features and correct the number of output channels for upsampling of the model. Dilated convolution can increase the receptive field, which can alleviate the lack of completeness of road extraction, and then, strip convolution operation and convolution attention module are adopted to strengthen road detail features, which can better extract complete roads.

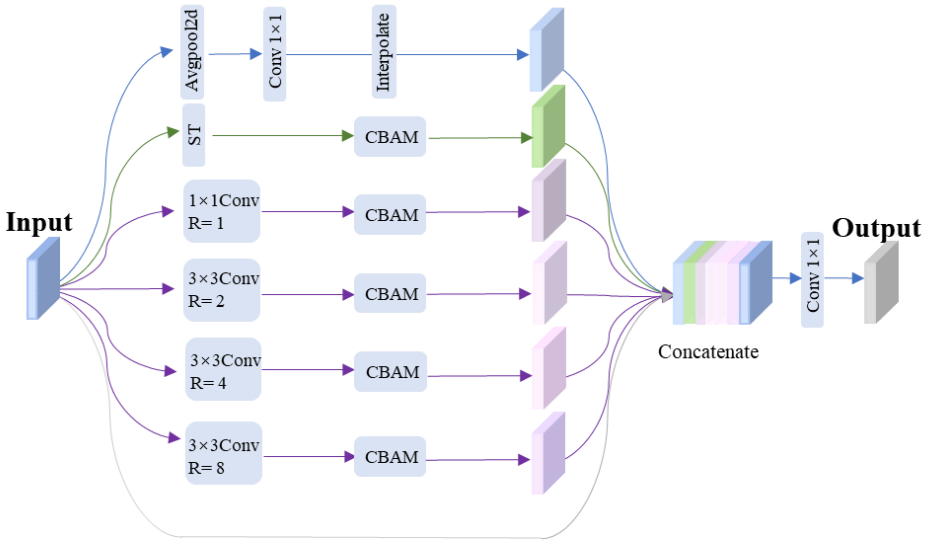

Figure 7. PDSA module.

X is used to represent the input feature map of the module. *Avgpool* indicates global average pooling, $f^{n\times n}$ indicates convolution operation with different convolution kernels, F^{in} indicates interpolate operation, f^{ST} indicates strip convolution module, $f_m^{n\times n}$ indicates dilated convolution with dilated rate m and convolution kernel n, F^{cb} indicates convolution attention module, and *Concatence* represents the connection operation. The PDSA module can be expressed as follows, in Equations (2)–(5):

$$Z_1 = F^{in}(f^{1\times 1}(AvgPool(X))) \tag{2}$$

$$Z_2 = F^{cb}(f^{\mathrm{ST}}(X)) \tag{3}$$

$$Z_3 = F^{cb}(f_1^{1\times 1}(X); f_2^{3\times 3}(X); f_4^{3\times 3}(X); f_8^{3\times 3}(X)) \tag{4}$$

$$Out = Concatenate\,(X, Z_1, Z_2, Z_3) \tag{5}$$

2.6. Compound Loss Function

Road extraction can be regarded as the problem of distinguishing roads from the background, namely a binary classification problem. Road pixels often only occupy one-tenth or even less of the entire image area, and the category imbalance is very serious. The compound loss function of dice loss and focal loss is introduced to guide the model training, which can make the loss and gradient value of pixel values in a certain region correlate not only with its label but also with other pixel points, which can better excavate road region. Overall, the weight of hard-to-classify samples was increased in the loss function and the learning of such samples was strengthened, which can help improve the accuracy

of hard-to-classify samples. The proportions of the two functions are λ_1 and λ_2, according to the proportion of road and background in the whole region, respectively. The compound loss function is described in Equation (6):

$$Loss = \lambda_1 \times L_{Dloss} + \lambda_2 \times L_{FL} \tag{6}$$

L_{Dloss} and L_{FL} represent the dice loss and focal loss, and the expressions of the function are calculated in Equations (7) and (8):

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

$$L_{FL}(pt) = -(1 - pt)^{\gamma} \log(pt) \tag{8}$$

X represents the pixel label of the ground truth, and Y represents the segmented image predicted by the model. pt reflects the similarity between the predicted value and the real value, and $\gamma > 0$ is an adjustable factor.

3. Experimental Results and Analysis

In this section, the model performance of DPSDA-Net is evaluated, and the experimental setup and experimental results are described in detail below. Firstly, our model is compared with other mainstream road extraction models, including U-Net [10], SegNet [29], D-LinkNet [13], NL-LinkNet [14], R-BasicUNet [17], and DDU-Net [28]. Then, the ablation experiments of strip positional attention modules, long-range connectivity operations, and pyramidal dilated strip attention modules are performed.

3.1. Dataset

To verify the effectiveness of the designed model, the Massachusetts Road Dataset [30] and LRSNY (Large Road Segmentation Dataset from Optical Remote Sensing Images of New York) Dataset [17] were used for experimental analysis. The Massachusetts Road Dataset is constructed by Mnih and Hinton and contains 1171 aerial images of Massachusetts. Each image has a size of 1500 × 1500 pixels and covers 2.25 square kilometers. The data is randomly divided into a training set of 1108 images, a validation set of 14 images, and a test set of 49 images. The data covers a wide range of areas, with more than 2600 square kilometers of urban, suburban, and rural areas in the United States. The ground truth of the dataset images contains binary images of road segmentation: road and non-road (background).

The LRSNY dataset is a large-scale satellite image dataset released by the team of Professor Chen Ziyi from the School of Computer Science and Technology of Huaqiao University. The dataset contains manually marked, pixel-level high-resolution satellite images, covering the center of New York, with a resolution of 0.5 m. Depending on road coverage, ground truth includes road surfaces that are obscured by trees, buildings, and cars. White (255, 255, 255) and black (0, 0, 0) are used in the labeled image to denote the road area and background area. The LRSNY dataset has a total of 1368 labeled images with a size of 1000 × 1000, including a training set of 716 images, a validation set of 220 images, and a test set of 432 images.

A few images are randomly selected in two datasets in order to show the datasets more clearly, as in Figure 8.

