applied sciences

MDPI

Article

Dual Parallel Branch Fusion Network for Road Segmentation in High-Resolution Optical Remote Sensing Imagery

Lin Gao * and Chen Chen

School of Information Science and Engineering, Shenyang Ligong University, Shenyang 110159, China; chenc_1998@163.com

* Correspondence: gaolin324@sylu.edu.cn

Abstract: Road segmentation from high-resolution (HR) remote sensing images plays a core role in a wide range of applications. Due to the complex background of HR images, most of the current methods struggle to extract a road network correctly and completely. Furthermore, they suffer from either the loss of context information or high redundancy of details information. To alleviate these problems, we employ a dual branch dilated pyramid network (DPBFN), which enables dual-branch feature passing between two parallel paths when it is merged to a typical road extraction structure. A DPBFN consists of three parts: a residual multi-scaled dilated convolutional network branch, a transformer branch, and a fusion module. Constructing pyramid features through parallel multi-scale dilated convolution operations with multi-head attention block can enhance road features while suppressing redundant information. Both branches after fusing can solve shadow or vision occlusions and maintain the continuity of the road network, especially on a complex background. Experiments were carried out on three datasets of HR images to showcase the stable performance of the proposed method, and the results are compared with those of other methods. The OA in the three data sets of Massachusetts, Deep Globe, and GF-2 can reach more than 98.26%, 95.25%, and 95.66%, respectively, which has a significant improvement compared with the traditional CNN network. The results and explanation analysis via Grad-CAMs showcase the effective performance in accurately extracting road segments from a complex scene.

Keywords: remote sensing imagery; transformer mechanism; dilated convolution; road segmentation; explanation analysis

check for updates

Citation: Gao, L.; Chen, C. Dual Parallel Branch Fusion Network for Road Segmentation in High-Resolution Optical Remote Sensing Imagery. *Appl. Sci.* **2023**, *13*, 10726. https://doi.org/10.3390/app131910726

Academic Editor: Christos Bouras

Received: 22 August 2023
Revised: 15 September 2023
Accepted: 20 September 2023
Published: 27 September 2023

1. Introduction

Road network information serves as a vital data source in various applications, including geographic information system (GIS), unmanned vehicles, and urban planning [1]. The high spatial resolution of images not only provides clearer ground object information, but also introduces three challenges. (a) Road network structures exhibit complexity, appearing as heterogeneous areas with substantial intra-class variability and ambiguous inter-class distinctions. (b) Roads often appear slender and elongated in high-resolution imagery, spanning hundreds of pixels in remote sensing images, which can make their detection challenging. (c) High-resolution images accentuate noise factors such as vehicles and pedestrians on the road, making their presence more pronounced [1]. Consequently, the automatic detection of road network information using advanced algorithms over large areas is of significant interest to the research community.

Meanwhile, roads are categorized into urban roads and rural highways based on administrative regions, and their manifestation in remote sensing imagery differs accordingly. Urban roads exhibit complex road network structures, often affected by various noise interferences such as pedestrians and vehicles along the roadside. Additionally, the distinctions between roads and other land features tend to be relatively subtle. On the other hand, rural roads are characterized by their elongated shape, featuring unclear road boundaries and significant obstruction from roadside trees. Therefore, with increasing image resolution, the intra-class differences of roads become more prominent, posing a greater challenge for road interpretation tasks.

In recent years, the transformer mechanism has gained increasing popularity in the field of optical remote sensing imagery interpretation, owing to its demonstrated advantages in natural language processing tasks. The transformer mechanism achieves global context information by utilizing multi-head attention, enabling it to establish connections between different patches of the input sequence and capture comprehensive contextual details [2]. However, there are several challenges when directly applied to the aspect of road extraction from HR remote sensing images. The most significant problem arises from the transformer's limited ability to capture fine-grained local features, resulting in a lack of detailed information in the outputs and leading to blurry or less sharp features. Transformers are primarily designed to focus on global context and long-range dependencies, which may not be well-suited for capturing the intricate local details inherent to HR remote sensing images. As a remedy, the integration of local feature pyramids and global context information becomes indispensable to boost the performance of road segmentation from HR remote sensing images.

In this part, we introduce a dual parallel branch fusion network (DPBFN) to address the challenges of broken and incomplete road segmentation results. The DPBFN tackles issues related to shadows and occlusions by employing a dual branch network architecture. The network comprises a transformer branch and a residual multi-scaled dilated convolution network branch, fused together through a fusion module. The contributions can be summarized as follows:

(1) We employed a dual parallel branch architecture to synthesize the features capturing from depth convolutional and transformer branches, which is able to maintain more unique characteristics of the road network. Moreover, dual branches features as a complementary can discriminate the road segments by the spatial and spectral dependency in the feature space for better inferring the occluded roads.
(2) We enhanced the residual encoder branch by incorporating parallel multi-scaled dilated convolution operations with an attention block, named as the dilated convolution pyramid with attention (DCPA) module. The DCPA module helps the network to capture localized road edges features more effectively.
(3) Extensive experiments were carried out on three road challenge datasets; the results show the efficacy and powerful generalization capacity of the proposed method. In addition, we provide a comprehensive explanation analysis using gradient class activation maps.

The experimental results showcase the effectiveness of the proposed method in handling challenging road extraction scenarios, including shadows and occlusions. The DPBFN achieves superior performance compared to existing methods, making it a promising solution for road segmentation in HR remote sensing imagery.

2. Related Work

In this section, we present a comprehensive review of the existing road extraction works that harness the deep learning techniques. We specifically focus on three categories: CNN-based road methods, transformer-based [3] road methods, and hybrid CNN-Transformer methods. These approaches have shown some promising results in road extraction tasks by leveraging the power of deep-learning techniques. By analyzing and comparing these methods, we aim to offer enhanced and valuable insights into the advancements and prospects of road extraction techniques based on deep learning.

2.1. CNN-Based Road Extraction Methods

CNN-based methods have been widely adopted in road extraction tasks because of their ability to learn hierarchical features from input images. These methods typically utilize convolutional neural networks to extract the discriminative features, followed by classification or segmentation modules for road identification. CNNs inherently possess translation invariance, meaning they can recognize road patterns regardless of their location in the image. This property is particularly advantageous for road extraction, where road layouts may vary in different parts of the image. Several classic CNN architectures, such as U-Net [4] SegNet [5], and DeepLab [6], have been successfully applied to road extraction tasks, and these methods have proven their effectiveness in accurately capturing road patterns and structures. In recent years, CNN models such as DDU-Net [7], SDUNet [8], and L-DeepLabv3+ [9] have also shown very good results in road extraction. And some networks like MECA-Net [10], ASPP-U-Net [11], and MSMT-RE [12] with multi-scale features also have good performance in road extraction. However, most of these CNN-based methods only focus on the multi-scale encoder structure or multiple branches of the neural network, while ignoring some inherent characteristics of the road surface. And compared with a transformer, CNN (convolution neural network) has some unique advantages: local perception, parameter sharing, and translation invariance. Some methods include DeepWindow [13], FuNet [14], and HsgNet [15]. And some networks like CADUNet [16], VNet [17], DA-RoadNet [18], and the literature [19] have proved that the integration of an attention mechanism in CNN architecture also works in road extraction tasks.

2.2. Transformer-Based Road Extraction Methods

Transformer-based methods, originally introduced for natural language processing tasks, have recently gained attention in remote sensing imagery interpretation tasks, including road extraction. Transformers can capture long-distance dependencies in images, thus effectively modeling global context information. This enables the model to better understand the location and shape of the road in the whole image and improve the quality of road extraction. In the field of road extraction, transformer-based models leverage self-attention mechanisms to capture contextual information and generate accurate road predictions. Examples of transformer-based architectures applied to road extraction include Vision Transformer (ViT) [20], Swin Transformer [21], Swin-UNet [22], BDTNet [23], and RoadFormer [24]. These models have shown promising results in capturing global context and improving the accuracy of road extraction. However, due to the large number of parameters in Vision Transformer, the generalization ability of the model may be limited for small data sets, and it is easy to over-fit.

