remote sensing

MDPI

Article

RoadFormer: Road Extraction Using a Swin Transformer Combined with a Spatial and Channel Separable Convolution

Xiangzeng Liu [1], Ziyao Wang [1], Jinting Wan [1], Juli Zhang [2], Yue Xi [3], Ruyi Liu [1] and Qiguang Miao [1,*]

[1] School of Computer Science and Technology, Xidian University, Xi'an 710071, China
[2] Academy of Advanced Interdisciplinary Research, Xidian University, Xi'an 710071, China
[3] Guangzhou Institute of Technology, Xidian University, Guangzhou 510555, China
* Correspondence: qgmiao@xidian.edu.cn

Abstract: The accurate detection and extraction of roads using remote sensing technology are crucial to the development of the transportation industry and intelligent perception tasks. Recently, in view of the advantages of CNNs in feature extraction, its related road extraction methods have been proposed successively. However, due to the limitation of kernel size, they perform less effectively at capturing long-range information and global context, which are crucial for road targets distributed over long distances and highly structured. To deal with this problem, a novel model named RoadFormer with a Swin Transformer as the backbone is developed in this paper. Firstly, to extract long-range information effectively, a Swin Transformer multi-scale encoder is adopted in our model. Secondly, to enhance the feature representation capability of the model, we design an innovative bottleneck module, in which the spatial and channel separable convolution is employed to obtain fine-grained and globe features, and then a dilated block is connected after the spatial convolution module to capture more integrated road structures. Finally, a lightweight decoder consisting of transposed convolution and skip connection generates the final extraction results. Extensive experimental results confirm the advantages of RoadFormer on the Deepglobe and Massachusetts datasets. The comparative results of visualization and quantification demonstrate that our model outperforms comparable methods.

Keywords: road extraction; remote sensing image; Swin Transformer; separable convolution

Citation: Liu, X.; Wang, Z.; Wan, J.; Zhang, J.; Xi, Y.; Liu, R.; Miao, Q. RoadFormer: Road Extraction Using a Swin Transformer Combined with a Spatial and Channel Separable Convolution. *Remote Sens.* **2023**, *15*, 1049. https://doi.org/10.3390/rs15041049

Academic Editor: Johannes R. Sveinsson

Received: 27 December 2022
Revised: 2 February 2023
Accepted: 13 February 2023
Published: 15 February 2023

1. Introduction

The extraction of the road from remote sensing images has long been a hot research topic owing to its essential role in applications including automatic driving, vehicle navigation, and road monitoring [1,2]. In the past decades, researchers have achieved good results with high-contrast images using traditional methods involving mathematical morphology and texture analysis [3–5]. However, these methods are usually limited by fixed parameters and have been proven to underperform when applied to low-contrast images [6–8].

From the machine learning perspective, the road extraction work can be regarded as a classification task with two categories (road and background), which is equivalent to a binary segmentation task. Considering the excellent performance of deep learning in recent years for computer vision tasks, researchers nowadays prefer to use deep learning methods to deal with road extraction tasks. Some recent works have explored CNN-based road extraction techniques [9–13], which outperform traditional methods by overcoming the shortcomings mentioned above. However, these works only simplify road extraction to a semantic segmentation problem and ignore the inherent structure of the road. Extracting roads is not an exact segmentation problem due to two reasons. First, the resolution of remote sensing images is usually lower than that of images in general tasks, which means that road segmentation networks should have a large receptive field. Second, since the road areas in remote sensing images are often slender and complicated, the network is supposed to retain the fine-grained feature of the image. CNN-based models are not effective enough

to solve the problem because the receptive field is usually determined by the convolution kernel size. The current CNN-based models mainly use a 3×3 convolution kernel, which is far from satisfying the demand of road extraction tasks, while further increasing the size of the convolution kernel will increase the computational cost with little improvement. Moreover, the pooling will lose image details during image downsampling. Therefore, a new structure is still needed for solving road extraction tasks.

Fortunately, the Vision Transformer (ViT) [14] shows that transformer architecture has excellent potential to face the problems mentioned above. The attention mechanism fuels the transformer to better build long-range dependence so that global information can be utilized at both deep and shallow layers [15]. An increasing number of transformer structures have been developed in different computer vision applications, especially Swin Transformer [16], which has made important achievements in semantic segmentation tasks. Compared with CNN-based models, the Swin Transformer has stronger contextual semantic relevance and a wider receptive field, owing to its shifted windowing scheme and hierarchical architecture. Therefore, the motivation of our model is that introducing the transformer mechanism into the road extraction task may help to further improve the segmentation.

Based on the above discussion, a new road segmentation network with Swin Transformer as the backbone is proposed, named RoadFormer. Considering the distribution and morphological characteristics of roads, an innovative bottleneck is designed. The bottleneck generates the spatial and channel features through the separable convolution and a dilated convolution module in multi-scale is deployed to capture more integrated road structures. The major contributions of this paper can be described as follows:

(1) The proposed model is the first to apply the Swin Transformer as the backbone network to road extraction, achieving an effective perception of global and local road features.

(2) A bottleneck merging the spatial and channel separable convolution and dilated convolution is designed, which makes our model able to capture the local details and global structures of roads more effectively.

(3) Plenty of comparative experiments are implemented, and the visualization and quantitative results show that RoadFormer outperforms the related CNN-based methods [17–28] and Transformer-based methods [29].

The remaining parts of this paper are structured as follows. In Section 2, the overview of previous road extraction works is provided, and the differences between our method and the related methods are also analyzed. In Section 3, the architecture and design of the proposed model are described in detail. In Section 4, implementation details of the experiments are presented, and comparative experiments are conducted and analyzed. Finally, conclusions are given in Section 5.

2. Related Works

In this section, the related road extraction works are reviewed. Then, the structure of the transformer is introduced and its advantages in the road extraction task are analyzed.

2.1. Road Extraction Methods

Numerous approaches for extracting roads from remote sensing images have been presented in recent years, and they may be divided into two primary categories: traditional and deep learning-based methods [30]. Early traditional methods relied heavily on manually designed features or morphological features. Among these methods, the advanced directional morphological operator was presented to prevent the introduction of form biases and successfully retrieve the road shape features. [3]. In addition, linear features that resemble ribbons or ridges are extracted to categorize the road regions, which perform more robustly than previous methods [31]. However, these traditional methods usually lack robustness to incomplete structure, illumination, and contrast changes [6,7].

To solve the difficulties existing in the traditional methods mentioned, deep learning-based approaches were employed for road extraction. As a representative of CNN-based methods, a patch-based CNN model was proposed for road extraction from high-resolution remote sensing data [32]. Later, RoadNet [13] was presented to extract the road surface, centerlines, and edges in several tasks. In order to preserve more spatial detail information and enhance road integrity, a superpixel segmentation and graph convolutional network was recently developed [33]. The CNN-based methods above can accomplish high accuracy, while their processing speed has to be increased.

In order to address the shortcomings of CNNs, the fully convolutional network (FCN) substitutes the fully connected layer with deconvolution, which achieves end-to-end pixel-level classification. In the early works, it was established that the FCN approach was successful in maintaining the continuity and integrity of roads for road extraction tasks [17]. Later, it was suggested to use UFCN to extract roads from aerial images taken by UAV [34]. Subsequently, FCN-32 was applied for extracting the road in the high-resolution image [35]. To comprehensively utilize multi-scale information from images, U-Net series-equipped skip connection modules were developed [18,19,36]. SegNet [22] adopted the encoder–decoder structure, where the edge position can be restored in the decoder by the index value reserved in the encoder. Recently, to obtain better segmentation results, DeepLab series methods [21,37,38] employed dilated convolution to capture long-range information and developed a pyramid-shaped pooling layer to retain the spatial structure.

Although FCN models improve the efficiency of road extraction, they often misclassify road areas and backgrounds in highly complex scenarios. Meanwhile, FCN-based models will lose edge position information due to the existence of pooling layers. In addition, missing long-range information limits the segmentation accuracy of U-Net and SegNet. Additionally, dilation convolution makes Deeplab perform well in large target extraction but poorly in small targets. To solve the problems above, we introduce the transformer structure to our road segmentation task.