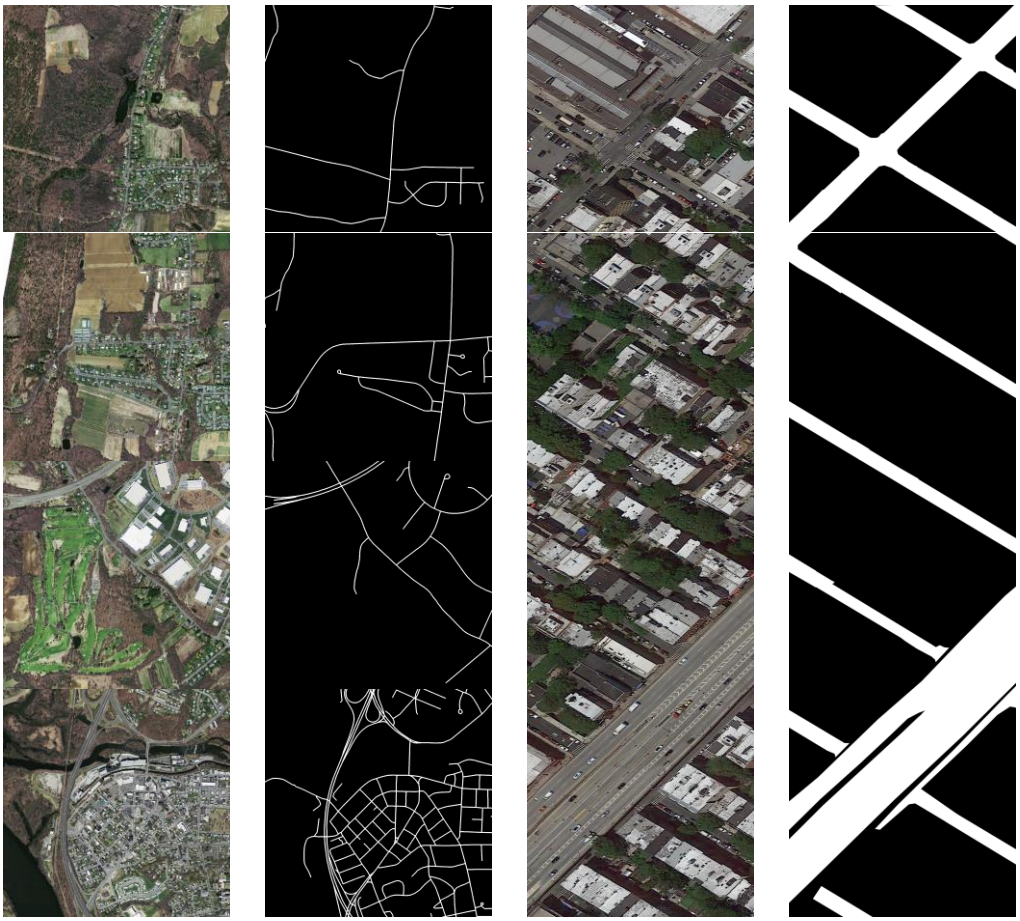

Figure 8. Massachusetts dataset: (**a**) satellite imagery; (**b**) ground truth. LRSNY dataset: (**c**) satellite imagery; (**d**) ground truth.

3.2. Evaluation Indicators

To quantitatively evaluate the reliability of the proposed model, four evaluation metrics were used to assess the performance of the model: precision, recall, F1 score, and IOU, expressed in Equations (9)–(12):

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$F1 - score = \frac{2 \times precision \times recall}{precision + recall} \tag{11}$$

$$IOU = \frac{TP}{TP + FP + FN} \tag{12}$$

where TP, FP, and FN represent true positives, false positives, and false negatives based on predicted images and ground truth, respectively.

3.3. Comparative Experiment with Other Networks

In the experiments, the original image dataset was cropped separately; 10,539 images of 512 × 512 size were finally obtained from the Massachusetts dataset, and a total of 5472 images of 512 × 512 size were acquired from the LRSNY dataset. The data augmenta-

tion methods of random hue–saturation value, random shift scaling, random horizontal flip, random vertical flip, and random rotation by 90° were adopted. Our model and the other six models are constructed using the Pytorch framework. Our model uses a learning rate of 5×10^{-4}, selects 20 epochs as one cycle, adjusts the learning rate using cosine annealing, and due to the dissimilarity of the two datasets, different size epochs are used. The training epochs for Massachusetts and LRSNY datasets are 180 and 160 epochs, respectively. The optimizer uses the Adam optimizer, which can update the variables based on the oscillation of the historical gradient and the historical gradient after filtering the oscillation. The hyperparameters used in the comparison models in this paper are consistent with those described in the original article, and all models were trained and tested on an NVIDIA GeForce RTX 3080 10 GB.

3.3.1. Comparative Experiment on Massachusetts Road Dataset

Using the Massachusetts road extraction dataset, our proposed DPSDA-Net was compared with other mainstream segmentation methods, namely U-Net [10], SegNet [29], D-LinkNet [13], NL-LinkNet [14], R-BasicUNet [17], and DDU-Net [28]. Table 1 shows the results of the comparison on the Massachusetts dataset, where the best values are shown in bold. As can be seen from Table 1, although DPSDA-Net ranks second only to NL-LinkNet in precision and second only to D-LinkNet in recall evaluation criteria, both precision and recall of this model remain at a relatively high level. DPSDA-Net performs best in terms of F1 score and IOU. It can be concluded that the overall performance of our model is superior to that of other mainstream methods.

Table 1. Model performance comparison on the Massachusetts roads dataset (the best results are displayed in bold).

Method	Precision	Recall	F1 Score	IOU
U-Net	77.29%	72.13%	73.19%	59.46%
SegNet	72.79%	77.41%	74.26%	60.11%
D-Linknet	74.57%	**78.85%**	75.58%	61.75%
NL-LinkNet	**79.14%**	74.17%	75.64%	62.19%
R-BasicUnet	77.29%	76.07%	75.76%	62.18%
DDU-Net	77.24%	78.29%	76.96%	63.63%
DPSDA-Net	78.28%	78.49%	**77.66%**	**64.46%**

Figure 9 lists the visual segmentation results of different methods tested on the Massachusetts dataset. The first to ninth columns are the original image, ground truth, U-Net, SegNet, D-Linknet, NL-LinkNet, R-BasicUNet, DDU-Net, and our results, respectively. From the first row of pictures, it can be seen that for the obscured roads, the connectivity of the roads in the results extracted by the proposed model was the best. The model also achieved superior extraction results for the occluded part of roads in more dense areas of buildings, and its effect can be proved from the results of the second row of pictures. For the third row of pictures, the proposed model also demonstrated superior performance for the dense road scenes. Finally, from the images in the last three rows, it can be seen that our method has strong anti-interference ability. Compared with the results of other road extraction models, the results extracted by the proposed model contain less noise and are more robust.

Both the evaluation indicators and the visualized images show that DPSDA-Net can fully excavate the local and global contextual semantic information that conforms to road structure characteristics, retaining, to a great extent, the detailed spatial information of the image. It has a better extraction effect than other mainstream models for small roads, complex roads, and occlusion caused by surrounding objects.