2.3. Hybrid CNN-Transformer Methods

To take advantage of the complementary advantages of CNN and transformers, a hybrid method for road extraction has been proposed. These methods combine the depth feature extraction ability of the neural network and the context information modeling from transformers. By integrating the two discriminative features, the hybrid model can capture fine-grained details and long-range dependencies, thus improving the performance of road extraction tasks. In recent years, many universal models for semantic segmentation have been proposed, such as TransFuse [25], TransUnet [26], Ds-TransUnet [27], and SegFormer [28], which have good results in the related fields. And some noteworthy hybrid CNN–transformer models for road extraction include TransLinkNet [29], Seg-Road [30], and DCS-TransUperNet [31]. In the above models, local and global features can be adaptively integrated through specific strategies and attention mechanisms, so that the model can better capture the structure and context information of the road.

In summary, different approaches to road extraction have shown their strengths in handling specific aspects of the task. CNN-based road methods have excelled in capturing local features and spatial patterns, allowing them to accurately delineate road structures at a fine-gained level. On the other hand, transformer-based road methods have demonstrated the ability to model global context and long-dependencies, enabling them to understand the overall road layout and context in the entire image. Hybrid CNN-transformer methods have shown enhanced performance by leveraging the strengths of both approaches. The next section will delve into the details of the hybrid CNN–transformer mechanism and provide an in-depth analysis of its advantages in road extraction tasks.

3. Materials and Methods

We introduced a spatial information inference structure aimed at effectively modeling road-specific contextual information at length. The multi-scaled feature inference structure was designed to exploit the inherent spatial dependencies present in road networks, thereby providing a more comprehensive understanding of road patterns and structures from the images.

3.1. Dual Parallel Branch Network Structure

The dual parallel network architecture comprised a CNN branch, a transformer branch, and a fusion module. Figure 1 provides an overview of the entire structure. The refined CNN branch was specifically designed to capture multi-scale features by integrating various scales of parallel dilated spatial pyramid (PDSP) modules and a convolution block attention module. The transformer branch was focused on modeling extensive relationships and capturing the global context information, allowing for the effective handling of long-range dependencies. Finally, the two paralleled branches were fused using the BiFusion module referred to in [25]. The fusion module integrated the complementary information extracted by the CNN and transformer branches, enabling the network to leverage both local and global contextual cues for road extraction tasks.

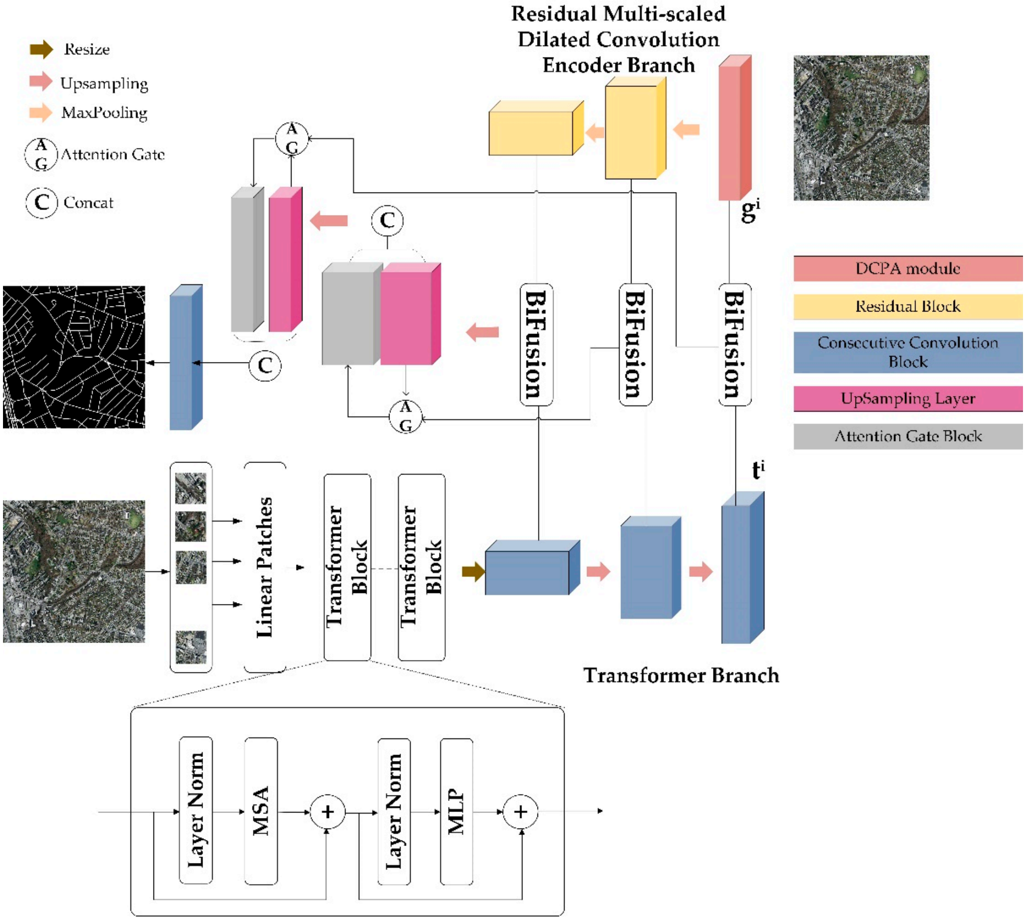


Figure 1. The overall the network structure.

3.2. Transformer Encoder Branch

As inspired by [25], the architecture of the transformer branch adhered to the conventional encoder–decoder design. Initially, the input image is uniformly partitioned into $N = \frac{H}{S} \times \frac{W}{S}, S = 16$. Those patches were subsequently flattened and forwarded through a linear embedding layer, yielding an output dimension of D_0, resulting in the original embedding sequence $e \in \mathbb{R}^{N \times D_0}$. To incorporate the spatial information, a tunable positional embedding of the same dimension was added to the raw embedding e. These embeddings $z^0 \in R^{N \times D_0}$ were fed into the transformer branch, which consisted of L layers of multi headed self-attention (MSA) and multi-layer perception (MLP). The self-attention (SA) mechanism, a key principle of transformer, was essential in updating the states of each embedded patch during every layer of the transformer encoder. SA enabled the model to aggregate information globally by considering the relationships between all patches in the input image. Unlike traditional CNNs that have local receptive fields, the SA mechanism allowed for the network to capture long-range dependencies and contextual information from the whole image.

$$SA(z_i) = softmax(\frac{q_i k^T}{\sqrt{D_h}})v \tag{1}$$

where $[q, k, v] = zW_{qkv}, W_{qkv} \in R^{D_0 \times 3D_h}$ is the projection matrix and vector $z_i \in R^{1 \times D_h}$ are the i^{th} row of z, respectively.

MSA is an extension of SA that merges multiple SA operations and projects the latent dimension back to R^{D_0}. This enabled the model to capture diverse and complementary patterns and relationships within the input. Additionally, the MLP is a stack of dense layers [20] that processes the outputs of the MSA, further refining and transforming the extracted features. As the global context features can be fused with the corresponding feature maps of convolution encoder branch, we used the progressive upsampling method referred to as SETR [32].

3.3. Residual Multi-Scaled Dilated Convolution Branch

The residual multi-scaled dilated convolution branch employed the dilated convolution pyramid with attention (DCPA) module to capture localized road features effectively. The DCPA module comprised four parallel dilated convolution blocks (PDCB), a global average pooling layer, a channel attention block, and a spatial attention block. Due to the sparsity of HR remote sensing images, incorporating pyramid features without reducing resolution can enhance the model's ability to identify objects and improve overall performance.