2.2. Transformer-Based Approaches

Lately, transformer architecture [39] has become vibrant in the computer vision field in view of its special attention mechanism. Transformer's attention mechanism can enable it to learn long-range features and model global information, in contrast to CNN's emphasis on local features. The Vision Transformer (ViT) [14] accomplished satisfactory results in image classification and showed great potential in computer vision, where the image patches are considered the token of the transformer module. Although the design is feasible, there are still many apparent disadvantages [29]. The quadratic computational load imposed by transformers brings a considerable cost that is intolerable in segmentation tasks for large-size images. Furthermore, although transformer could capture long-range information and global context, it is difficult to capture low-level information needed in segmentation [40].

To reduce the memory requirements of transformers, Liu et al. [16] conceived the Swin Transformer, which adopts a strategy of merging neighboring patches to build a hierarchical representation structure. With these hierarchical representations, the model can easily make dense predictions using a feature pyramid network. Meanwhile, the Swin Transformer computes self-attention in non-overlapping windows with only linear computational complexity. These advantages make it suitable as a segmentation backbone. In view of the global feature-capturing capability and lower computational complexity of the Swin Transformer, we introduce it as an encoder into our network.

2.3. Feature Separation

For the road extraction task, an obvious challenge is that the distribution of roads requires the model to have strong long-range information acquisition ability, while the slender and complex road characteristics require the model to have enough detail processing ability. Having both of these capabilities for general convolution operations would be contradictory. According to Tao et al. [41], the spatial and channel features of roads exist

apparent differences, and thus processing the features of different dimensions separately can improve the accuracy of segmentation. From the perspective of information representation, the channel features can reflect the image's local details, and spatial features can help the network capture long-range information. Therefore, for road extraction, it is necessary to distinguish spatial and channel properties.

In previous works, depth-wise (DW) separable convolution was intended to divide the conventional convolution into depth-wise and point-wise, effectively reducing the computational complexity [42]. Compared with traditional convolution, DW separable convolution has fewer numbers of parameters and a lower cost of operation but still achieves almost the same results. Zhou et al. [28] use DW separable convolution combined with a graph convolution network (GCN) to achieve feature separation. Motivated by the previous work above, we replaced the original DW separable convolution series structure with a parallel structure to obtain the channel and spatial features.

3. Method

This section provides a detailed description of the proposed model's architecture. In Section 3.1, the overall design of RoadFormer is introduced. Then, the workflow of the encoder is described in Section 3.2, and the design of the bottleneck for road feature refining is presented in Section 3.3. Lastly, Section 3.4 provides the decoder and loss function.

3.1. RoadFormer Overall Design

We provide a road extraction model called RoadFormer to overcome the receptive field constraints and capture detailed information in remote sensing images. The architecture of RoadFormer is divided into three sections, as displayed in Figure 1:

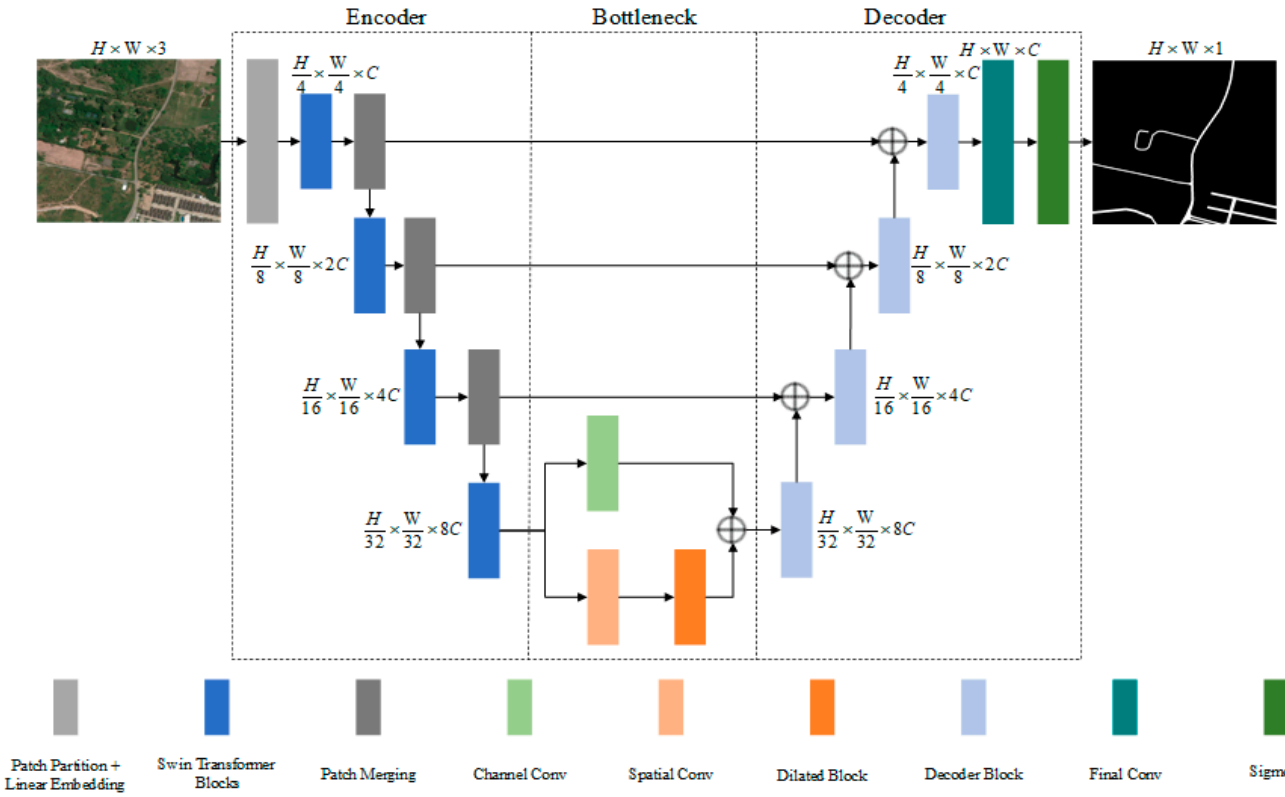

Figure 1. RoadFormer architecture consists of an encoder, bottleneck, and decoder. Multi-scale feature representation can be produced by the encoder. The high-dimension feature maps can be obtained by the separable convolution and dilated block in the bottleneck. The final results are given by the decoder.

(1) Swin Transformer-based encoder: the encoder downsamples and encodes the input RGB image into multi-scale high-dimensional feature maps, which are necessary inputs for the decoder and bottleneck.

(2) Feature separation bottleneck: the bottleneck separates the high dimensional input feature maps into channel and spatial features. Meanwhile, a dilated block consisting of four dilated convolution layers is applied to the spatial feature to expand the receptive field.

(3) Lightweight decoder: bottleneck-generated feature maps are alternately upsampled and merged with encoder-generated feature maps to the top decoder block. Then, the segmentation result is obtained from the top decoder by using transposed convolution and a sigmoid.

In the subsequent sections, each network component will be described in detail.

3.2. Encoder

Without loss of generality, the distribution of roads should be continuous and throughout the whole image, and the model is supposed to have a great capacity to collect long-range information. We adopted the Swin Transformer as the encoder for the suggested model because of its prowess in modeling long-range information relationships. Different from the transformer, the Swin Transformer replaces the multi-head self-attention module (MSA) with a block that can be made up of shifted window-based MSA, MLP, LayerNorm, and a residual connection. Continually and alternately, the W-MSA and SW-MSA (MSA with regular and shifted windowing configurations) are applied in a block. The structure of the Swin Transformer blocks is presented in Figure 2.

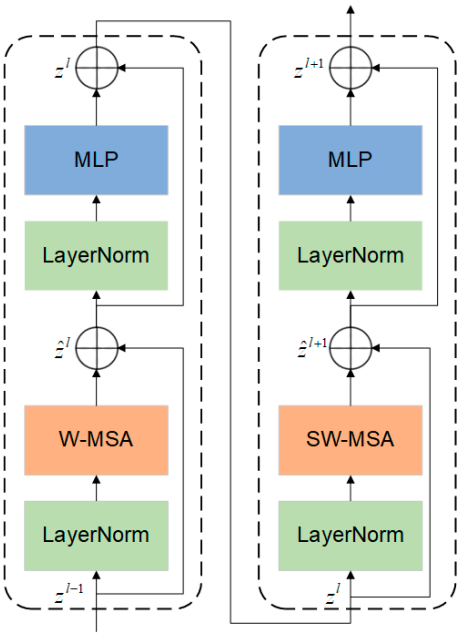

Figure 2. The architecture of Swin Transformer blocks.