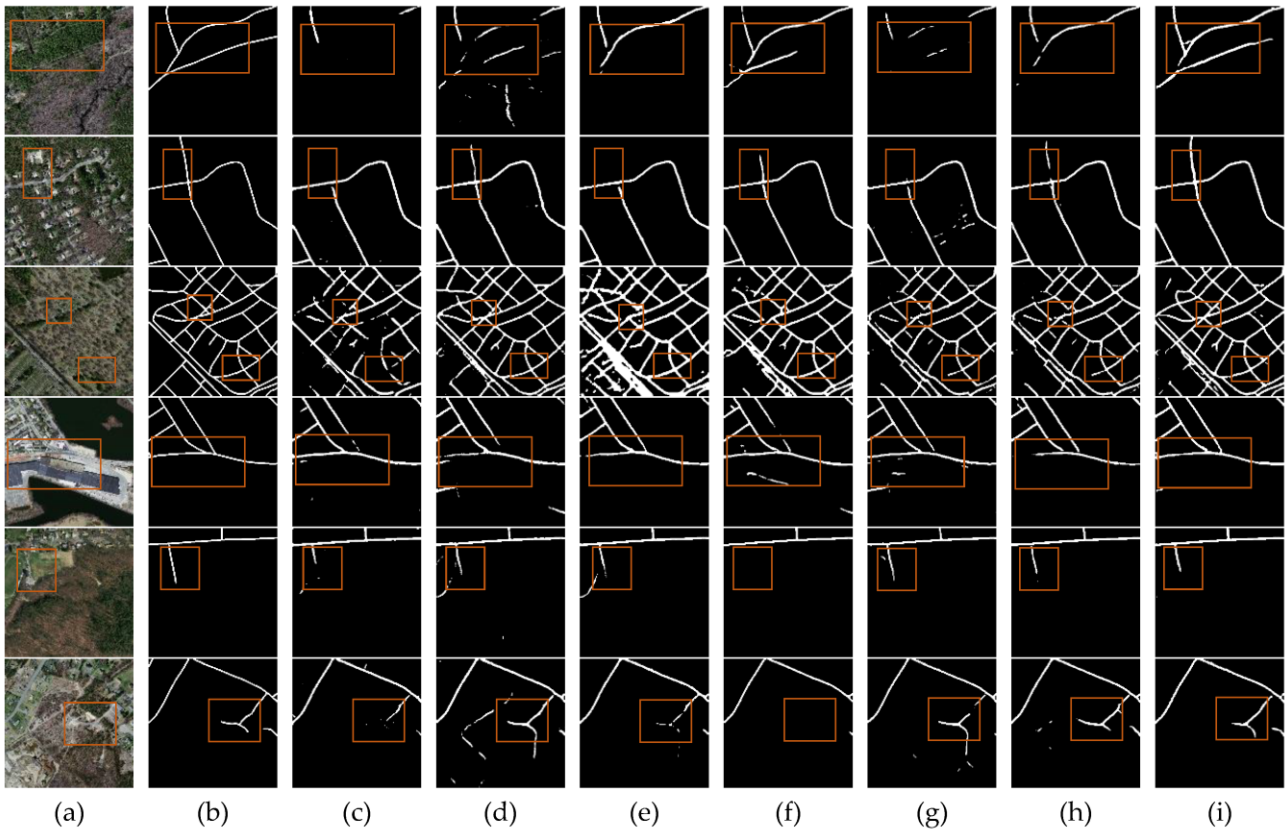

Figure 9. Road extraction results using the Massachusetts roads dataset. (**a**) Satellite imagery. (**b**) Ground truth. (**c**) U-Net. (**d**) SegNet. (**e**) D-LinkNet. (**f**) NL-LinkNet. (**g**) R-BasicUNet. (**h**) DDU-Net. (**i**) Our model.

3.3.2. Comparative Experiment on LRSNY Road Dataset

To further demonstrate the generalization ability of the proposed model, experimental verification was carried out on the LRNSY road dataset. Compared with the Massachusetts dataset, the road performance in the LRSNY dataset was clearer, but the road forms were different: the overall road structure in the Massachusetts dataset is slender and the roads in the LRSNY dataset are wider. Extracting two different types of road datasets is an important requirement for the generalization ability of the model. Table 2 shows the performance comparison of DPSDA-Net and six mainstream methods on the LRSNY dataset; the optimal values are shown in bold. The first column shows the different tested methods compared, which are U-Net [10], SegNet [29], D-LinkNet [13], NL-LinkNet [14], R-Bias-UNet [17], DDU-Net [28], and DPSDA-Net. The second to fifth columns exhibit four different evaluation criteria to show the differences between our proposed model and the other six testing methods. It can be seen from Table 2 that, although the performance of DPSDA-Net on precision is not optimal, it is significantly better than that of U-Net, SegNet, D-LinkNet, NL-LinkNet, R_Basic_UNet, and DDU-Net in terms of IOU, recall, and F1 scores, maintaining a high accuracy, which proves that our model is more reliable.

Figure 10 shows the results of road extraction in the LRSNY dataset with different methods and different complex cases. It can be clearly seen that DPSDA-Net visually outperforms the other tested methods when dealing with different types of roads. As can be seen in the first to third rows, most of the roads in the original image are obscured by trees on both sides, which greatly tests the model's ability to extract the obscured roads, and the integrity extracted by this model is optimal compared with other test methods; from the fourth to sixth rows, for road intersections, the model proposed in this paper performs very well when dealing with such complex regions, with strong road boundary segmentation ability.

Table 2. Model performance comparison on the LRSNY roads dataset (the best results are displayed in bold).