The DCPA module, inspired by [11,33], is detailed in Table 1 and visualized in Figure 2. The module comprised the parallel dilated convolution blocks (PDCB), which consisted of multi-scale parallel convolution layers with different dilation rates, along with a global pooling layer. The dilated rates used in four parallel paths were 1, 6, 12, 18, respectively, aiming to capture pyramid features and context information of road pixels. The output of the parallel paths was then concatenated and processed by a 1 × 1 convolution layer to produce the final output. The dilated convolution layers control the size of the effective receptive field of the convoluted kernels and dilated rate. A larger dilated rate corresponds to a larger scale without reducing feature resolution. Furthermore, the global pooling layer, which operates at the image level, averaged the feature maps across the spatial dimensions to improve the model's ability to recognize global context and capture long-range dependencies. Furthermore, each path in the DCPA module consisted of a convolution layer with a specific dilated rate, followed by the operations of batch normalization (BN) and activation (ReLU).

Table 1. The settings of the PDSA module.

Items	Layer	Kernal Size	Dilated Rate
PDCB-1	ReLU(BN(Conv))	1 × 1	1
PDCB-2	ReLU(BN(Conv))	3 × 3	6
PDCB-3	ReLU(BN(Conv))	3 × 3	12
PDCB-4	ReLU(BN(Conv))	3 × 3	18
AvgPooling	ReLU(BN(Conv))	1 × 1	-
CA	Pooling-MLP-ReLU	-	-
SA	Pooling-Conv	7 × 7	-

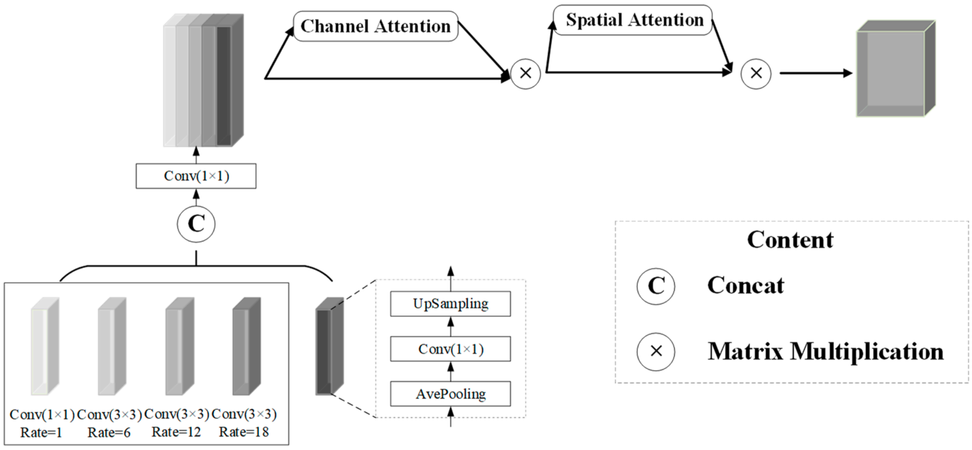


Figure 2. The structure of the DCPA module.

The spatial attention component learned to selectively emphasize informative spatial locations by computing an SA map. The map was computed by applying a convolution layer with a 7 × 7 kernel and a stride set at 1, followed by a ReLU layer and another convolution layer with a 1 × 1 kernel size and a stride of 1. The resulting spatial attention output was then multiplied with the refined feature maps obtained from the channel attention component. The final output of the DCPA module was acquired by concatenating the refined feature maps with the spatial attention map. The DCPA modules are expressed as follows:

Dilated convolution blocks:

$$f_{DCP}(X) = \sum_{path} conv(X, dilated\_rate_{path}) \tag{2}$$

Channel attention blocks:

$$f_{channel}(X) = \sigma(MLP(AvePool(X))) \times X \tag{3}$$

where X is the input tensor, AvgPool represents the global average pooling operation, MLP is a two-layer fully connected network with ReLU activation layers correspondingly, and σ utilizes the sigmoid function.

Spatial attention blocks:

$$f_{spatial}(X) = \sigma(conv_1(\mathrm{ReLU}(conv_2(X)))) \times X \tag{4}$$

where conv1 and conv2 are convolutional layers with kernel sizes of 7 × 7 and 1 × 1, respectively, and σ is also the Sigmoid layer. The final output of the DCPA module is obtained by merging the refined feature maps with the SA map:

$$f_{DCPA}(X) = f_{channel}(f_{spatial}(f_{DCP}(X))) \tag{5}$$

4. Results

4.1. Datasets and Processing

The validation experiments utilized three datasets: the Massachusetts road dataset [34], the Deep Globe road extraction sub-challenge dataset [35], and the GF-2 dataset [36]. Within the three datasets, the Massachusetts road dataset consists of 1171 images. The size of each image is 1500 × 1500 pixels, and its resolution is 1.2m. The Deep Globe dataset comprises 1113 satellite images accompanied by corresponding masks delineating road labels. The dimension of the raw images is 1024 × 1024 × 3 pixels, and the resolution is 0.5 m. The GF-2 road dataset contains 200 images, obtained from GF-2 satellite, with a size of 1500 × 1500 pixels. Each of the three datasets is divided into three parts: training set, validation set, and test set. The details are described in Table 2.

Table 2. Datasets Allocation Details.

Dataset	Training Set	Validation Set	Test Set
Deep Globe	780	111	222
Massachusetts	1108	14	49
GF-2	170	10	20

To maximize the effective utilization of the constrained training set, we incorporate geometric transformation techniques, encompassing random clipping as well as horizontal and vertical flip transformations. In the context of the trainable road datasets, the predominant contributor to the loss value stems from the misclassification of road pixels as background pixels. This phenomenon arises due to the inherent disparity in quantity between background and road pixels within the HR remote sensing images. Therefore, while optimization endeavors may lead to loss reduction, the fine-tuned semantic segmentation networks are susceptible to classifying uncertain pixels as background rather than road. To mitigate this challenge, we address the matter by a rebalancing strategy, where we adjust the distribution of road and non-road samples to counteract the issue.

4.2. Implementation Details

The training sets obtained by data augmentation were preprocessed by a series of common data enhancement methods. For the purpose of spectral augmentation, we performed random cropping of both the image and its corresponding mask, altering size and aspect ratio. This approach was adopted to enhance the diversity of the dataset. Subsequently, due to constraints posed by GPU memory, all the images were resized to dimensions of 256 × 256 × 3 prior to being inputted into the networks. Afterwards, all the compared models were trained with the same training environment. Specifically, the networks were optimized by the Adam algorithm [37], whose moments of 0.9 and 0.999 correspond to the two parameters, in Windows 10 with one GTX3060 (memory 12GB) that allows for a batch size of 2 images. All experiments were performed using PyTorch 1.11.0.