The encoder of the proposed model is composed of four stacked Swin Transformer modules. The original image ($H \times W \times 3$) is transported to the patch partition in the first layer and divided into patches ($\frac{H}{4} \times \frac{W}{4} \times 48$). Then, these patches are converted into tokens by linear embedding layer mapping. After that, the tokens are fed successively alternately into Swin Transformer blocks and patch merging layers to create a hierarchical representation. To be specific, Swin Transformer blocks produce feature maps at the current layer scale while patch merging layers downsample these maps. Notably, the output of patch merging layers is simultaneously supplied by skip connection to the relevant layer of the decoder and handled as the input of the next Swin Transformer block.

3.3. Bottleneck

To obtain the spatial and channel features effectively, a parallel structure combined with DW and PW convolution is developed in RoadFormer. The process of the separable convolution module is shown in Figure 3. Specifically, parallel connections between spatial convolution and channel convolution are made after the encoder. In the channel convolution part, a 1D convolution kernel is used to convolute the feature map along the channel direction. In the spatial convolution part, each feature map is convoluted by a $k \times 1 \times 1$ kernel and concatenated as spatial feature maps. The refined feature maps by spatial convolution and channel convolution have a size of $H \times W \times N$, which is consistent with the input.

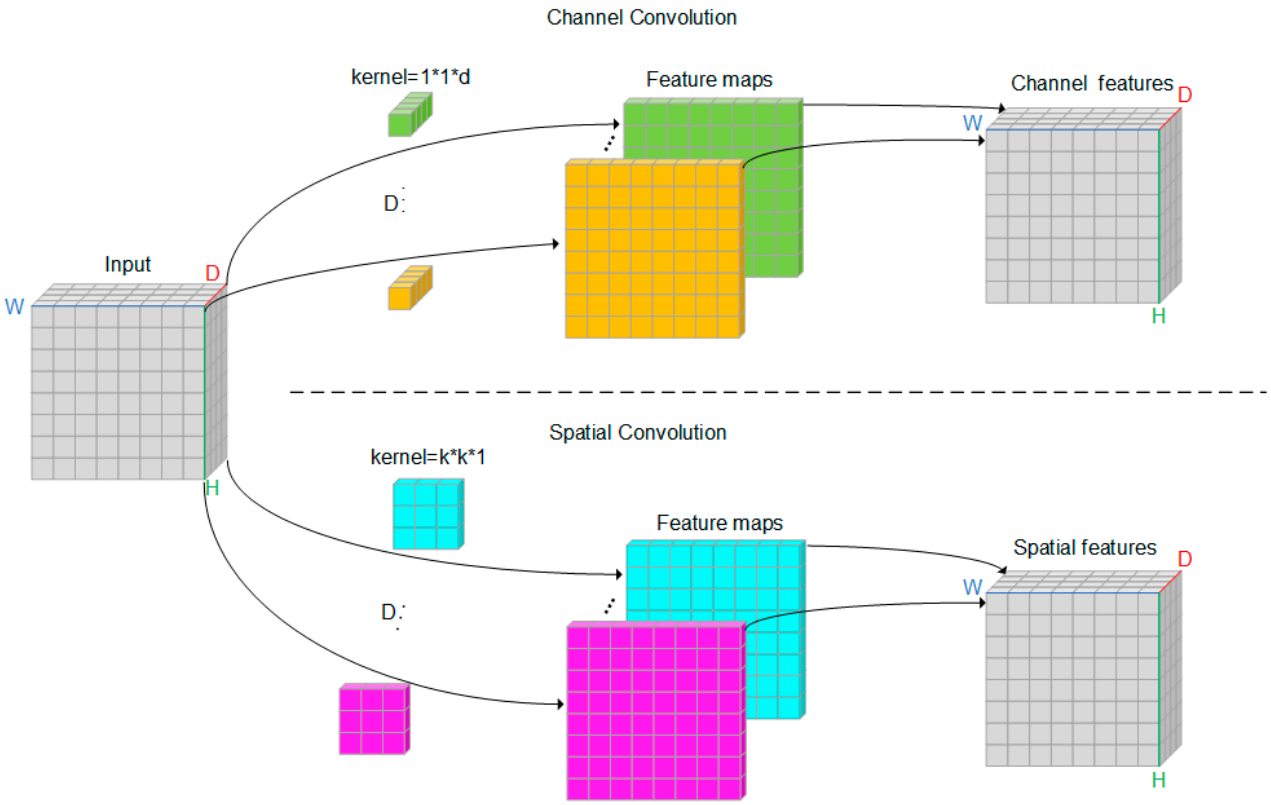

Figure 3. Structure of the parallel separable convolution. The separable convolution split the features into channel features and spatial features.

Previous works have proved that traditional convolution tends to have a finite receptive field, which performs not well in segmentation tasks [20,43]. Fortunately, dilated convolution can effectively expand the receptive fields while keeping the resolution of feature maps. Referring to D-LinkNet [24], we set a cascade and parallel structure of dilated convolution after the spatial convolution module. The receptive fields of each layer will be 3, 7, 15, and 31 if each layer's dilation rates are set to 1, 2, 4, and 8, as demonstrated in Figure 4. The Swin Transformer encoder downsamples the original input with one reduction of $\frac{1}{4}$ and three reductions of $\frac{1}{2}$. For an image with a size of 1024×1024, the output feature map size of the encoder will be 32×32. In this case, the receptive field of the dilated block can cover almost the entire range of the feature map. Through this design, this architecture considerably enhances our model's capacity to capture long-range information.

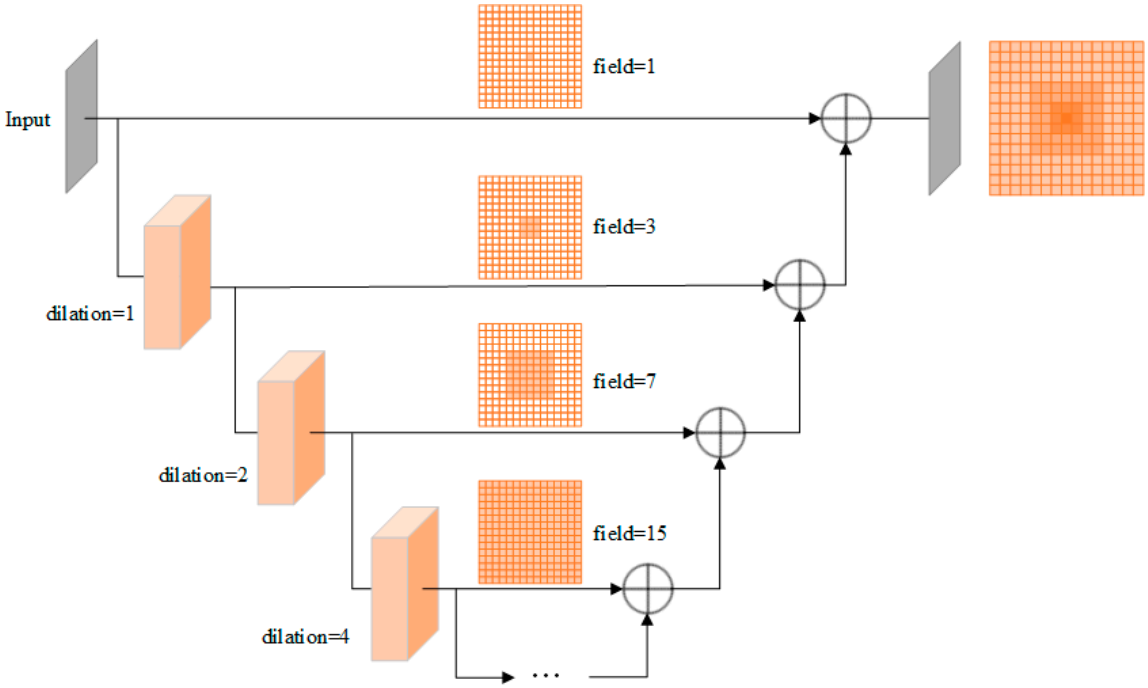

Figure 4. Structure of the dilated block. Multi-scale receptive fields are constructed through dilated convolution layers, which makes the network can extract different scale features.