Method	Precision	Recall	F1 Score	IOU
U-Net	**95.18%**	85.86%	89.47%	83.29%
SegNet	93.42%	82.93%	86.13%	79.65%
D-Linknet	93.56%	89.29%	90.69%	84.85%
NL-LinkNet	93.55%	89.02%	90.64%	84.64%
R-BasicUnet	91.73%	89.74%	89.63%	84.1%
DDU-Net	91.69%	90.38%	90.16%	84.58%
DPSDA-Net	92.34%	**91.72%**	**91.76%**	**86.36%**

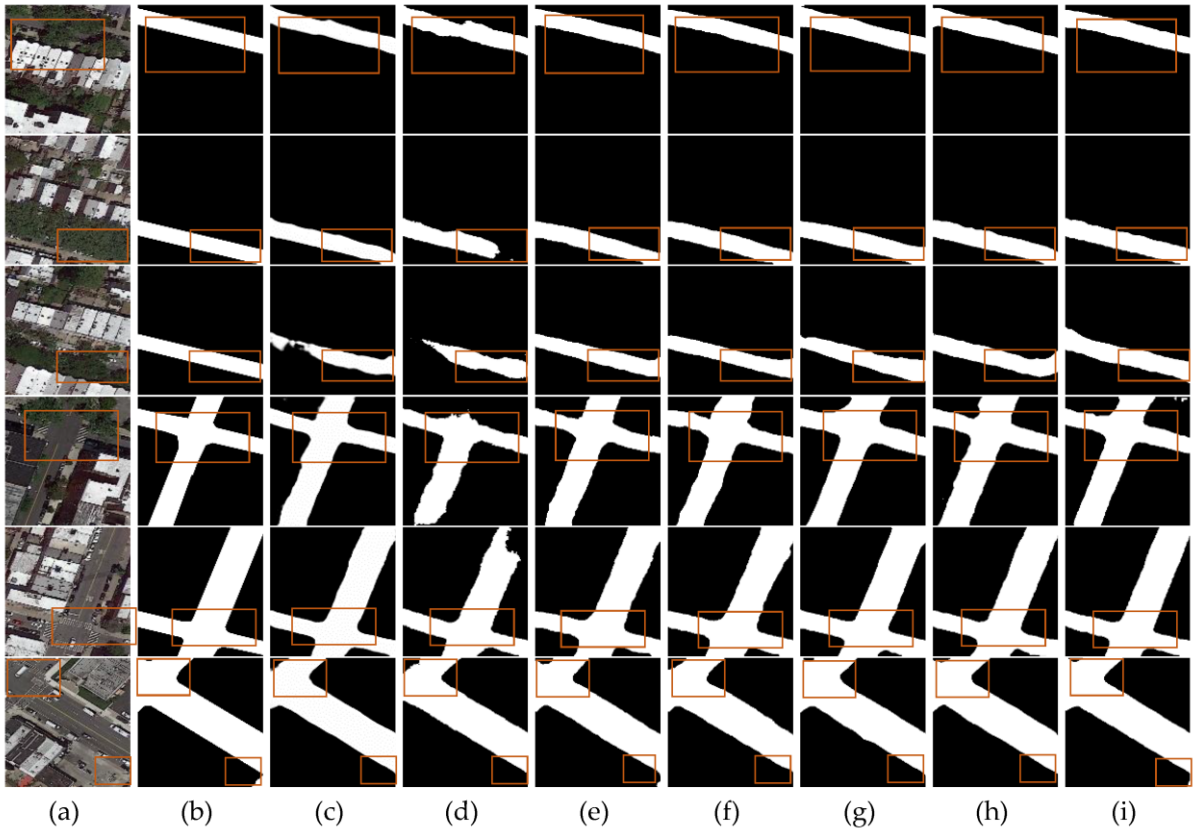

Figure 10. Visual segmentation result from exhibition of different methods tested on LRSNY dataset. (**a**) Satellite imagery. (**b**) Ground truth. (**c**) U-Net. (**d**) SegNet. (**e**) D-LinkNet. (**f**) NL-LinkNet. (**g**) R-BasicUNet. (**h**) DDU-Net. (**i**) Our model.

3.3.3. Analysis for Difference Performance on Massachusetts and LRSNY Datasets

As can be seen from Tables 1 and 2, the F1 scores of DPSDA-Net for the Massachusetts dataset and the LRSNY dataset are 77.66% and 91.76%, and IOU is 64.66% and 86.36%, respectively. Our method achieves the best results compared to the other methods, but the accuracy gap between the two datasets is huge. By comparing the images in the two datasets in Figure 8, it can be seen that the roads in the Massachusetts dataset are mostly slender. The structure of the road is relatively complex, and some roads are not very obvious. In the LRSNY dataset, the distribution density of roads and complexity of the road network are much smaller than those in Massachusetts dataset. Consequently, the model has a huge difference in the extraction results of the two datasets. The accuracy is higher on the LRSNY dataset than that on the Massachusetts dataset.

3.4. Ablation Experiments

The effects of different modules on the performance of the DPSDA-Net model were experimentally analyzed and evaluated in detail to verify the effectiveness of the modules. The U-Net model with the ResNet50 encoder was used to improve the information feature extraction ability of the model.

The results of the ablation experiments are shown in Tables 3 and 4, on the Massachusetts dataset and LRSNY dataset, with the strip position attention module (STPA), long-distance shortcut connectivity operation (LDSC), pyramidal dilated strip convolution module (PDSA), and the combination between any two modules and the three modules, respectively. STPA focuses on location attention perception, which can better learn context information that conforms to road structure features and improve the connectivity of road extraction. The PDSA module can extract the features of different scales of multi-level attention perception and enhance the integrity of the road extraction results. The long-distance shortcut connection operation LDSC module helps the network to retain detailed spatial semantic information during the downsampling process and improve the integrity of road extraction. From Tables 3 and 4, it can be concluded that the design of STPA, LDSC, and PDSA modules improves the F1 score and IOU of the network while adding different modules produces better performance than individual modules, which are indispensable to improve the overall performance of the network.

Table 3. Ablation results on Massachusetts roads dataset ("√" indicates that the module is added to the baseline, the best results are displayed in bold).

STPA	LDSC	PDSA	F1 Score	IOU
√			76.62%	63.16%
	√		76.26%	62.79%
		√	76.49%	63.07%
√		√	77.47%	64.19%
	√	√	77.23%	63.94%
√	√		77.41%	64.01%
√	√	√	**77.66%**	**64.46%**

Table 4. Ablation results on LRSNY roads dataset ("√" indicates that the module is added to the baseline, the best results are displayed in bold).

STPA	LDSC	PDSA	F1 Score	IOU
√			90.83%	85.05%
	√		90.91%	85.38%
		√	91.16%	86.03%
√		√	91.27%	86.11%
	√	√	91.35%	85.93%
√	√		91.17%	86.02%
√	√	√	**91.76%**	**86.72%**

4. Discussion

In this section, the STU-Net, LDU-Net, and PDU-Net single-module network models are obtained by embedding the STPA module, LDSC module, and PDSA module into the baseline network, and experiments are conducted on the Massachusetts dataset and LRSNY dataset to discuss and analyze the roles of different modules.

4.1. STPA Module Discussion Analysis

The STPA module was combined with the baseline network to form the STU-Net, which was trained on both datasets. By comparing the results with the benchmark network model, the influence of the STPA module on the model was explored. The road extraction visualization results are shown in Figure 11.