The learning rate was initially set to be 0.0001 and reduced by a factor of 0.1 every 30 epochs. Since the raw size datasets were divided into many patches, the model needed a significantly higher number of iterations within each epoch across those three datasets. Moreover, we used the fine-tuned strategy in our network, which can accelerate the convergence. Training is conducted by optimizing structure loss function employing the Adam algorithm. Therefore, the last part of the proposed network utilized Sigmoid activation to output the results. Then, structure loss with binary cross entropy loss and the weighted IoU loss function [17,25] can be described as:

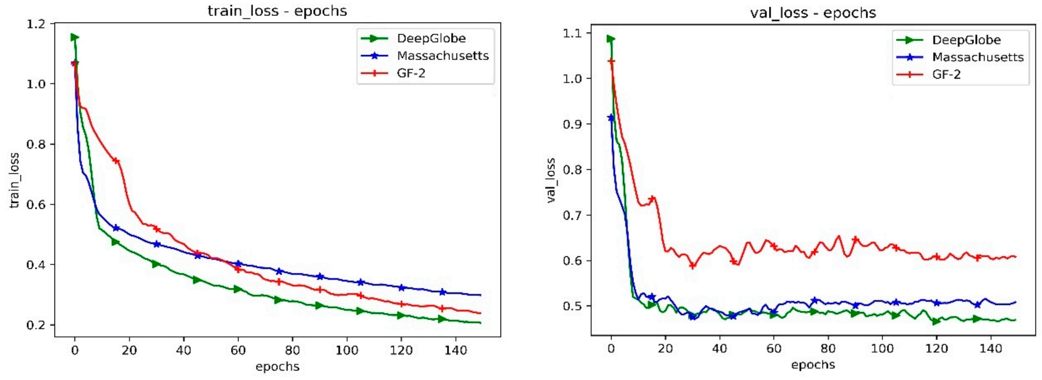
$$StructureLoss = BCELoss + WeightedIoULoss \tag{6}$$

$$BCELoss(p, m) = -{}^{1}\!/_{N}\sum_{i=1}^{N}\left[m_i \cdot \log(p_i) + (1-m)\log(1-p_i)\right] \tag{7}$$

$$WeightedIoULoss(p,m) = 1 - \frac{\sum_{i=1}^{N}(p_i \times m_i \times w_i) + \varepsilon}{\sum_{i=1}^{N}(p_i + m_i + p_i \times m_i) \times w_i + \varepsilon} \tag{8}$$

In Equations (7) and (8), N represents the number of samples. m represents the label of the sample index ith, usually 0 (negative category) or 1 (positive category). p represents the model predicted value of sample i, usually between 0 and 1, indicating the model's probability estimate that the sample belongs to a positive category.

Figure 3 shows the convergence tendency of loss function values in the process of the training. Evidently, the error exhibits a gradual reduction over time. In Figure 3b, the validation loss value becomes stable after the 30 epochs among the three datasets.



(**a**) Training Loss Curve (**b**) Validation Loss Curve

Figure 3. The loss curves of DPBFN training process from three datasets.

4.3. Evaluation Metrics

The road networks are commonly evaluated by complexness and correctness [19]. Because the metrics mentioned above can be approximated as precision and recall, the extracted road results are evaluated by the overall accuracy (OA), precision (P), recall (R), F_1, and intersection-over-union (IoU). The F1 score and IoU serve as two comprehensive evaluation metrics. These metrics [34] can be computed as follows:

$$OA = (TP + FN)/(TP + TN + FP + TN) \tag{9}$$

$$P = {}^{TP}/_{(TP+FP)} \tag{10}$$

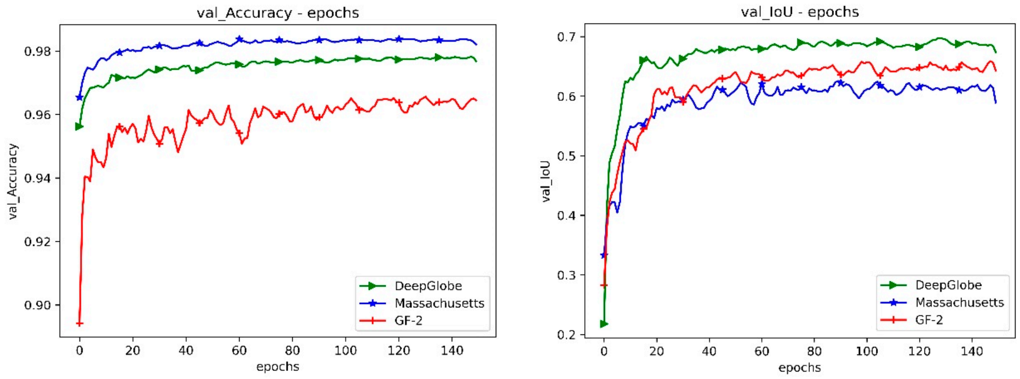
$$R = {}^{TP}/_{(TP+FN)} \tag{11}$$

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{12}$$

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

where TPs (true positives) and TNs (false positives) are the number of road pixels and non-road pixels, respectively, which are correctly classified. FPs (false positives) represent the number of background pixels that are identified as road. FNs (false negatives) are the number of road pixels that are identified as background. Meanwhile, those above evaluated metrics are powerful for the harmonic means.

Figure 4 exhibits the accuracy and IoU curves of the DPBFN from three validation sets. From the leftmost graph, it can be observed that the accuracy of all three datasets start stabilizing after 20 epochs. The Massachusetts dataset achieves the highest accuracy. The accuracy of the Massachusetts and Deep Globe datasets remains relatively stable, while that the GF-2 dataset exhibits more significant fluctuations. However, as shown in Figure 4b, the Deep Globe dataset achieves the best IoU. The IoU curves of the Massachusetts and GF-2 datasets exhibit a phenomenon of crossing during the convergence process, especially in the beginning of 10 epochs. Moreover, even after reaching a stable state, the IoU value of GF-2 is higher than that of the Massachusetts dataset.

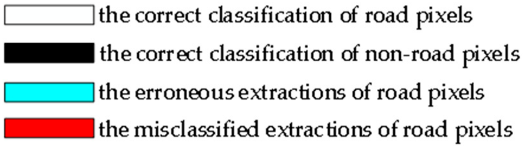


(**a**) Validation accuracy curve (**b**) Validation IoU curve

Figure 4. Validation accuracy and IoU the curves of DPBFN from three datasets.

4.4. Results and Comparison

We mainly validate the performance of our network with other different road extraction methods through the three challenging datasets mentioned above. These datasets are optical remote sensing imagery datasets. To demonstrate the advantages of the DPBFN, we compare our method with some classic methods selected from CNN-based, transformer-based, and hybrid based. To facilitate visual interpretation and the analysis of extraction outcomes yielded by diverse algorithms, accurately classified road segments are represented in white, while accurately classified non-road pixels are depicted in black. In cases of erroneous extractions and misclassified pixels, these specific segments will be emphasized by using red and blue shades, as illustrated in Figure 5.



□ the correct classification of road pixels
■ the correct classification of non-road pixels
🟦 the erroneous extractions of road pixels
🟥 the misclassified extractions of road pixels

Figure 5. The legends for results outputting.

4.4.1. Massachusetts Data

Figure 6 showcases two instances derived from the Massachusetts dataset, following a comparison involving various methodologies. Notably, the proposed approach demonstrates enhanced performance in terms of finer details, as indicated in Figure 6. The corresponding metrics are detailed in Table 2. For ease of observation and analysis, a specific region within the yellow frame was chosen from the image. The extraction outcomes pertaining to this selected area are visualized in Figure 7.

As exhibited in Table 3, the OA of the DPBFN is significantly higher than that of other models using CNN and integrated transformer architecture. The F1 score and IoU are improved, with an average of 99.22% and 98.35%, respectively. The additional DCPA module can obviously reinforce the performance of decapitation extraction and reduce the misclassification of pixels. However, in the part of depth feature extraction, more scale features can be extracted through the DCPA module, and these improvements help to extract roads (by using the fused features) of different sizes.