3.4. Decoder and Loss Function

To recover the segmentation details, a decoder is employed in RoadFormer. Symmetrically with an encoder, four decoder blocks and a final convolution layer are adopted to upsample the feature maps. Figure 5 depicts the decoder block's structural layout. Specifically, in each decoder block, the features are filtered by a 3×3 convolution layer first and are upsampled by a transposed convolution layer subsequently. Then, the features are filtered by a 3×3 convolution layer again. After the convolution, the upsampled features are added with the results of the encoder in the corresponding scale. After going through four decoder blocks, one transposed convolution layer and two convolution layers with 3×3 kernels will process the feature maps to be the same size as the source image. Lastly, a sigmoid classifier is applied to extract road areas by mapping the output to a range of 0 to 1, where the threshold is set to 0.5 to classify the road areas and background.

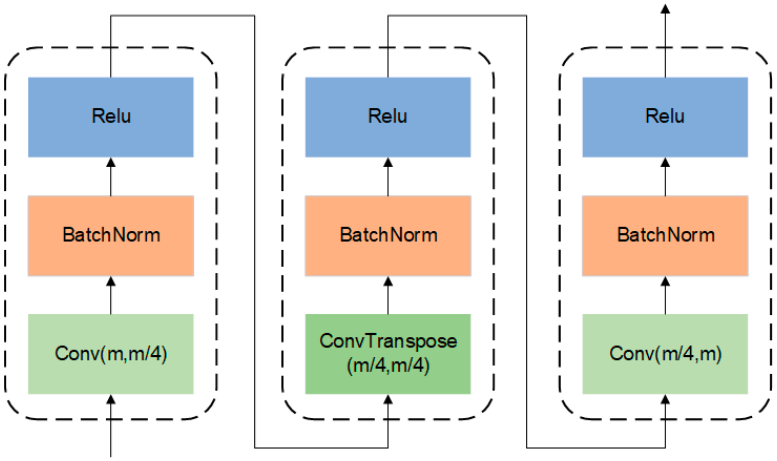

Figure 5. Structure of the decoder block.

Binary cross entropy (BCE) loss and dice coefficient loss make up RoadFormer's loss function. The BCE loss, which is most frequently employed in the binary segmentation task, is defined as follows:

$$L_{BCE} = -\frac{1}{N}\sum_i \left(t_i \cdot \log(o_i) + (1 - t_i) \cdot \log(1 - o_i)\right) \quad (1)$$

where o indicates the predicted results after sigmoid, t indicates the true label, and N indicates the batch size.

Road segmentation is a particular scenario where the foreground and background are severely imbalanced. Therefore, the loss function should have adaptability for unbalanced data distribution. Dice loss is more focused on the mining of foreground regions during training, whose supervised contribution to the network does not vary with the size of the image. Therefore, it is suitable for solving the situation where the foreground accounts for a relatively small amount. The formulation of the dice loss is:

$$L_{Dice} = 1 - \frac{2\sum_i o_i \cdot t_i}{\sum_i o_i + \sum_i t_i} \quad (2)$$

To prevent a zero in the denominator, we added a smooth parameter s. The optimized L_{Dice} can be described as follows:

$$L_{Dice} = 1 - \frac{2\sum_i o_i \cdot t_i + s}{\sum_i o_i + \sum_i t_i + s} \quad (3)$$

The smooth parameter avoids the zero division problem and prevents the overfitting of the model. The total loss can be computed as:

$$L_{total} = \alpha L_{BCE} + \beta L_{Dice} \quad (4)$$

where α and β denote the weights that could balance the two loss functions.

The loss function designed above makes the model extract roads accurately and retain road connectivity. Through the loss function design, the feature information is most effectively conveyed to the segmentation result, which could ensure road extraction accuracy and retain road connectivity simultaneously.

4. Experimental Results and Analysis

In this section, the dataset and model training details are introduced first. Subsequently, the evaluation metrics commonly used in road extraction tasks are presented. Next, the ablation experimental results are analyzed, which confirms the validity of our model design. Finally, visualization and quantitative results of our approach in comparison to other SOTA methods are then shown.

4.1. Datasets and Experiment Implementation

Datasets: In this paper, the Deepglobe dataset and the Massachusetts road dataset are used for the experiment, as shown in Figure 6. The following is a detailed description of the two datasets:

1. Deepglobe Dataset: Deepglobe is the dataset prepared for the 2018 Deepglobe road extraction challenge. This dataset includes 6226 images with a resolution of 0.5 m and a size of 1024 × 1024 pixels. These RGB images in JPG format cover Thailand, India, and Indonesia, and include roads of cement, asphalt, and mountain. Each annotation image is a three channel binary image in PNG format, which uses (255, 255, 255) and (0, 0, 0) to present roads and backgrounds, respectively. In the experiment of our model, the dataset was split into the training set (4987 images) and the test set (1246 images).

2. Massachusetts dataset: The Massachusetts road dataset consists of 1108 images for training, 14 images for validation, and 49 images for testing, all of which are 1500 × 1500 in size. According to [44], the resolution of Massachusetts can be inferred to be about 1.5 m. The source image in TIF format is three channel color image and its label in TIFF format is a binary image that uses white and black to distinguish roads and backgrounds. The roads of cement and asphalt are the main types in this dataset.

Figure 6. Some samples of the Deepglobe dataset and the Massachusetts dataset, the first and second rows are for the Deepglobe dataset, and the third and fourth rows are for the Massachusetts dataset.

Data augmentation: In order to demonstrate that our model works effectively on large-size remote sensing images, we directly use uncropped images with 1024 × 1024 size as the input of the network. To comprehensively utilize the limited training set, we employ geometric transformation and photometric distortion to augment the data. The geometric transformation method includes random clipping and horizontal and vertical flip transformation. In the photometric distortion part, random luminance transformation and random contrast transformations are used. Saturation and hue transformation are applied after the RBG image is converted to HSV space. In addition, test time augmentation, including horizontal flip and vertical flip, is adopted in the testing phase. In this phase, the predicted results are restored to match the origin direction, and the final predicted results are given according to the average of augmentation outputs.

Implementation detail: All the experiments are implemented on an NVIDIA GeForce RTX 3090 GPU using Pytorch in a Linux environment. To make the model have better results, the learning rate schedule strategy is employed. Specifically, we adopt a poly strategy to modify the learning rate dynamically to make the model have a better convergence speed. An adaptive moment estimation (Adam) optimizer is applied in the training phase of our model. Meanwhile, multiple sets of learning rate parameters are set, and according to the convergence of the model, 2e-4 was selected as the start learning rate. In the ablation experiments, different pretrained Swin Transformers are employed to test the performance

of road extraction. We train RoadFormers with Swin-T, Swin-S, and Swin-B as the backbone using 4, 4, and 2 as the the batch size, respectively.

4.2. Evaluation Metrics

Road extraction can be approached as a segmentation problem with two classes of roads and backgrounds [30]. Therefore, the effectiveness of the road extraction models is assessed using the evaluation metrics of binary segmentation. Precision (Pr), recall (Rc), F1-score, and intersection over union (IoU) are the four performance evaluation measures that are most frequently utilized. Precision reflects the percentage of road extraction results that are correctly classified, which can be formulated as:

$$\mathrm{Pr} = \frac{\mathrm{TPre}}{\mathrm{TPre} + \mathrm{FPre}} \tag{5}$$

where true positive and false positive of road extraction (TPre and FPre) represent the numbers of pixels correctly and incorrectly classified as road areas, respectively. Different from precision, recall represents the percentage of properly recognized pixels in the whole road label, which can be formulated as:

$$\mathrm{Rc} = \frac{\mathrm{TPre}}{\mathrm{TPre} + \mathrm{FNre}} \tag{6}$$

where the false negative of road extraction (FNre) denotes the number of road pixels extracted as other areas. In addition, the F1-score, which offers a more thorough evaluation of the model's performance, is the harmonic mean of Pr and Rc. It can be calculated as follows:

$$\text{F1-Score} = \frac{2\mathrm{TPre}}{2\mathrm{TPre} + \mathrm{FPre} + \mathrm{FNre}} \tag{7}$$

Without loss of generality, IoU is the intersection of ground truth and road extraction findings divided by their union, which can be calculated as follows:

$$\mathrm{IoU} = \frac{\mathrm{TPre}}{\mathrm{TPre} + \mathrm{FNre} + \mathrm{FPre}} \tag{8}$$

The four evaluation metrics mentioned above are adopted in our quantitative experiments.

