remote sensing

MDPI

Article

Deep-Separation Guided Progressive Reconstruction Network for Semantic Segmentation of Remote Sensing Images

Jiabao Ma [1], Wujie Zhou [1,2,*], Xiaohong Qian [1] and Lu Yu [2]

[1] School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China
[2] College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China
* Correspondence: 109029@zust.edu.cn

Abstract: The success of deep learning and the segmentation of remote sensing images (RSIs) has improved semantic segmentation in recent years. However, existing RSI segmentation methods have two inherent problems: (1) detecting objects of various scales in RSIs of complex scenes is challenging, and (2) feature reconstruction for accurate segmentation is difficult. To solve these problems, we propose a deep-separation-guided progressive reconstruction network that achieves accurate RSI segmentation. First, we design a decoder comprising progressive reconstruction blocks capturing detailed features at various resolutions through multi-scale features obtained from various receptive fields to preserve accuracy during reconstruction. Subsequently, we propose a deep separation module that distinguishes various classes based on semantic features to use deep features to detect objects of different scales. Moreover, adjacent middle features are complemented during decoding to improve the segmentation performance. Extensive experimental results on two optical RSI datasets show that the proposed network outperforms 11 state-of-the-art methods.

Keywords: digital surface model; multimodal; multi-scale supervision; feature separation; reconstruction refinement

Citation: Ma, J.; Zhou, W.; Qian, X.; Yu, L. Deep-Separation Guided Progressive Reconstruction Network for Semantic Segmentation of Remote Sensing Images. *Remote Sens.* **2022**, *14*, 5510. https://doi.org/10.3390/rs14215510

Academic Editor: Gwanggil Jeon

Received: 15 September 2022
Accepted: 30 October 2022
Published: 1 November 2022

1. Introduction

Semantic segmentation aims to semantically classify the pixels in an image [1]. In remote sensing, semantic segmentation is crucial in several applications, such as scene understanding [2], land cover classification [3], and urban planning [4]. Owing to the success of deep learning (DL) and the promising results obtained on multiple semantic segmentation benchmarks containing natural images [5–7], semantic segmentation of remote sensing images (RSIs) increasingly adopts DL approaches [8–10]. However, an RSI is substantially larger than a typical natural image for computer vision applications; it contains objects of different sizes and shows complex scenes. Moreover, during data acquisition, the tilted perspective of RSIs can lead to scale variations in objects captured at different distances [11,12], exacerbating problems related to multi-scale changes.

With the continuous development of DL, convolutional neural networks have ushered in a new era of computer vision. The full convolution was proposed by Long et al. [13] to replace a fully connected layer with a convolutional layer in a classification network. However, decoding relies on the deep semantic features obtained from upsampling to obtain an output prediction map. Accordingly, using U-shaped architectures, Ronneberger et al. [14] and Vijay et al. [15] proposed UNet and SegNet, respectively, which use upsampling and continuous convolutions to complete decoding; each layer splices features from the encoding stage. The method of supplementing the features extracted from the encoder to the decoder is also often used in the later semantic segmentation methods, enhancing the complementarity of features in a different phase. Inspired by the above methods, Jiang et al. [16] proposed RedNet in 2018 with the same decoding approach, obtaining intermediate prediction maps at each stage to supervise the network at different resolutions.

Moreover, Chen et al. [17] used the splicing of deep semantic features with shallow features and upsampling for the prediction map. This study proposed the idea of atrous convolution to expand the receptive field in the convolution process. Meanwhile, Chen et al. [18] proposed the feature separation and aggregation models for fusing multimodal features and fully exploring the characteristics of different stages. However, the architecture of the decoder makes it simple to fully use the features extracted by the encoder. In addition, Yu et al. [19] used deep features from two encoder branches to construct prediction maps during decoding for fast inference and the real-time performance of the proposed method. Xu et al. [20] used an attention mechanism and multi-branch parallel architecture to build lightweight networks for real-time segmentation. Such structures helped obtain representations of objects at different scales. For the multimodal data, Zhou et al. [21] and Seichter et al. [22] used RGB and depth multimodal data to complete the semantic segmentation task of indoor scenes. The proposed network followed the encoder–decoder architecture combined with the last three high-level features of the encoder to construct the prediction map, which was simple in structure. However, this approach failed to make full use of shallow features. Some details need to be included in the refactoring process. Hu et al. [23] used five decoder blocks with the same encoding structure and applied upsampling to each block to restore the resolution of the prediction map. This is one of the most widely used architectures in semantic segmentation in recent years. Middle feature streams are deployed in multimodal data to deal with fused features, which can enhance the representation of multimodal features. In addition, researchers have widely favored multimodal data in different fields [24–26]. In particular, to handle quality variations across multimodality RSI datasets, Zheng et al. [27] use a DSM (Digital Surface Model) as auxiliary information to improve the segmentation performance of the model on single-modal data. Nevertheless, the method only applies self-attention to the deepest feature, and the structure of the decoder is relatively simple. Thus, it fails to detect the object in the complex scene. Similarly, Ma et al. [28] used powerful encoding features with a transformer to extract multimodality information. In this approach, the transformer is fully combined with CNN to deal with multi-scale features.