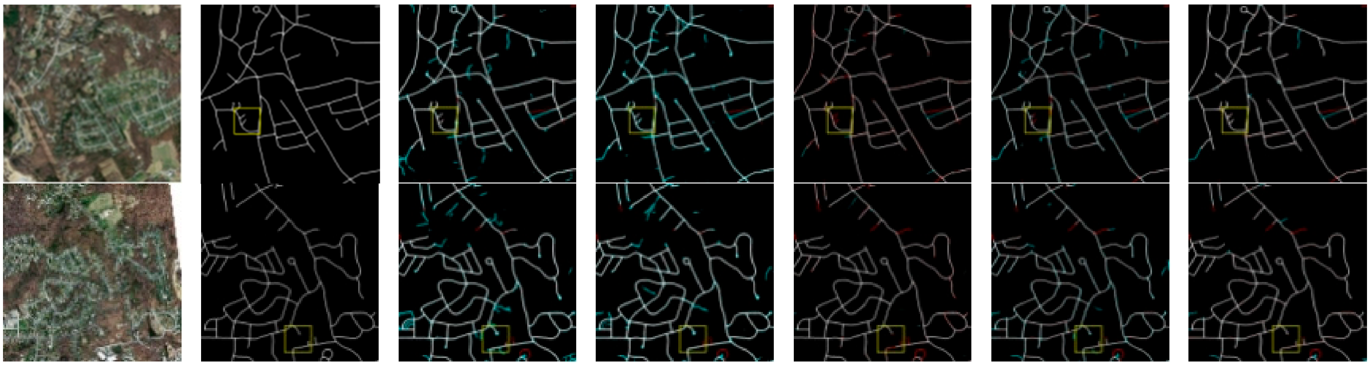

(**a**) Image (**b**) Label (**c**) U-Net (**d**) Deeplabv3+ (**e**) Swin-UNet (**f**) TransFuse (**g**) DPBFN

Figure 6. Different results outputting of five methods from Massachusetts dataset.

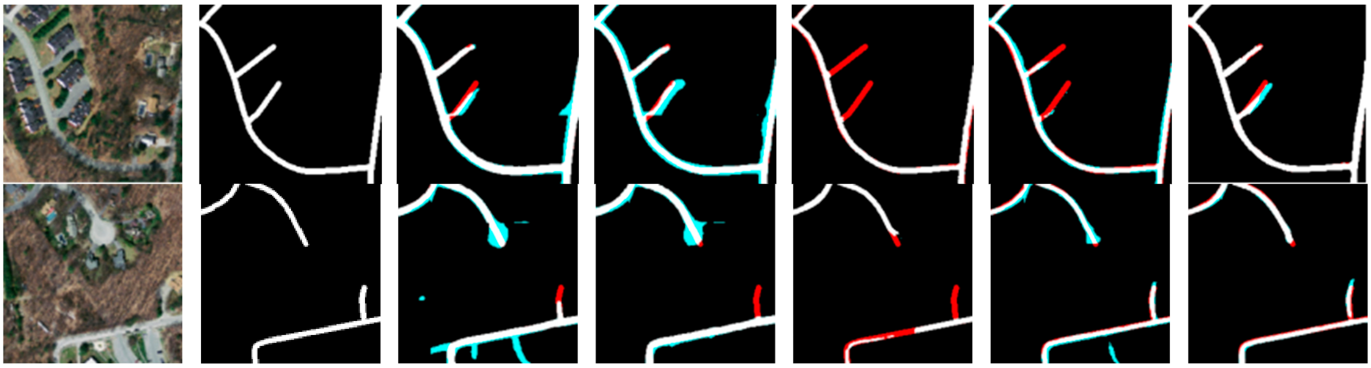

(**a**) Image (**b**) Label (**c**) U-Net (**d**) Deeplabv3+ (**e**) Swin-UNet (**f**) TransFuse (**g**) DPBFN

Figure 7. Local amplification results for Figure 6.

Table 3. Quantitative evaluation results in Figure 6.

Methods	**Image 1**					**Image 2**				
	OA (%)	**P (%)**	**R (%)**	**F1 (%)**	**IoU (%)**	**OA (%)**	**P (%)**	**R (%)**	**F1 (%)**	**IoU (%)**
U-Net	96.85	96.95	99.15	98.04	96.72	96.72	96.77	98.78	97.76	96.57
DeepLabv3+	97.06	97.13	99.79	98.44	96.93	96.88	96.93	99.79	98.34	96.74
Swin-UNet	98.26	99.53	98.56	99.04	98.21	98.35	99.48	98.51	99.00	98.30
TransFuse	98.17	98.81	99.28	99.04	98.10	98.32	98.88	99.35	99.11	98.25
DPBFN	98.35	99.28	99.20	99.24	98.29	98.47	99.26	99.13	99.20	98.41

Table 4 outlines the performance evaluation of the Massachusetts test dataset using metrics such as OA, P, R, F1 score, and IoU at breakeven points. Notably, the primary distinctions between the DPBFN and alternative approaches encompass factors such as kernel filter dimensions, the count of convolution layers, and the depth of the network structure. The DPBFN can achieve good performance via the DCPA module, which is an effective way to filter the validation information of road characteristics. The parallel multi-scaled dilated convolution operators can not only enlarge the receptive field without losing the feature resolution, but also build the feature pyramid to capture multi-scaled context information. The attention block can widen the network configuration with channel and spatial attention computed by maxpooling, averagepooling, and MLP, which confirms the theory reported in [35]. A statistical examination of Table 3 reveals that the classification accuracy achieved by the DPBFN surpasses that of U-Net, DeeplabV3+, RoadFormer, Swin-UNet, and TransFuse. Among these methods, the proposed approach demonstrates the highest classification accuracy, boasting an IoU of 66.78%, while U-Net registers the last classification rank with an IoU of 59.57%. As demonstrated in this part, we employ comprehensive statistics from the confusion matrix to elucidate the classification accuracy of these five methods across various experimental regions.

Table 4. Various compared results on the Massachusetts road test set.

Dataset	Methods	OA (%)	P (%)	R (%)	F1 (%)	IOU (%)
Massachusetts	U-Net	97.26	76.91	74.00	74.66	59.57
	DeepLabv3+	97.34	80.01	70.46	74.94	59.92
	RoadFormer [23]	-	80.70	77.60	79.2	65.50
	Swin-UNet	98.16	79.03	76.84	77.92	65.48
	TransFuse	98.09	76.46	79.38	77.89	65.76
	DPBFN	98.26	80.79	79.38	80.08	66.78

4.4.2. Deep Globe Data

Road extraction in Deep Globe data is obviously challenging due to the large intra-class variances and small inter-class distinctions. The DPBFN is effectively applied, as illustrated in Figure 8, with focused details in Figure 9 using yellow rectangles. Two images with ambiguous road-class distinctions were deliberately chosen. The test images, found in both figures and maps, predominantly encompass rural terrains where the roads are narrow, and their texture resembles that of cultivated land.

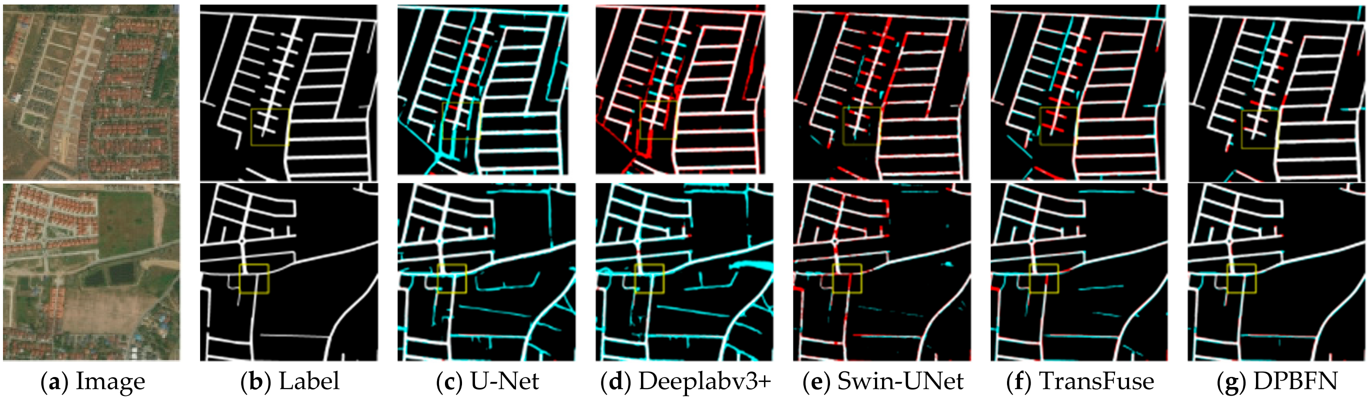

(**a**) Image (**b**) Label (**c**) U-Net (**d**) Deeplabv3+ (**e**) Swin-UNet (**f**) TransFuse (**g**) DPBFN

Figure 8. Different results outputting of five methods from the Deep Globe dataset.