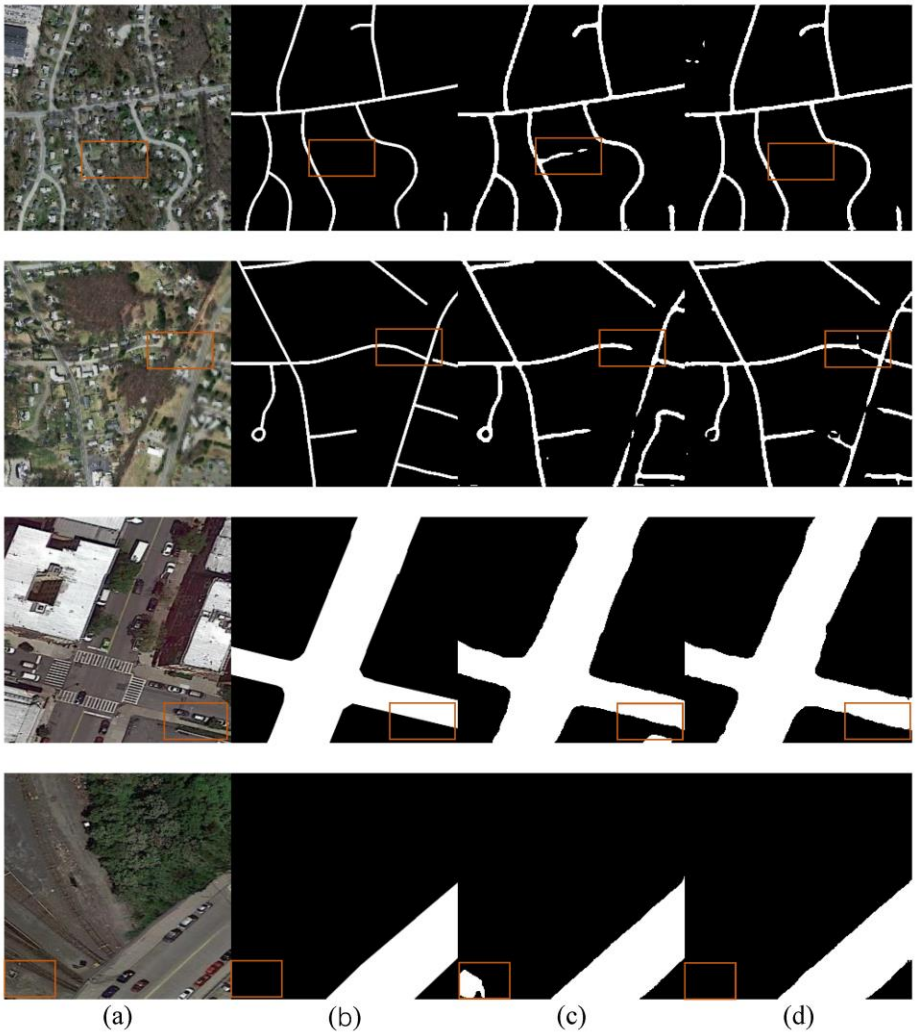

Figure 11. The visual segmentation results of different models tested on the two datasets show that the first two rows are from the Massachusetts dataset and the last two rows are from the LRSNY. (**a**) Satellite imagery. (**b**) Ground truth. (**c**) Baseline. (**d**) STU-Net.

As can be seen in Figure 11, the STPA module can highlight the characteristics of roads and suppress the influence of similar features and noise, improving the accuracy of extraction results. It has a better extraction ability for roads connected to two roads at the same time, which can improve the connectivity of such roads. The STPA module not only focuses on the position of a single pixel, but also uses a striped convolution kernel, which is more consistent with the road shape, to extract long striped semantic data of the road, enabling the network to selectively focus on the important features related to road information, so as to improve the poor road connectivity of the network model and obtain a better visual effect on the connectivity and integrity of the road.

4.2. LDSC Module Discussion Analysis

For the experimental analysis of the long-distance shortcut connection operation (LDSC), LDU-Net, formed by the combination of LDSC and the benchmark network, was carried out to train and test on the two road datasets, from which several representative images are selected, as shown in Figure 12.

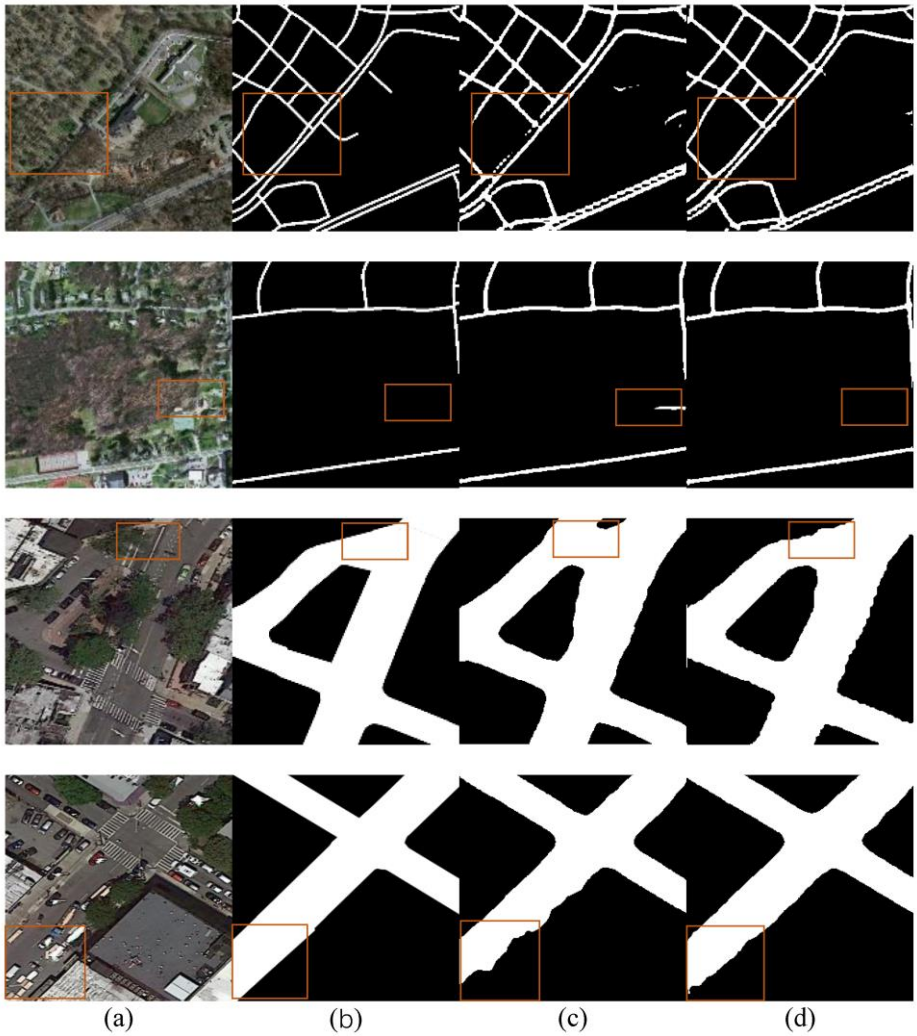

Figure 12. The visual segmentation results of different models tested on the two datasets show that the first two rows are from the Massachusetts dataset and the last two rows are from the LRSNY dataset. (**a**) Satellite imagery. (**b**) Ground truth. (**c**) Baseline. (**d**) LDU-Net.