Most networks for tasks such as semantic segmentation have encoder–decoder architectures. Common encoders include VGG [29], ResNet [30] and, recently, the transformer [31]. However, the feature extraction ability of these encoders is limited to some extent. In particular, network performance improvements depend on how to handle the above features and, most importantly, how to reconstruct the features in the decoder. For developing decoders, different architectures have been devised; however, the bottom-up approach is typically used after feature extraction. A typical decoder is UNet [14], which is the basis for several subsequently developed networks. Various studies [32] have performed a fusion of features with different scales after extraction to improve feature reconstruction. In such methods, the decoder contains common convolutional and upsampling layers. Although its implementation is simple, this type of decoder lacks efficiency. Moreover, the features extracted by the encoder contain different levels of meaning at different resolutions. That is, current methods cannot take advantage of these features. How to reconstruct features efficiently and cooperate with each other is crucial in the design model.

To solve the abovementioned problems, we propose a deep-separation-guided progressive reconstruction network (DGPRNet) comprising a deep separation module (DSEM) for semantic segmentation of RSIs. In particular, to improve feature reconstruction, we design a progressive reconstruction block (PRB) based on atrous spatial pyramid pooling (ASPP) [33] with multiple convolutional layers combining various receptive fields for refactoring characteristics at each resolution. Unlike other methods based on upsampling to increase the resolution [34], the PRBs use deconvolution to adjust the resolution, increasing through each block until the input image is solved. Moreover, to enhance the forward guidance of deep semantic features to shallow layers, the proposed deep separation module (DSEM) processes semantic features such that pixels of the same class are clustered, whereas the separation between pixels from different classes is maximized. The prediction map is

multi-supervised. Thus, the expression ability of deep semantic features is enhanced, and the PRB provides positive feedback.

The study's contributions are as follows:

1. A PRB based on ASPP [33] was embedded in the decoder to strengthen the feature reconstruction and reduce the error in this process. Five features with different resolutions were processed serially using atrous convolution layers with different ratios, and the feature resolution was expanded by deconvolution to obtain the decoding output of each block.
2. The proposed DSEM processed the last three semantic features from the decoder to emphasize semantic information to use deep semantic features. Intraclass separation was minimized, while interclass separation was maximized. Meanwhile, multi-supervision was applied to DGPRNet for segmentation, improving the reconstruction ability of each module.
3. Experiments on two RSI datasets showed that the proposed model outperforms 11 state-of-the-art methods, including current semantic segmentation methods.

2. Proposed DGPRNet

Figure 1 shows the architecture of the proposed DGPRNet. In particular, the architecture comprised symmetric ResNet-50 [30] backbones for feature extraction and the novel decoder consisting of PRBs and DSEM for processing semantic features. As seen in Figure 1, the DGPRNet adopts an encoder–decoder architecture. The two symmetric ResNet-50 backbones constituted the encoder processing input images by extracting features at five different resolutions from RGB (red–green–blue)/DSM (digital surface model) RSIs. According to the features extracted by the encoder, the adjacent modules from [35] were used between features F_2, F_3, and F_4 for feature aggregation. During decoding, inspired by ASPP [33], we used the proposed PRBs to reconstruct and combine the features at various resolutions, and each PRB provided a prediction map at the corresponding resolution. The DSEM classified the last three deep semantic features. Finally, we obtained the prediction map from five scales.