4.3. Ablation Experiments

In this part, the ablation experiment is carried out to evaluate the performance of encoder modules with different backbones. We use ResNet-50 and the Swin Transformer series as the encoder of the network. According to different configurations, the Swin Transformer can be formed as Swin-T (tiny size), Swin-S (small size), and Swin-B (base size). As shown in Table 1, Swin-T achieved better results under the condition that the number of parameters of ResNet-50 is comparable. Among them, Swin-B achieved the best performance with four times the number of parameters of ResNet-50. To trade off the performance and cost of the model, Swin-S was selected for subsequent ablation experiments. In the comparison experiments, we mainly use the results of Swin-B for comparison because of its better performance. In practices with computational efficiency needs, Swin-T is a good choice because of its small size and fast speed.

Table 1. Quantitative comparison of different backbones for RoadFormer using the Deepglobe dataset.

Backbone	Params	Precision	Recall	IoU	F1-Score
ResNet-50	21.66 M	84.91	78.61	68.14	81.64
Swin-T	28.30 M	84.05	81.34	69.91	82.67
Swin-S	49.59 M	85.29	82.51	72.18	83.88
Swin-B	88.07 M	85.76	83.17	73.11	84.50

We conduct another ablation experiment to demonstrate that the bottleneck part is valid. The quantitative results of different module configurations are shown in Table 2. As is seen from the results, the spatial and channel separable convolution significantly enhances the model's overall performance. We set dilated blocks after spatial convolution and channel convolution, respectively. Obviously, it makes sense to treat global and detailed features separately. The model with feature separation performs significantly better with recall, F1-score, and IoU than the model without such structure. Meanwhile, dilated block after spatial convolution improves IoU and F1-score. In contrast, although the addition of dilated block after channel convolution increases the precision, the other performance metrics are reduced. This is due to the fact that the effect of the dilated block is to expand the receptive fields, which are compatible with the spatial features being separated out. However, the channel features separated by a 1×1 convolution focus on the information of the pixel itself, and it is meaningless to expand its receptive fields. The above results confirm that spatial convolution followed by dilated convolution improves the performance, while that becomes worse after channel convolution.

Table 2. Quantitative comparison of different configurations of the bottleneck.

Methods	Feature Separation	Dilated Block	Precision	Recall	IoU	F1-Score
RoadFormer	×	×	83.71	80.70	69.66	82.18
RoadFormer	√	×	82.83	**83.68**	71.28	83.26
RoadFormer	√	channel	**86.79**	79.45	70.86	82.95
RoadFormer	√	spatial	85.29	82.51	**0.7218**	**83.88**
RoadFormer	√	channel + spatial	86.47	80.80	71.77	83.54

The results of the ablation experiment with different configurations are presented in Figure 7. Among different configured models, the road extraction results with feature separation and dilated block have better continuity and detail. In summary, the feature separation module effectively enhances the comprehensive performance of the model.

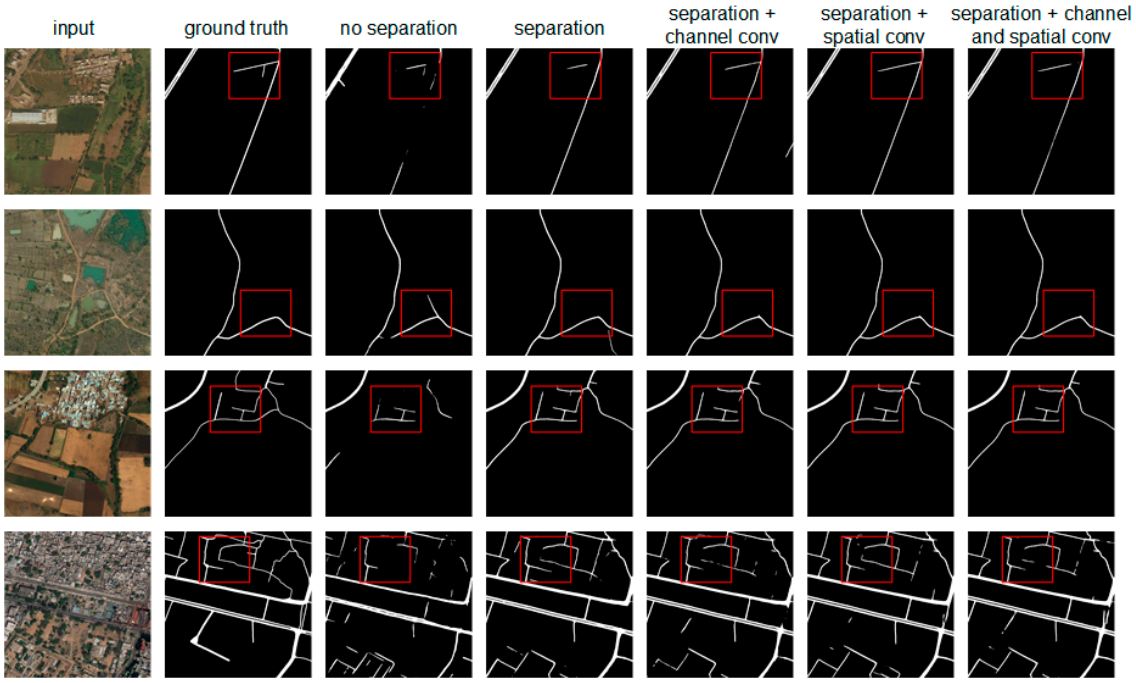

Figure 7. Visualization results of different configurations of the bottleneck. In the 1024×1024 size image, the 350×320 size red boxes highlight the areas where RoadFormer's road extraction results are better.

4.4. Comparative Experiments

We conduct the experiments via a comparison with SOTA approaches on the Deepglobe dataset and Massachusetts road dataset in terms of accuracy, recall, IoU, and F1-score to completely evaluate the effectiveness of the proposed approaches. Visualization results of the proposed model with five representative models are presented, and quantitative analysis and results are given in this section.

4.4.1. Experiments on the Deepglobe Dataset

On the Deepglobe dataset, RoadFormer was compared with FCN, U-Net, PSPNet [20], DeeplabV3, Seg-Net [22], LinkNet [23], D-LinkNet [24], HourGlass [26], Batra et al. [27], and SwinUnet [29]. Among the methods above, FCN and U-Net are representatives of the classic segmentation models. PSPNet employed a pyramid pooling structure to gather information about the context. DeeplabV3 developed dilated convolution to enlarge the receptive field and aggregate the multi-scale features by using an ASSP module. SwinUnet is a novel transformer-based model originally used for medical image segmentation. We show the visualization results obtained by RoadFormer and these five presentative methods above. For the other methods, we quoted the quantitative results from their source, so visualization results are missing as they were not available.

For an intuitive evaluation of road extraction performance, eight representative images with different scenes were chosen from the test set. Figure 8 shows the road extraction results of these images by using six different methods, respectively. The extracted roads of eight images are listed in eight rows and eight columns. The input images, ground truth images, and results of FCN, U-Net, PSPNet, DeeplabV3, SwinUnet, and RoadFormer are displayed in the left-to-right columns. For the image of the town scene (first to third rows), the results obtained by U-Net and DeeplabV3 miss much road information, while other methods work well. In obscured scenes (fourth to sixth rows), the discontinuous road structures all appeared in the results of other methods, and RoadFormer's extraction results remain complete. For low-contrast scenes (the seventh and eighth rows), none of the five methods can extract the road structure completely, while RoadFormer is able to extract road areas precisely. It is worth noting that the roads extracted by SwinUnet perform better than other CNN-based models in terms of continuity, which is due to the long-range dependence established by the transformer. However, SwinUnet is missing some of the slender roads, while RoadFormer still performs well in this case due to its bottleneck design. The visualization results above show that the performance of RoadFormer outperforms the other methods. The integrity and continuity of the road are well preserved due to the long-range information-capturing ability and feature separation strategy of RoadFormer.