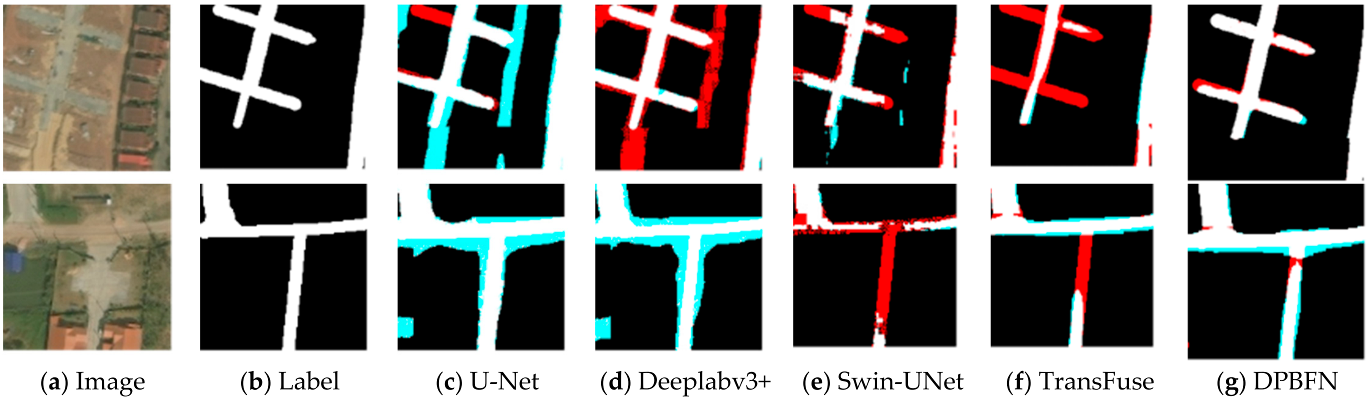

(**a**) Image (**b**) Label (**c**) U-Net (**d**) Deeplabv3+ (**e**) Swin-UNet (**f**) TransFuse (**g**) DPBFN

Figure 9. Local amplification results for Figure 8.

The statistical analysis of Table 5 shows that the overall accuracy of the DPBFN is higher than that of the other methods. Our method shows that the classification accuracy of the two shown images is the highest, with an average OA of 96.57% and an average IoU of 96.11%. U-Net holds the last classification rank, with an average OA of 92.62% and an average IoU of 91.52%. The rest of the network performance is between the two, the second models are TransFuse and Swin-UNet, the average OA is 95.89% and 95.33%, respectively, and the average IOU is 95.39% and 95.35%, respectively. Although the DPBFN has achieved better results, the fracture and mis-extraction of rural road extraction results will be more obvious than in urban roads.

Table 5. Quantitative evaluation results in Figure 8.

Methods	Image 1					Image 2				
	OA (%)	**P (%)**	**R (%)**	**F1 (%)**	**IOU (%)**	**OA (%)**	**P (%)**	**R (%)**	**F1 (%)**	**IOU (%)**
U-Net	91.65	90.95	99.11	94.85	90.21	93.59	93.00	99.91	96.33	92.92
DeepLabv3+	92.19	91.48	99.24	95.21	90.85	92.42	91.74	99.87	95.64	91.64
Swin-UNet	94.74	97.39	94.87	96.11	94.27	95.92	97.93	97.75	97.83	96.43
TransFuse	94.88	97.69	96.30	96.99	94.16	96.90	97.68	98.88	98.27	96.61
DPBFN	95.76	97.46	97.51	97.49	95.10	97.38	97.71	99.38	98.54	97.12

In the Table 6, it can be seen more clearly that the DPBFN has better extraction accuracy when facing the bifurcation road, and the phenomenon of mis-extraction and fracture of the bifurcation road is clearly less than that of other network models. And in the figure-the overall extraction comparison chart, the extraction accuracy of the DPBFN is also higher, and the data performance in the table is also better. The OA and IOU reach their highest at 95.25% and 72.08%, respectively, which is significantly improved compared with other models.

Table 6. Various compared results in the Deep Globe Road test dataset.

Dataset	Methods	OA (%)	P (%)	R (%)	F1 (%)	IOU (%)
Deep Globe	U-Net	89.95	65.26	87.78	74.16	60.73
	RoadFormer [23]	-	85.80	83.2	84.50	73.10
	DeepLabv3+	90.46	66.40	81.30	72.82	61.65
	Swin-UNet	93.07	81.67	71.58	76.29	61.67
	TransFuse	94.37	86.42	79.69	82.92	70.83
	DPBFN	95.25	86.27	86.56	86.42	72.08

4.4.3. GF-2 Data

We employed the DPBFN on GF-2 data, as illustrated in Figure 10, with the yellow rectangles strategically, as depicted in Figure 11. Two distinct images were chosen, each from different settings—rural and urban areas. The initial row of Figure 10 features a test image encompassing rural regions, where the roads are narrow and share a texture reminiscent of cultivated land. Moving to the second row in Figure 10, and further highlighted in the zoomed-in depiction in Figure 11, the image covers an urban expanse.

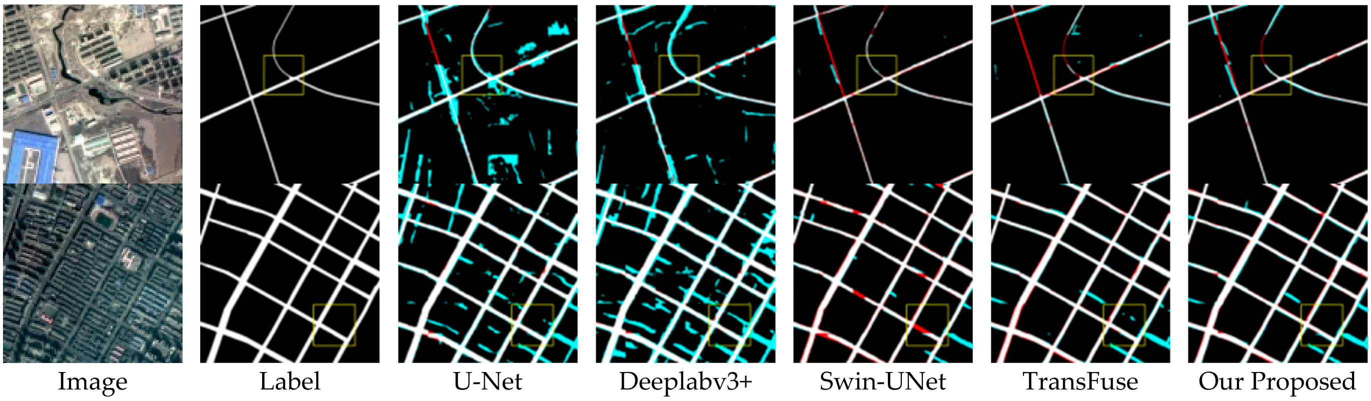


Figure 10. Different results outputting of five methods from the GF-2 dataset.