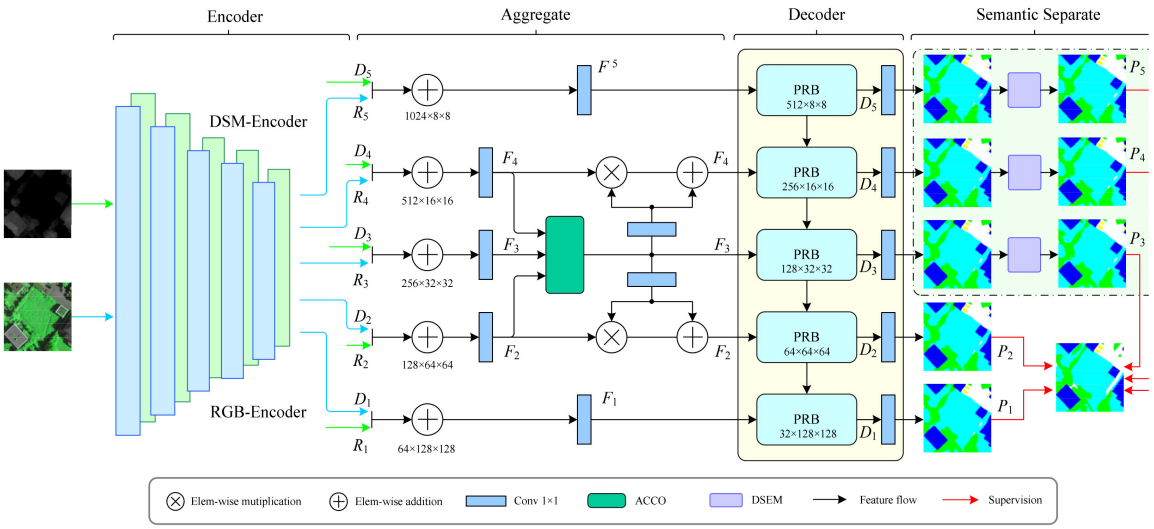

Figure 1. Overall architecture of the proposed DGPRNet. The network includes four stages: encoding, feature aggregation, decoding, and semantic separation.

2.1. Encoder

RGB and DSM images contained unreliable information and objects of different sizes due to the complexity of real scenes and the diversity of RSIs. In particular, feature extraction was essential in existing image semantic segmentation methods based on DL.

We used ResNet-50 as the encoder to obtain five features with different resolutions, R_i and D_i, $i \in \{1, 2, 3, 4, 5\}$, from the RGB and DSM images, respectively, and fused the features by simple pixel-wise addition [36–39], obtaining feature F_i. Shallow features contained details such as object boundaries, and deep features reflected semantic information such as the class and location of an object. RSIs contained various objects of different sizes. In particular, detecting these objects is crucial. Moreover, we used the module from [34] to aggregate multi-scale information. We retained the original information of the shallowest and deepest features and only used the features of the middle three resolutions to obtain the aggregated representation of adjacent features. The aggregated features were supplemented with features at the corresponding resolution. The encoder was formulated as follows:

$$F_i = Conv(R_i \oplus D_i), \tag{1}$$

$$\begin{cases} F_3 = ACCO(F_2, F_3, F_4) \\ F_2 = F_2 \otimes Conv(F_3) \oplus Conv(F_3) \\ F_4 = F_4 \otimes Conv(F_3) \oplus Conv(F_3), \end{cases} \tag{2}$$

where $\oplus$ denotes pixel-wise addition, $\otimes$ denotes pixel-wise multiplication, and *Conv* represents a convolutional layer with batch normalization and rectified linear unit activation. Subsequently, the aggregated representation of an object at different resolutions can be obtained. In this way, multi-scale objects can be accurately detected.

2.2. PRB

Universal networks work well on all datasets. Therefore, a critical problem in applying DL to computer vision is reconstructing the features extracted by the encoder according to the characteristics of a specific dataset, and finally providing an accurate prediction map. Therefore, we proposed the PRB (shown in Figure 2), where the encoder extracted features with different resolutions. Based on ASPP, we used dilated convolutions with different rates in series to enlarge the receptive field at each resolution. In each block, objects of different sizes and those with different dimensions were detected at different resolutions.

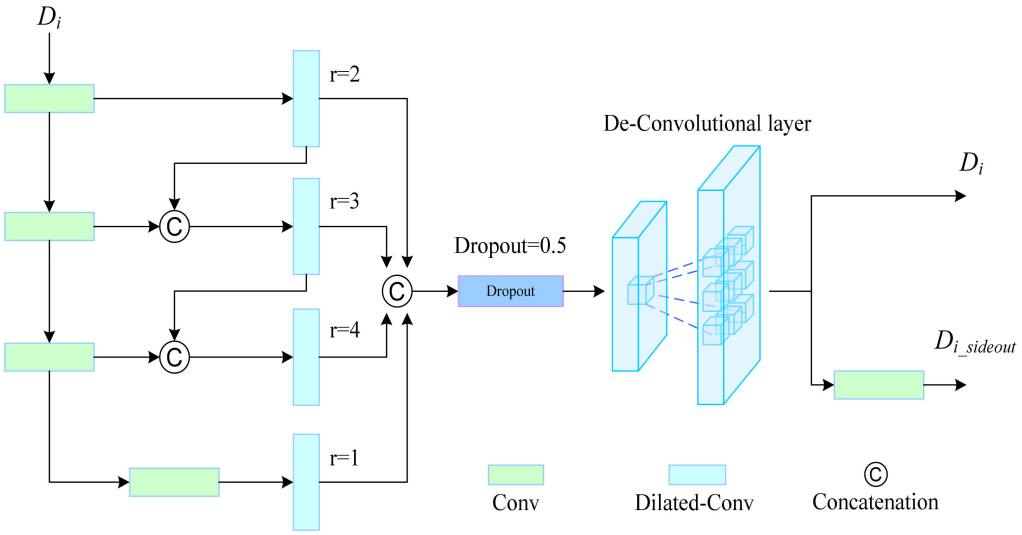

Figure 2. Architecture of the proposed PRB.