For making a more thorough evaluation of the proposed method, we quantified and compared RoadFormer with SOTA methods, including FCN, U-Net, PSPNet, DeeplabV3, Seg-Net, LinkNet, D-LinkNet, HourGlass, Batra et al., and SwinUnet. Table 3 displays the quantitative performance results of these methods on the Deepglobe dataset. RoadFormer (with a Swin-B backbone) obtains the best results for precision (85.8%), IoU (73.1%), and F1-score (84.5%), and the second-best result for recall (83.2%), which is only less than Batra et al. It is worth pointing out that the method of Batra et al. employed a strategy of multi-task learning considering road direction information. The method enhances the correlation between the extracted segments but also leads to an increase in cost. The lightweight RoadFormers (using Swin-T and Swin-S as the backbone) still outperform most SOTA methods in terms of performance metrics. This result substantiates the reliability of the model structure we suggested.

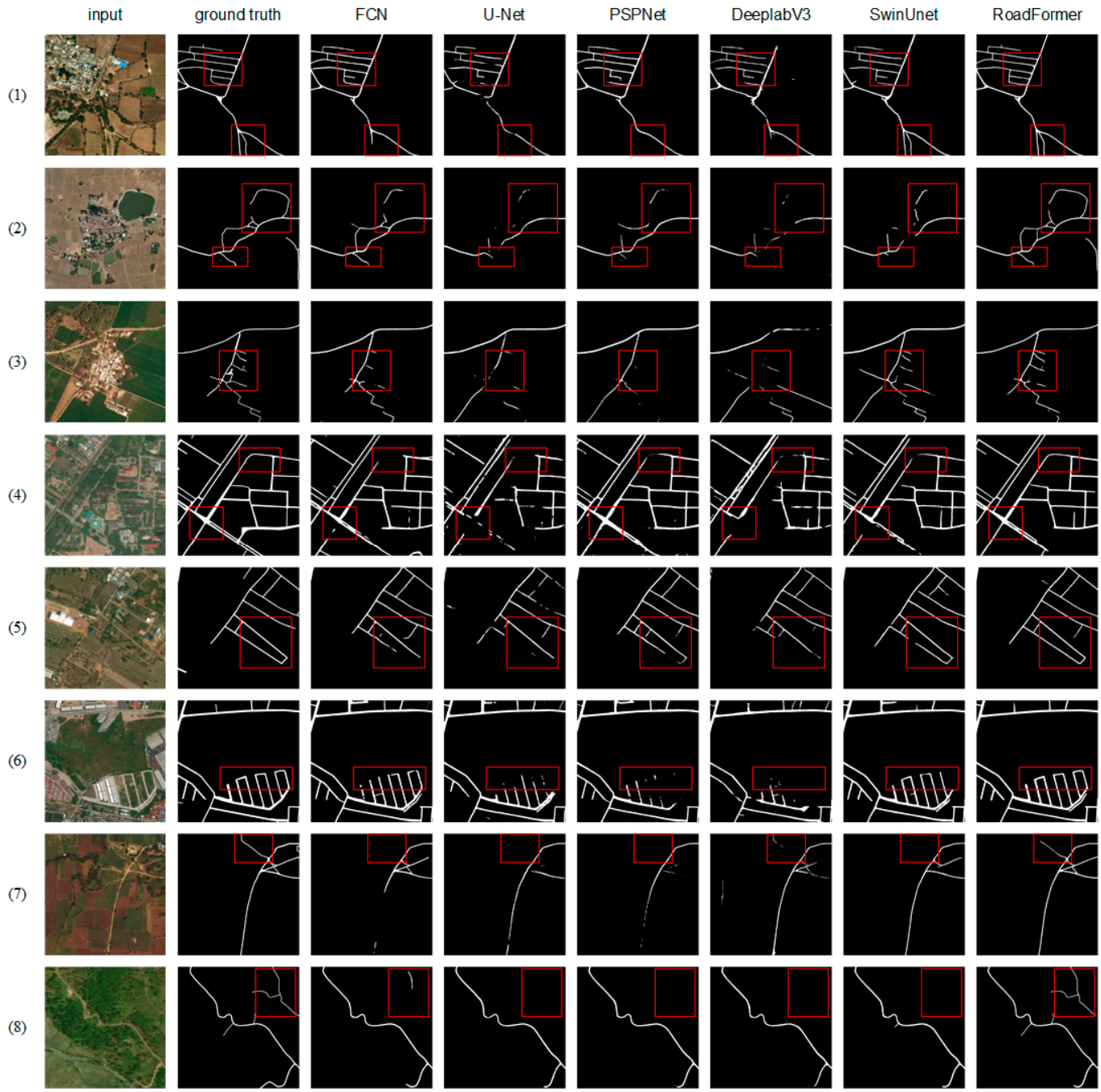

Figure 8. Visualization results of RoadFormer and other deep-learning models on the Deepglobe dataset. The red boxes highlight the regions where our method outperforms other methods. The size of the red boxes in each image from top to bottom are (**1**) 320 × 272, 282 × 256, (**2**) 410 × 410, 298 × 160, (**3**) 320 × 336, (**4**) 246 × 202, 282 × 272, (**5**) 432 × 426, (**6**) 608 × 192, (**7**) 320 × 240, and (**8**) 226 × 403.

Table 3. Results of the Deepglobe dataset's quantitative performance.

Methods	Precision	Recall	IoU	F1-Score	Param	FLOPs
FCN	83.1	75.5	64.8	79.1	47.1 M	197.7 G
U-Net	82.6	64.0	55.3	72.1	29.1 M	202.5 G
PSPNet	84.7	70.1	60.1	76.7	49.0 M	178.4 G
DeeplabV3	78.9	58.3	50.0	67.0	65.7 M	270.0 G
Seg-Net	69.5	73.0	55.3	71.2	-	-
LinkNet	78.3	78.8	64.7	78.6	-	-
D-LinkNet	84.9	78.6	68.1	81.6	-	-
HourGlass	79.4	80.1	66.3	79.8	-	-
Batra et al.	83.8	**84.1**	72.4	84.0	-	-
SwinUnet	82.1	73.3	62.9	77.7	27.1 M	254.8 G
RoadFormer (Swin-T)	84.1	81.3	69.9	82.7	28.3 M	176.5 G
RoadFormer (Swin-S)	85.3	82.5	72.2	83.9	49.6 M	269.4 G
RoadFormer (Swin-B)	**85.8**	83.2	**73.1**	**84.5**	89.0 M	447.7 G

4.4.2. Experiments on the Massachusetts Dataset

On the Massachusetts dataset, RoadFormer was compared to FCN, U-Net, U-Net++ [19], PSPNet, DeepLabV3, Seg-Net, CADUNet [25], Batra et al., SGCN, and SwinUnet. Compared to the Deepglobe dataset, the Massachusetts dataset is more challenging due to (1) the image resolution of the Massachusetts road dataset having a lower resolution (1.5 m) than Deepglobe (0.5 m), which means that the road shape in the image will be slenderer and (2) the images in the training set are fewer, which makes difficulties for the model to converge. Moreover, to fully validate RoadFormer's ability to process large-size images, we directly feed the uncropped images to the network. Consistent with the experiments on the Deepglobe dataset, we visualize the extracted roads of FCN, U-Net, PSPNet, DeeplabV3, SwinUnet, and RoadFormer, and the quantitative results of the other five above approaches are quoted directly from the source.

Six images with different representative scenes were chosen from the test set for visualization comparison. Figure 9 shows the road results extracted by FCN, U-Net, PSPNet, DeeplabV3, and RoadFormer in six rows. Among the extraction results, RoadFormer has stronger adaptability for complex scenes. For example, roads are obscured and have low contrast with the surrounding environment in the mountain road scene (first row). In this case, the extraction results of CNN-based methods lose part of the road information, while Swin-Unet and RoadFormer could extract valid road structures. Suffering from the interference of dense roads in town scenes (second to fourth row), the other five methods cannot extract road features effectively, whereas our method can still extract clear road structures owing to the introduction of the Swin Transformer. For slender road scenes (the seventh and eighth rows), none of the other methods could extract road structures completely, while RoadFormer is able to extract these fine roads accurately. The visualization results in Figure 9 further demonstrate that RoadFormer has better adaptability than the other methods in complex scenes.