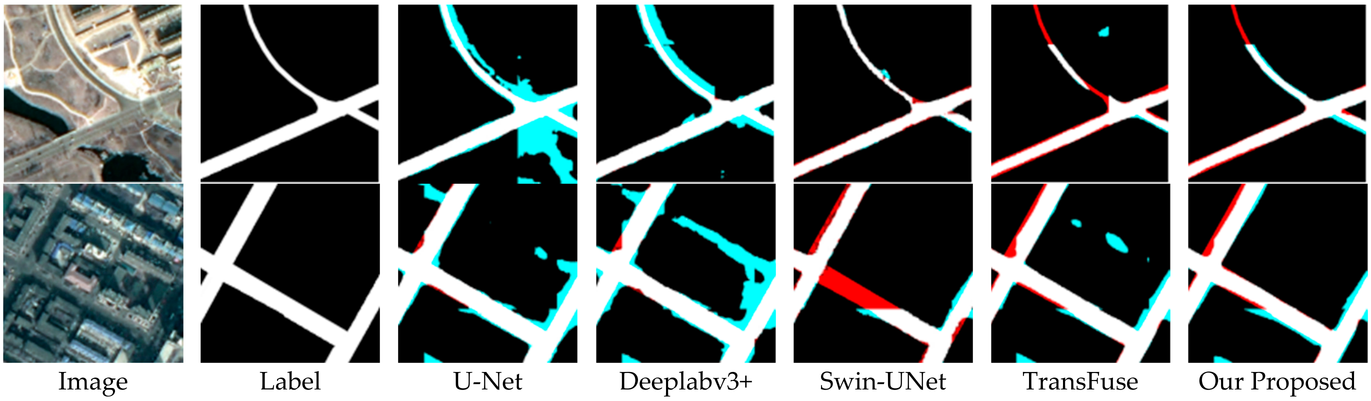


Figure 11. Local amplification results for Figure 10.

The presence of HR remote sensing images' noise, including cars, trees, and buildings' shadow, significantly hampers the extraction results and poses a major challenge in the aspect of road segmentation. Table 7 shows the comparison of the road extraction accuracy with various methods in the same experimental settings.

Table 7. Quantitative evaluation results in Figure 10.

Methods	**Image 1**					**Image 2**				
	OA (%)	**P (%)**	**R (%)**	**F1 (%)**	**IoU (%)**	**OA (%)**	**P (%)**	**R (%)**	**F1 (%)**	**IoU (%)**
U-Net	93.08	93.11	99.66	96.28	92.83	94.32	93.31	99.67	96.39	93.03
DeepLabv3+	95.77	96.01	99.57	97.76	95.62	89.45	87.09	99.82	93.02	86.96
Swin-UNet	98.64	99.52	99.22	99.36	98.61	96.71	98.48	97.65	98.05	96.11
TransFuse	98.55	99.40	99.09	99.23	98.50	96.67	97.33	98.57	97.94	96.01
DPBFN	98.50	99.16	99.25	99.21	98.42	96.60	97.10	98.65	97.87	95.87

In the GF-2 dataset, the DPBFN was compared with U-Net, DeeplabV3+, Swin-UNet, and TransFuse. Among the aforementioned methods, U-Net and DeeplabV3+ are representatives of the classic convolution neural network methods. SwinUnet and TransFuse are typical transformer-based and hybrid-based models, respectively, originally used for medical image segmentation. We present the visualization outcomes generated by the DPBFN and the aforementioned representative methods. Pertaining to the comprehensive accuracy analysis encompassing all images within the test set, Table 8 displays the metrics computed from the test set at the breakeven point. We present the visualization outcomes generated by the DPBFN and the aforementioned representative methods. Pertaining to the comprehensive accuracy analysis encompassing all images within the test set, Table 8 displays the metrics computed from the test set at the breakeven point. The hybrid-based method ranks first (with a recall of 65.05%), followed by Swin-UNet (with a recall of 63.23%), Deeplabv3+ (with a recall of 78.38%), and U-Net (with a recall of 75.11%). However, in the GF-2 dataset, the advantages of the DPBFN are not obvious with TransFuse in the respect of other metrics. Although this approach represents a significant improvement compared to other methods, it does not confer an advantage over Transfuse. Further investigation is needed to elucidate the underlying reasons for the phenomenon.

Table 8. Various compared results on the GF-2 road test dataset.

Dataset	**Methods**	**OA (%)**	**P (%)**	**R (%)**	**F1 (%)**	**IoU (%)**
GF-2	U-Net	90.52	56.52	75.11	64.50	56.24
	DeepLabv3+	90.69	54.87	78.38	64.55	58.16
	Swin-UNet	93.87	72.30	63.23	67.46	72.46
	TransFuse	95.71	76.01	65.03	70.12	74.27
	DPBFN	95.66	75.34	65.05	69.84	74.19

5. Discussion and Explanation Analysis

The higher-level visual constructs [38] and global context information [39] play significant roles in road extraction from HR remote sensing imagery with complex background. In terms of HR remote sensing images interpretation, it is vital to explain the effectiveness of the extracted method. The class activation map (CAM) is a widely adopted and efficacious approach for interpretation in various domains. Thus, we utilize Grad-CAM [40] generated class activation maps with gradient information, as shown in Figure 11. It leverages the gradient information flowing into the last convolutional layer of both branches to assign importance value to each neuron for a particular decision of interest [40].

To acquire the road-discriminating localization map Grad-CAM loss $L^{road}_{Grad-CAM}$

$$L^{road}_{Grad-CAM} = ReLU(\sum_k \alpha^{road}_k A^k) \tag{14}$$

we first compute the gradient of the score of road class y^{road}, with respect to feature map A^k of activated convolution layers. These gradients flowing back are global-average-pooled over the width and height dimensions to obtain the neuron importance weights α^{road}_k:

$$\alpha^{road}_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^{road}}{\partial A^k_{ij}} \tag{15}$$

To intuitively demonstrate the effectiveness of the DPBFN, we present Grad-CAMs for comparison between the CNN branch without DCPA (in Figure 12b), the CNN branch with DCPA (in Figure 12c), and the transformer branch (in Figure 12d). Compared with the second and the third column, the mixed pixels problem in the Grad-CAMs (of CNN without DCPA module) is severe, and the road pixels are indistinguishable from the roof of the surrounding buildings. As shown in Figure 12c, the CNN branch with the DPCA module can alleviate the phenomenon of mixed pixels effectively, but there are still two obstacles as follows: (1) There is noise caused by the misrepresentation of similar feature pixels; (2) the stability of the convolution branch to the shadowed road recognition is poor.

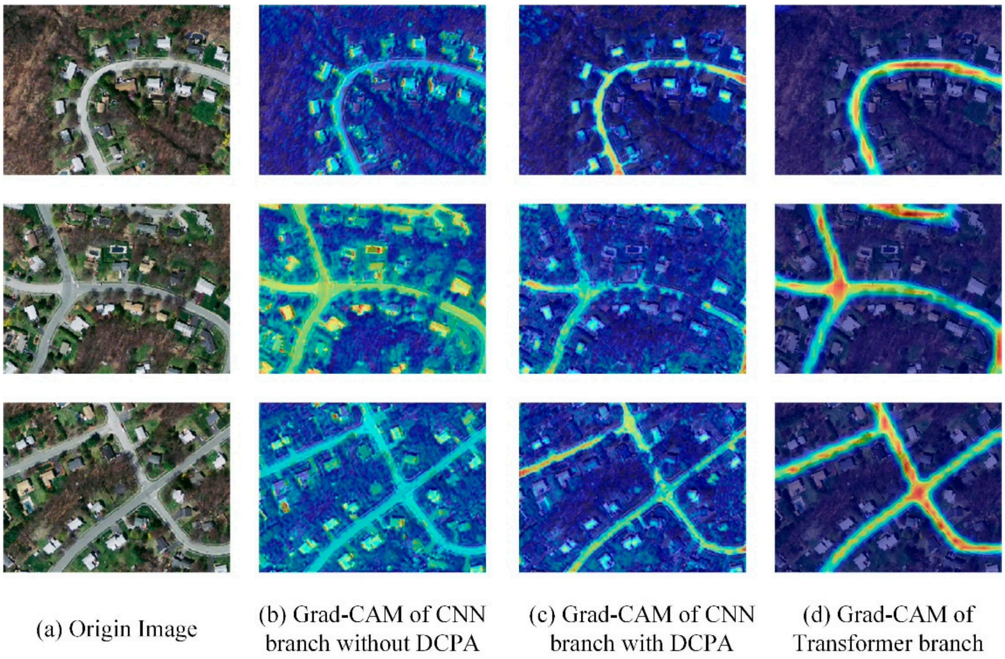

(a) Origin Image
(b) Grad-CAM of CNN branch without DCPA
(c) Grad-CAM of CNN branch with DCPA
(d) Grad-CAM of Transformer branch

Figure 12. Visualization of Grad-CAMs from the dual branches.