By analyzing the visualization results of the two datasets, it can be seen that the LDU-Net extraction effect is better for detailed road parts, such as the gap between road intersections, the junction of roads and pedestrian crossings, and the part of the main road covered by vehicles. The analysis also confirms the effectiveness of the LDSC module, enabling the model to focus on road detail extraction. The long-distance shortcut connection operation (LDSC) extracts road features from the original image by using average convolution pooling, standard convolution, and pointwise convolution operations, which can effectively retain detailed information. Finally, combining the output of the LDSC module with the advanced fusion of feature maps makes the model pay more attention to enhancing the details in the original image in road extraction work, which is convenient for guiding the restoration of image details in the process of model upsampling.

4.3. PDSA Module Discussion Analysis

The impact of the PDSA module on the overall model is explored through visual analysis of the experimental results. Using the combination of the PDSA module and the benchmark network, the PDU-Net network model was constructed, and the two datasets were trained and tested separately. Several representative images were selected; the results are shown in Figure 13.

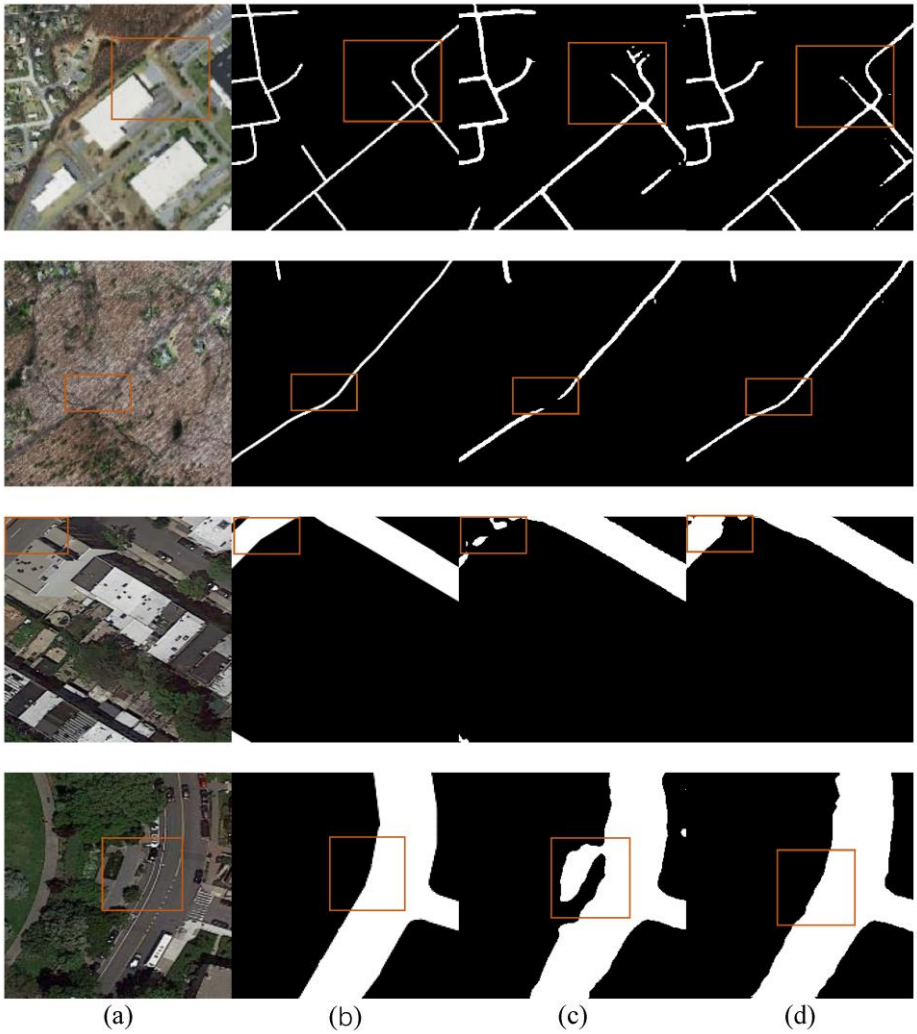

Figure 13. The visual segmentation results of different models tested on the two datasets show that the first two rows are from the Massachusetts dataset and the last two rows are from the LRSNY. (**a**) Satellite imagery. (**b**) Ground truth. (**c**) Baseline. (**d**) PDU-Net.

According to the visualization results of the two datasets, it can be seen that in the test result of the Massachusetts dataset, the road integrity extracted by PDU-Net is better than that of the baseline network model. Although the width of the road in the LRSNY dataset is relatively large, the network model with the PDSA module also demonstrates splendid performance. This is because the downsampling results of the network model are input into the PDSA module, which can obtain global semantic information, generate global semantic information feature maps, pay more attention to the extraction of global information, and represent the linear relationship between different objects. At the same time, the attention mechanism in the PDSA module reinforces the expression of road feature information and adopts the striped convolution operation to further extract semantic information that conforms to the road shape. PDU-Net also achieved better performance, in terms of integrity, for roads with large spans and widths.

4.4. Combined Module Performance Discussion and Analysis

From the above analysis, it can be seen that the single module plays a role in promoting the connectivity, detail, and integrity of the network extraction results, and whether the network, after the combination of modules, maintains high performance is also a major point. It can be seen from Tables 3 and 4 that adding the designed modules has improved the results of the benchmark network. In order to explore the impact of the combined

modules on the performance of the network and analyze the effects between modules, in the meantime, the three modules are embedded into the baseline network. Several representative images are selected for comparison, as shown in Figure 14.