To further evaluate the model performance, quantitative comparison results between RoadFormer and the other nine methods are given in Table 4. We can observe from the table that the other methods achieve a recall rate lower than 75%, except for PSPNet, CADUNet, and RoadFormer. More importantly, other methods achieve IoU rates lower than 65% and F1-scores lower than 78%, except SGCN and RoadFormer. Obviously, recall, IoU, and F1-score performance are all best on RoadFormer. Among these SOTA methods, SGCN also uses the technique of feature separation. RoadFormer achieves better results owing to its ability to capture long-range information and a larger receptive field. Notably, our model uses the entire image as the input, while the other models use the cropped patches as the input. Thus, it is evident that RoadFormer is more capable of handling high-resolution images. In comparison to other approaches, RoadFormer consistently outperforms them in criteria such as recall, IoU, and F1-score, proving that it is extremely superior.

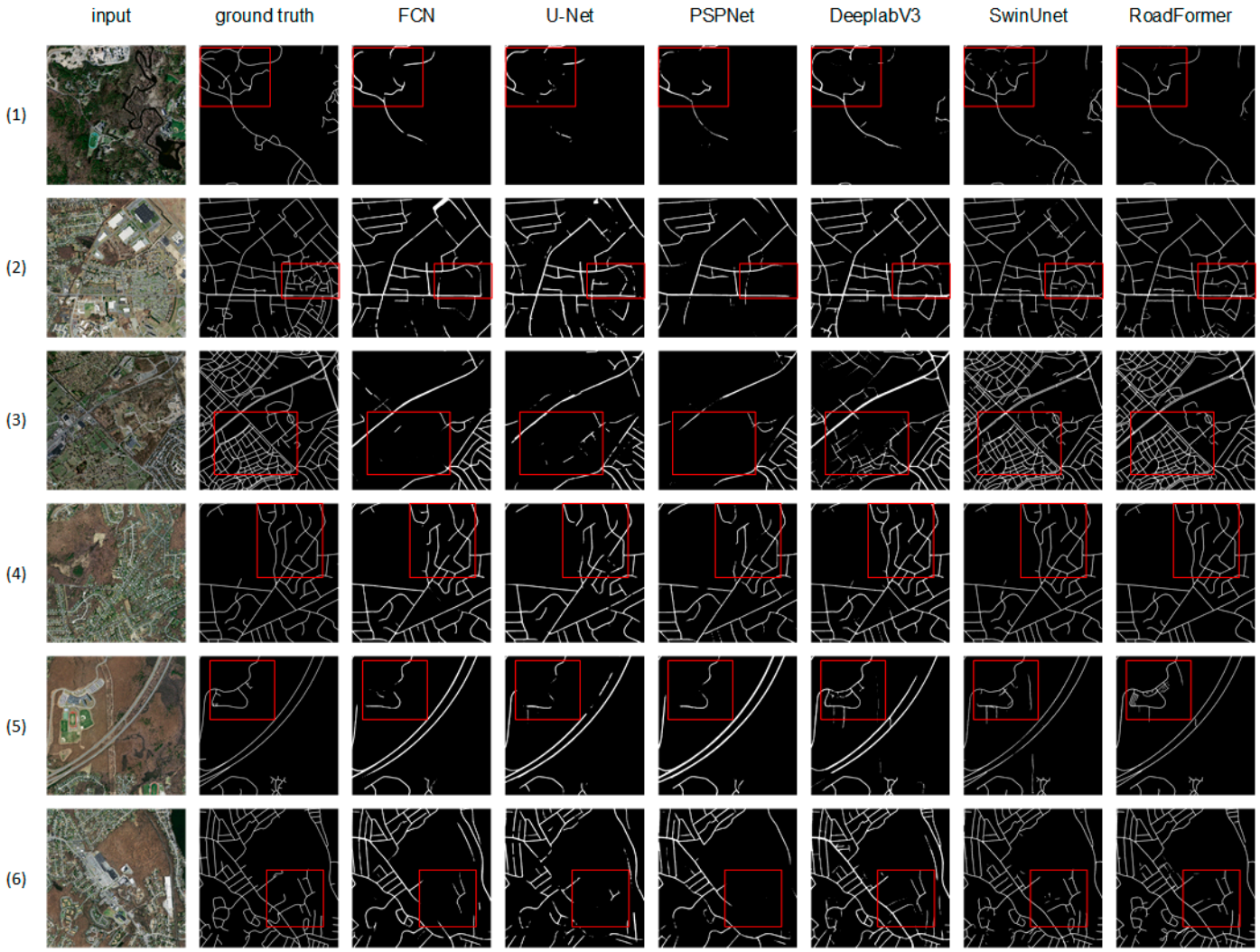

Figure 9. Visual comparison of RoadFormer and other SOTA methods for the Massachusetts road dataset. The red boxes highlight the regions where our method outperforms other methods. The size of the red boxes in each image from top to bottom are (**1**) 750 × 633, (**2**) 629 × 375, (**3**) 891 × 675, (**4**) 703 × 797, (**5**) 694 × 633, and (**6**) 609 × 609.

Table 4. Results of the Massachusetts road dataset's quantitative performance.

Methods	Precision	Recall	IoU	F1-Score	Param	FLOPs
FCN	82.8	68.1	59.7	74.7	47.1 M	197.7 G
U-Net	82.3	70.37	61.1	75.9	29.1 M	202.5 G
U-Net++	80.9	72.4	61.8	76.4	-	-
PSPNet	77.9	76.3	62.7	77.1	49.0 M	178.4 G
DeepLabV3	78.3	74.0	61.4	76.1	65.7 M	270.0 G
Seg-Net	82.5	72.1	62.5	76.9	-	-
CADUNet	79.5	76.6	64.1	77.9	-	-
Batra et al.	81.9	69.3	60.1	75.1	-	-
SGCN	**84.8**	73.9	65.3	79.0	-	-
SwinUnet	78.5	75.8	62.8	77.1	27.1 M	254.8 G
RoadFormer (Swin-B)	80.7	**77.6**	**65.5**	**79.2**	89.0 M	447.7 G

The visualization and quantitative comparison results above confirm that the proposed method has a higher capacity to extract roads. The visualization results intuitively show that RoadFormer is more adaptable to complex road and fine road scenarios. In Tables 3 and 4,

we can see that the RoadFormer with a tiny size has lower computational complexity and a smaller number of parameters, but still has good performance. Although RoadFormer with base size has higher complexity, it has the best performance. For 1024 × 1024 image input, the inference time of the base model on RTX3090 is less than 0.3 s per frame, which can meet the practical application requirements. The following three factors are mainly accountable for RoadFormer's superiority. (1) Swin Transformer modules help the model establish long-range dependence. (2) By feature separation module, a refined feature map allows the model to perceive different features separately. (3) The model's receptive field is further expanded by the dilated block that results from spatial convolution. Furthermore, according to the quantitative results, our model has the highest IoU and F1-score on the Deepglobe and Massachusetts datasets, further demonstrating the superiority of our approach.

5. Conclusions

For road extraction tasks, we present a novel model called RoadFormer that uses a Swin Transformer as its backbone. The spatial and channel separable convolution are combined in the design of RoadFormer to improve the feature representation of the model. In addition, a dilated block is adopted after the spatial convolution, which effectively helps the model capture better global contextual information and obtain larger receptive fields. Ablation experiments demonstrate the validity of our module design. The Deepglobe and Massachusetts datasets were used in experiments that were thoroughly assessed. The proposed method outperforms previous SOTA methods, as shown by the comparison of visualization and quantitative results, which supports the proposed model's superiority and effectiveness.