As shown in Figure 12d, based on the Grad-CAMs of the transformer branch, it can be observed that the multi-head attention module has a positive impact on addressing local shadow occlusion and mixed pixels. Specifically, areas shown in Figure 12a that are significantly obscured by roadside trees can be effectively connected in trans-feature maps. The main reason is that embedding the patches sequence can focus on the context order; furthermore, as shown in Figure 12d, the transformer branch is not affected by mixing pixels. Even so, the transformer still has some problems; the trans module has the problem of boundary blurring due to insufficient edge refining in the road interpretation task of large-scale high-resolution remote sensing images. Therefore, the fusion of a transformer (focus on context information) branch and a CNN branch with a DPCA module (focus on edge and details) can effectively complement each other's features, resulting in improved robustness and accuracy.

6. Conclusions

In this paper, we employ a hybrid method for extracting roads from HR remote sensing imagery using the dual parallel branch fusion network. The DPBFN model consists of dual parallel branches (transformer branch and residual multi-scaled dilated convolution encoder branch). The convolutional branch is based on a dilated the convolution pyramid with an attention module and residual blocks. The major contributions are the refinement of the CNN branch, and the fused dual branch features to segment road regions more accurately. In addition, we utilized the Grad-CAMs to analyze our explanation of the method. The DPBFN was evaluated using three challenging datasets. The experiments prove that DPBFN achieves commendable performance in the task of road extraction from complex high-resolution images (complex backgrounds and mixed pixels).

In conclusion, it is apparent that the domain of road extraction is currently experiencing significant evolution, driven by advancements in sensor technologies and the increasing accessibility of high-resolution imagery. While our research has yielded appreciable improvements in road extraction accuracy, it is important to acknowledge that there are several promising avenues for further investigation in this field.

Author Contributions: Conceptualization, L.G.; Methodology, L.G. and C.C.; Writing—original draft, L.G. and C.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Liaoning Provincial Department of Education Youth Project, Gao Lin (NO.1030040000560).

Data Availability Statement: Data sharing not applicable.

Acknowledgments: Thanks to Yongxin Feng and Junchang Xin for their supervision and promotion of this manuscript. Meanwhile, thanks to the editors and reviewers.

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

References

1. 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]
2. Xu, Y.; Chen, H.; Du, C.; Li, J. MSACon: Mining Spatial Attention-Based Contextual Information for Road Extraction. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–17. [CrossRef]
3. 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*, 2–6.
4. 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: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18; Springer International Publishing: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
5. 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]
6. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
7. 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*, 1–12. [CrossRef]

8. Yang, M.; Yuan, Y.; Liu, G. SDUNet: Road extraction via spatial enhanced and densely connected UNet. *Pattern Recognit.* **2022**, *126*, 108549. [CrossRef]
9. Xie, G.; He, L.; Lin, Z.; Zhang, W.; Chen, Y. Road extraction from lightweight optical remote sensing image based on LMMI DeepLabv3 + [J/OL]. *Laser J.* **2023**, 1–8.
10. Jie, Y.; He, H.; Xing, K.; Yue, A.; Tan, W.; Yue, C.; Jiang, C.; Chen, X. MECA-Net: A Multiscale Feature Encoding and Long-Range Context-Aware Network for Road Extraction from Remote Sensing Images. *Remote Sens.* **2022**, *14*, 5342. [CrossRef]
11. He, H.; Yang, D.; Wang, S.; Wang, S.; Li, Y. Road extraction by using atrous spatial pyramid pooling integrated encoder-decoder network and structural similarity loss. *Remote Sens.* **2019**, *11*, 1015. [CrossRef]
12. Lu, X.; Zhong, Y.; Zheng, Z.; Liu, Y.; Zhao, J.; Ma, A.; Yang, J. multi-scale and multi-task deep learning framework for automatic road extraction. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 9362–9377. [CrossRef]
13. Lian, R.; Huang, L. DeepWindow: Sliding window based on deep learning for road extraction from remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 1905–1916. [CrossRef]
14. Zhou, K.; Xie, Y.; Gao, Z.; Miao, F.; Zhang, L. FuNet: A novel road extraction network with fusion of location data and remote sensing imagery. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 39. [CrossRef]
15. 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]
16. 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]
17. Abdollahi, A.; Pradhan, B.; Alamri, A. VNet: An end-to-end fully convolutional neural network for road extraction from high-resolution remote sensing data. *IEEE Access* **2020**, *8*, 179424–179436. [CrossRef]
18. 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]
19. Alshaikhli, T.; Liu, W.; Maruyama, Y. Simultaneous extraction of road and centerline from aerial images using a deep convolutional neural network. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 147. [CrossRef]
20. 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 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929, 2020.
21. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021. [CrossRef]
22. Ge, C.; Nie, Y.; Kong, F.; Xu, X. Improving Road Extraction for Autonomous Driving Using Swin Transformer Unet. In Proceedings of the 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), IEEE, Macau, China, 8–12 October 2022; pp. 1216–1221.
23. Luo, L.; Wang, J.X.; Chen, S.B.; Tang, J.; Luo, B. BDTNet: Road extraction by bi-direction transformer from remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
24. 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. [CrossRef]
25. Zhang, Y.; Liu, H.; Hu, Q. Transfuse: Fusing transformers and cnns for medical image segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1 2021; Proceedings, Part I 24; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; pp. 14–24.
26. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
27. Lin, A.; Chen, B.; Xu, J.; Zhang, Z.; Lu, G.; Zhang, D. Ds-transunet: Dual swin transformer u-net for medical image segmentation. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–15. [CrossRef]
28. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
29. Miao, C.; Liu, C.; Zhang, Z.; Tian, Q. TransLinkNet: LinkNet with transformer for road extraction. In Proceedings of the International Conference on Optics and Machine Vision (ICOMV 2022), Guangzhou, China, 14–16 January 2022.
30. Tao, J.; Chen, Z.; Sun, Z.; Guo, H.; Leng, B.; Yu, Z.; Wang, Y.; He, Z.; Lei, X.; Yang, J. Seg-Road: A Segmentation Network for Road Extraction Based on Transformer and CNN with Connectivity Structures. *Remote Sens.* **2023**, *15*, 1602. [CrossRef]
31. Zhang, Z.; Miao, C.; Liu, C.A.; Tian, Q. DCS-TransUperNet: Road segmentation network based on CSwin transformer with dual resolution. *Appl. Sci.* **2022**, *12*, 3511. [CrossRef]
32. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.S.; et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890.
33. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018. [CrossRef]

34. Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; Raska, R. Deepglobe 2018: A challenge to parse the earth through satellite images. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, Salt Lake City, UT, USA, 18–20 June 2018; pp. 172–181.
35. Mnih, V. *Machine Learning for Aerial Image Labeling*; University of Toronto: Toronto, ON, Canada, 2013; p. 109.
36. Gao, L.; Song, W.; Dai, J.; Chen, Y. Road Extraction from High-Resolution Remote Sensing Imagery Using Refined Deep Residual Convolutional Neural Network. *Remote Sens.* **2019**, *11*, 552. [CrossRef]
37. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
38. Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef]
39. Mahendran, A.; Vedaldi, A. Visualizing Deep Convolutional Neural Networks Using Natural Pre-Images. *Int. J. Comput. Vis.* **2015**, *120*, 233–255. [CrossRef]
40. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the IEEE International Conference on Computer Vision, IEEE, Venice, Italy, 22–29 October 2017. [CrossRef]

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