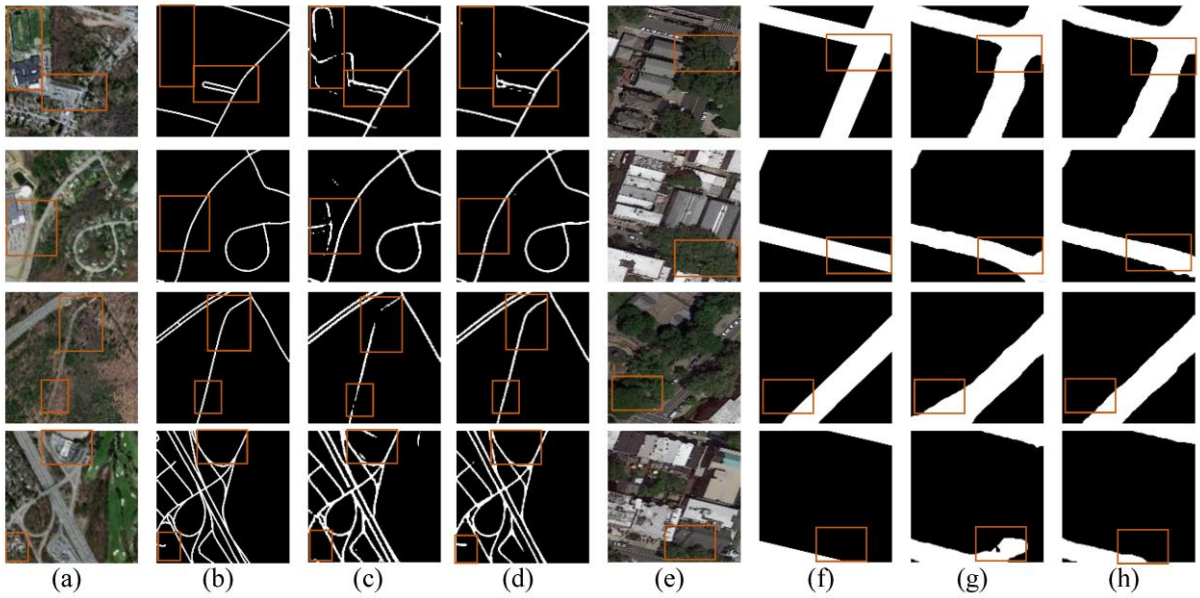

Figure 14. Visual segmentation results' exhibition of complete models tested on Massachusetts and LRSNY datasets. (**a**) Satellite imagery of Massachusetts. (**b**) Ground truth of Massachusetts. (**c**) Baseline results on Massachusetts dataset. (**d**) DPSDA-Net results on Massachusetts dataset. (**e**) Satellite imagery of LRSNY. (**f**) Ground truth of LRSNY. (**g**) Baseline results on LRSNY dataset. (**h**) DPSDA-Net results on LRSNY dataset.

Figure 14 shows that the final network model achieves good performance in terms of road details, connectivity, and integrity extraction. The STPA module uses a combination of strip convolution and standard convolution, focusing on extracting different contextual information so that the information of various elements of the road is more distinguishable. The strip convolution has a narrow shape and is able to capture the local details. The design of the LDSC module helps to retain the detailed information in the original road image and reduces the information loss caused by multiple downsamplings. The PDSA module is embedded between the encoder and decoder for extracting multi-scale contextual features. Meanwhile, it can be seen from Figures 9 and 10 that, compared with other mainstream models, our model shows a significant improvement in road extraction. The extraction of global semantic information and the establishment of long-distance linear relationships are essential. Due to the consideration of the road structural features and the long-distance dependence of information, STPA is used to extract the context semantic information that conforms to the road structural features. The LDSC operation is adopted to retain the more primitive image details, and the bridge connectivity PDSA module is utilized to combine the global information with the detailed information. It further retains the road structure and strengthens the attention to the roads' detail. Each module has a mutually reinforcing role, which can effectively improve the connectivity and integrity of the road and make the model more robust to the extraction of the road.

4.5. Discussion and Analysis of Model Parameter Efficiency

The results of the experiments on the two datasets are shown in Tables 1 and 2, showing that the method in this paper has a significant accuracy improvement over the mainstream methods. The efficiency of road extraction in remote sensing image datasets is also very important. In this subsection, the number of parameters and FLOPS of the analysis model will be discussed; the results are shown in Table 5.

Table 5. Params, FLOPS, and training time of different models with the corresponding ranking marked in the upper right corner.

Methods	Params (M)	FLOPS (G)	Massachusetts (Time/h)	LRSNY (Time/h)
U-Net	31.04 (4)	109.48 (6)	15.5 (5)	6.2 (5)
SegNet	24.94 (3)	66.99 (4)	12.5 (3)	4.5 (3)
D-Linknet	21.64 (1)	13.73 (1)	11.5 (2)	4.3 (2)
NL-LinkNet	21.82 (2)	15.72 (2)	10.2 (1)	4.2 (1)
R-BasicUnet	37.29 (5)	72.83 (5)	13.9 (4)	6.0 (4)
DDU-Net	171.0 (7)	131.65 (7)	19.7 (7)	8.0 (7)
DPSDA-Net	69.81 (6)	36.44 (3)	16.1 (6)	6.9 (6)

Table 5 shows that DPSDA-Net has more parameters than other models. This is mainly due to the design of the STPA and PDSA modules. Among them, in the STPA module, the three feature maps obtained from the input feature maps by 1×1 convolution are subjected to strip convolution operations in four directions, followed by multiplication operations. It increases the complexity of the module and occupies a large amount of memory during the model computation process. In the PDSA module, the strip convolution operation is utilized and the computation of this module is greater than that of the normal bridge module.

Although DPSDA-Net is long in terms of total training time spent, it is superior to DDU-Net, which adopts a double decoding network structure. In terms of model FLOPS, DPSDA-Net remains at a lower level, although U-Net, SegNet, and R-BasicUNet have fewer parameters than DPSDA-Net. U-Net and SegNet use maximum pooling operation and the U-Net decoder adopts deconvolution operation. R-BasicUNet utilizes different convolution kernel decoding branches for each decoder layer, while DPSDA-Net applies two-step convolution downsampling and bilinear interpolation upsampling to reduce the amount of model FLOPS. Although the number of parameters and FLOPS of the D-Linknet and NL-Linknet models were fewer than those of DPSDA-Net, their F1 score and IOU were unsatisfactory. Due to the network structure, DDU-Net also has a large amount of computation. In general, the method proposed in this paper can still achieve good efficiency while ensuring the accuracy of road extraction. However, designing a lightweight network to further improve the accuracy of road extraction while increasing the training speed will be the focus of future work.