The proposed model was trained on RGB remote sensing image datasets (Deepglobe and Massachusetts). In practice, the accuracy of road extraction could be further improved by the fusion of multimodal data. Specifically, DEM information and geological background are very important, which makes it easier to extract road features in special scenarios. In addition, the multiple channels of information in satellite remote sensing and radar imagery can offer different information on the road. Moreover, the architecture of the Swin Transformer block could be further optimized and tweaked, and the different loss functions could be investigated for improving the model performance. In the future, we will collect multimodal remote sensing datasets mentioned above and further improve the model performance by optimizing the architecture and loss function.

Author Contributions: X.L. conceived of and designed the experiments and wrote the paper; Z.W. performed the experiments and original draft preparation; J.W. and J.Z. analyzed the data; Y.X. and R.L. provided the review and editing; Q.M. supervised the study and reviewed the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China: 2018YFC0807504, The Key R&D Projects of Qingdao Science and Technology Plan: 21-1-2-18-xx.

Acknowledgments: We sincerely thank the authors of FCN, U-Net, PSPNet, DeeplabV3, Seg-Net, LinkNet, D-LinkNet, HourGlass, CADUNet, SGCN, and [38] for providing their algorithm codes to facilitate the comparative experiments. We also thank the authors of the Deepglobe and Massachusetts datasets for providing remote sensing road images for the comparative experiments.

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

References

1. Wei, Y.; Zhang, K.; Ji, S. Simultaneous Road Surface and Centerline Extraction from Large-Scale Remote Sensing Images Using CNN-Based Segmentation and Tracing. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8919–8931. [CrossRef]
2. Yang, F.; Wang, H.; Jin, Z. A Fusion Network for Road Detection via Spatial Propagation and Spatial Transformation. *Pattern Recognit.* **2020**, *100*, 107141. [CrossRef]
3. Valero, S.; Chanussot, J.; Benediktsson, J.A.; Talbot, H.; Waske, B. Advanced Directional Mathematical Morphology for the Detection of the Road Network in Very High Resolution Remote Sensing Images. *Pattern Recognit. Lett.* **2010**, *31*, 1120–1127. [CrossRef]

4. Chaudhuri, D.; Kushwaha, N.K.; Samal, A. Semi-Automated Road Detection From High Resolution Satellite Images by Directional Morphological Enhancement and Segmentation Techniques. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2012**, *5*, 1538–1544. [CrossRef]
5. Bae, Y.; Lee, W.-H.; Choi, Y.; Jeon, Y.W.; Ra, J.B. Automatic Road Extraction From Remote Sensing Images Based on a Normalized Second Derivative Map. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1858–1862. [CrossRef]
6. Zhang, Z.; Liu, Q.; Wang, Y. Road Extraction by Deep Residual U-Net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
7. Patil, D.; Jadhav, S. Road Extraction Techniques from Remote Sensing Images: A Review. In *Innovative Data Communication Technologies and Application*; Raj, J.S., Iliyasu, A.M., Bestak, R., Baig, Z.A., Eds.; Springer: Singapore, 2021; pp. 663–677.
8. 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]
9. Mendes, C.C.T.; Frémont, V.; Wolf, D.F. Exploiting Fully Convolutional Neural Networks for Fast Road Detection. In Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 16–20 May 2016; pp. 3174–3179.
10. Alshehhi, R.; Marpu, P.R. Hierarchical Graph-Based Segmentation for Extracting Road Networks from High-Resolution Satellite Images. *ISPRS J. Photogramm. Remote Sens.* **2017**, *126*, 245–260. [CrossRef]
11. Costea, D.; Leordeanu, M. Aerial Image Geolocalization from Recognition and Matching of Roads and Intersections. *arXiv* **2016**, arXiv:1605.08323.
12. Bastani, F.; He, S.; Abbar, S.; Alizadeh, M.; Balakrishnan, H.; Chawla, S.; Madden, S.; DeWitt, D. RoadTracer: Automatic Extraction of Road Networks from Aerial Images. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4720–4728.
13. Liu, Y.; Yao, J.; Lu, X.; Xia, M.; Wang, X.; Liu, Y. RoadNet: Learning to Comprehensively Analyze Road Networks in Complex Urban Scenes from High-Resolution Remotely Sensed Images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 2043–2056. [CrossRef]
14. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16 $\times$ 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
15. Liu, X.; Gao, H.; Miao, Q.; Xi, Y.; Ai, Y.; Gao, D. MFST: Multi-Modal Feature Self-Adaptive Transformer for Infrared and Visible Image Fusion. *Remote Sens.* **2022**, *14*, 3233. [CrossRef]
16. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 10–17 October 2021; pp. 9992–10002.
17. 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.
18. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Munich, Germany, 5–9 October 2015; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
19. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. *CoRR* **2018**. Available online: https://link.springer.com/chapter/10.1007/978-3-030-00889-5_1 (accessed on 12 February 2023).
20. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6230–6239.
21. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018; Springer International Publishing: Cham, Switzerland, 2018; pp. 833–851.
22. 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]
23. 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; pp. 1–4.
24. Zhou, L.; Zhang, C.; Wu, M. D-LinkNet: LinkNet with Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 192–1924.
25. 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]
26. Newell, A.; Yang, K.; Deng, J. Stacked Hourglass Networks for Human Pose Estimation. In Proceedings of the Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; Springer International Publishing: Cham, Switzerland, 2016; pp. 483–499.
27. Batra, A.; Singh, S.; Pang, G.; Basu, S.; Jawahar, C.; Paluri, M. Improved Road Connectivity by Joint Learning of Orientation and Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.

28. Zhou, G.; Chen, W.; Gui, Q.; Li, X.; Wang, L. Split Depth-Wise Separable Graph-Convolution Network for Road Extraction in Complex Environments From High-Resolution Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5614115. [CrossRef]
29. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. *arXiv* **2021**, arXiv:2105.05537.
30. Liu, P.; Wang, Q.; Yang, G.; Li, L.; Zhang, H. Survey of Road Extraction Methods in Remote Sensing Images Based on Deep Learning. *PFG* **2022**, *90*, 135–159. [CrossRef]
31. Shao, Y.; Guo, B.; Hu, X.; Di, L. Application of a Fast Linear Feature Detector to Road Extraction From Remotely Sensed Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2011**, *4*, 626–631. [CrossRef]
32. Alshehhi, R.; Marpu, P.R.; Woon, W.L.; Mura, M.D. Simultaneous Extraction of Roads and Buildings in Remote Sensing Imagery with Convolutional Neural Networks. *ISPRS J. Photogramm. Remote Sens.* **2017**, *130*, 139–149. [CrossRef]
33. Cui, F.; Feng, R.; Wang, L.; Wei, L. Joint Superpixel Segmentation and Graph Convolutional Network Road Extration for High-Resolution Remote Sensing Imagery. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; pp. 2178–2181.
34. Kestur, R.; Farooq, S.; Abdal, R.; Qadri, E.; Narasipura, O.; Mudigere, M. UFCN: A Fully Convolutional Neural Network for Road Extraction in RGB Imagery Acquired by Remote Sensing from an Unmanned Aerial Vehicle. *J. Appl. Remote Sens.* **2018**, *12*, 016020. [CrossRef]
35. Varia, N.; Dokania, A.; Senthilnath, J. DeepExt: A Convolution Neural Network for Road Extraction Using RGB Images Captured by UAV. In Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence (SSCI), Bengaluru, India, 18–21 November 2018; pp. 1890–1895.
36. 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.
37. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected Crfs. *arXiv* **2014**, arXiv:1412.7062.
38. 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] [PubMed]
39. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
40. Park, N.; Kim, S. How Do Vision Transformers Work? *arXiv* **2022**, arXiv:2202.06709.
41. Tao, C.; Qi, J.; Li, Y.; Wang, H.; Li, H. Spatial Information Inference Net: Road Extraction Using Road-Specific Contextual Information. *ISPRS J. Photogramm. Remote Sens.* **2019**, *158*, 155–166. [CrossRef]
42. Sifre, L.; Mallat, S. Rigid-Motion Scattering for Texture Classification. *arXiv* **2014**, arXiv:1403.1687.
43. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv* **2015**, arXiv:1511.07122.
44. Chen, Z.; Deng, L.; Luo, Y.; Li, D.; Marcato Junior, J.; Nunes Gonçalves, W.; Awal Md Nurunnabi, A.; Li, J.; Wang, C.; Li, D. Road Extraction in Remote Sensing Data: A Survey. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102833. [CrossRef]