5. Conclusions

In this study, a dual-path strip dilated convolutional network DPSDA-Net is proposed to extract road information from very high-resolution remote sensing images. The encoder of DPSDA-Net adopts a dual path in parallel and the two paths extract different mutual information. The first path utilizes strip convolutional modules according to the structural characteristics of narrow and long roads. It extracts semantic information in different directions to maintain the connectivity of the roads. The second path uses a long-distance shortcut connection operation to directly extract the road from the input image. This path preserves more spatial information and prevents information loss when conducting multiple downsampling. In addition, the dilated strip convolution pyramid module is designed in the bridge connection stage, which increases the perception field while ensuring road connectivity. Finally, the compound loss function is introduced to alleviate the imbalance between road and non-road regions in the training sample.

DPSDA-Net achieved significant improvements in overall evaluation metrics (e.g., F1 score and IOU) on the Massachusetts and LRSNY road datasets, achieving higher accuracy compared to other mainstream networks. Concretely, DPSDA-Net maintains the geometric structure of the road network and enhances the connectivity and integrity of the road. We will focus on the design of a lightweight network and plan to perform collaborative extraction for road segmentation and centerline tasks in the future.

Author Contributions: Conceptualization, L.Z.; fund acquisition, L.Z. and M.Z.; project management, L.Z., M.Z. and H.J.; methodology, L.Z. and L.Y.; software, L.Z. and L.Y.; validation, L.Z. and L.Y.; writing—original draft writing, L.Z., L.Y. and M.Z.; supervision and recommendations, H.J., Z.Y. and M.Y. 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 (41901276, 41901265), the Science and Technology Research Project of Henan Province (232102320348, 232102321057), Key Scientific Research Projects of Colleges and Universities in Henan Province (22A420001), the Self-Science Innovation Fund of Henan University of Technology (2021ZKCJ18), and the Cultivation Programme for Young Backbone Teachers in Henan University of Technology (21420198).

Data Availability Statement: The data and the code of this study are available from the corresponding author upon request (mizhang@whu.edu.cn).

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

References

1. Bajcsy, R.; Tavakoli, M. Computer recognition of roads from satellite pictures. *IEEE Trans. Syst. Man Cybern.* **1976**, *9*, 623–637. [CrossRef]
2. Yungang, C.A.; Zhipan, W.A.; Li, S.H.; Xue, X.I.; Lei, Y.A. Fusion of pixel-based and object-based features for road centerline extraction from high-resolution satellite imagery. *Acta Geod. Cartogr. Sin.* **2016**, *45*, 1231–1240.
3. Li, M.; Stein, A.; Bijker, W.; Zhan, Q. Region-based urban road extraction from VHR satellite images using binary partition tree. *Int. J. Appl. Earth Obs. Geoinf.* **2016**, *44*, 217–225. [CrossRef]
4. Lei, X.; Wang, W.; Lai, J. A Method of Road Extraction from High-resolution Remote Sensing Images Based on Shape Features. *Acta Geod. Cartogr. Sin.* **2009**, *38*, 457–465.
5. Netanyahu, N.S.; Philomin, V.; Azriel, R.; Stromberg, A.J. Robust detection of straight and circular road segments in noisy aerial images. *Pattern Recognit.* **1997**, *30*, 1673–1686. [CrossRef]
6. Pan, J.P. Road extraction based on multi-structure element of mathematical morphology. *Comput. Eng. Appl.* **2010**, *46*, 233–235.
7. Zang, Y.; Wang, C.; Yu, Y.; Luo, L.; Yang, K.; Li, J. Joint Enhancing Filtering for Road Network Extraction. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 1511–1525. [CrossRef]
8. Rui, Z.; Jixian, Z.; Haitao, L. Semi-automatic Extraction of Ribbon Roads from High Resolution Remotely Sensed Imagery Based on Angular Texture Signature and Profile Match. *J. Remote Sens.* **2008**, *2*, 224–232.
9. 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, 18–23 June 2018; pp. 207–210.
10. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015, Proceedings, Part III 18*; Springer International Publishing: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
11. Zhang, Z.; Liu, Q.; Wang, Y. Road extraction by deep residual u-net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
12. Chaurasia, A.; Culurciello, E. Linknet: Exploiting encoder representations for efficient semantic segmentation. In Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP), St. Petersburg, FL, USA, 10–13 December 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–4.
13. 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, 18–23 June 2018; pp. 182–186.
14. Wang, Y.; Seo, J.; Jeon, T. NL-LinkNet: Toward lighter but more accurate road extraction with nonlocal operations. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 3000105. [CrossRef]
15. Xu, M.; Li, Y.X.; Zhong, J.J.; Zuo, Z.C.; Xiong, W. L-UNet: Lightweight network for road extraction in cloud occlusion scene. *J. Image Graph.* **2021**, *26*, 2670–2679.
16. 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]
17. Chen, Z.; Wang, C.; Li, J.; Xie, N.; Han, Y.; Du, J. Reconstruction bias U-Net for road extraction from optical remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2284–2294. [CrossRef]
18. Ren, Y.; Yu, Y.; Guan, H. DA-CapsUNet: A dual-attention capsule U-Net for road extraction from remote sensing imagery. *Remote Sens.* **2020**, *12*, 2866. [CrossRef]
19. Ding, L.; Bruzzone, L. DiResNet: Direction-aware residual network for road extraction in VHR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 10243–10254. [CrossRef]
20. Filin, O.; Zapara, A.; Panchenko, S. Road detection with EOSResUNet and post vectorizing algorithm. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–23 June 2018; pp. 211–215.

21. Wang, S.; Mu, X.; Yang, D.; He, H.; Zhao, P. Road extraction from remote sensing images using the inner convolution integrated encoder-decoder network and directional conditional random fields. *Remote Sens.* **2021**, *13*, 465. [CrossRef]
22. Sun, T.; Chen, Z.; Yang, W.; Wang, Y. Stacked u-nets with multi-output for road extraction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–23 June 2018; pp. 202–206.
23. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
24. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
25. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
26. Li, J.; Liu, Y.; Zhang, Y.; Zhang, Y. Cascaded attention DenseUNet (CADUNet) for road extraction from very-high-resolution images. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 329. [CrossRef]
27. Wan, J.; Xie, Z.; Xu, Y.; Chen, S.; Qiu, Q. DA-RoadNet: A dual-attention network for road extraction from high resolution satellite imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 6302–6315. [CrossRef]
28. Wang, Y.; Peng, Y.; Li, W.; Alexandropoulos, G.C.; Yu, J.; Ge, D.; Xiang, W. DDU-Net: Dual-decoder-U-Net for road extraction using high-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4412612. [CrossRef]
29. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
30. Mnih, V. Machine Learning for Aerial Image Labeling. Ph.D. Thesis, University of Toronto (Canada), Toronto, ON, Canada, 2013.