After the encoded features were obtained, the PRB reconstructed the features at each resolution. In [33], ASPP modules were deployed at the bottom of the network, acting on the deepest semantic features to expand the receptive field. However, the intended effect was limited. Based on ASPP, we expanded the receptive field in each layer during decoding to detect objects at different resolutions. Specifically, the PRB contained four convolutional layers with different dilation rates. Moreover, the features were serially

transferred between convolutional layers. Then, they were concatenated in parallel for feature aggregation under different receptive fields. Moreover, upsampling enables us to increase the resolution of features [31]. Inspired by deconvolution, we merged upsampling and feature aggregation into one step, and a dropout layer was added to prevent overfitting. The PRB was formulated as follows:

$$D_i = Cat(F_i, D_{i+1}), \tag{3}$$

$$\begin{cases} D_{i,c1} = Conv(D_i) \\ D_{i,c2} = Conv(D_{i,c1}) \\ D_{i,c3} = Conv(D_{i,c2}), \end{cases} \tag{4}$$

$$\begin{cases} D_{i,d2} = DConv(D_i, 2) \\ D_{i,d3} = DConv(Cat(D_{i,d2}, D_{i,c2}), 3) \\ D_{i,d4} = DConv(Cat(D_{i,d3}, D_{i,c3}), 4), \end{cases} \tag{5}$$

$$D_i = DeConv(Dropout(Cat(D_{i,c3}, D_{i,d2}, D_{i,d3}, D_{i,d4}), 0.5)), \tag{6}$$

where *Cat*, *DConv*, and *DeConv* denote concatenation, a dilated convolutional layer, and a deconvolutional layer, respectively, and $i \in \{1, 2, 3, 4, 5\}$, with D_{i+1} being omitted for $i = 5$ in Equation (3).

2.3. DSEM

Deep semantic features represent the mapping of an image onto a semantic space. Moreover, the feature representation of pixels belonging to a class in complex scenes showed high variability, and RSIs corresponded to complex scenes. Consequently, different objects might be classified into the same class in some cases. To increase the classification accuracy, we proposed the DSEM that modeled intraclass and interclass features to strengthen their distinguishability and reduce ambiguity. First, high-level semantic feature map D_i, $i \in \{3, 4, 5\}$ was processed by a 1×1 convolutional layer to obtain feature maps α, β, $\gamma \in R^{C \times H \times W}$. Then, the features were processed to obtain different expressions within and between classes. The DSEM was formulated as follows:

$$\begin{cases} intra = Softmax(R(\alpha) \times T(R(\beta))) \times R(\gamma) \\ intra = F(intra) + D_i, \end{cases}, \tag{7}$$

$$\begin{cases} inter = Softmax(T(R(D_i)) \times R(D_i)) \\ inter = F(inter) \times D_i + D_i, \end{cases}, \tag{8}$$

$$P_i = (inter + intra) \otimes D_i, \tag{9}$$

where R denotes a resizing function from $R^{C \times H \times W}$ to $R^{C \times HW}$, T is the transposition from $R^{C \times HW}$ to $R^{HW \times C}$, F denotes the inverse mapping of R, and $\times$ denotes matrix multiplication.

The original semantic features were combined with the weights for intraclass and interclass features to obtain a deep separation prediction with higher resolution and more detailed feature classification performance while reducing feature redundancy. We applied the DSEM to features of the last three resolutions obtained from decoding. The network simultaneously performed prediction at five resolutions during training and supervised the network. Hence, the reconstruction ability during decoding was strengthened by integrating the DSEM.

2.4. Loss Function

We used binary cross-entropy as the loss function between the prediction map and the segmentation ground truth. The obtained prediction maps at five resolutions were resized

to the dimension of the ground truth to calculate the loss. Given the five prediction maps, the binary cross-entropy loss function was defined as follows:

$$Loss = \sum_{i=1}^{5} BCE(P_i, GT), \tag{10}$$

where GT and P denote the ground truth and a prediction map, respectively. During testing, P_1 is the segmentation result of DGPRNet.

3. Experiments and Results

3.1. Datasets and Performance Indicators

The Potsdam [40] and Vaihingen [41] RSI datasets were used in semantic segmentation experiments to verify the performance of the proposed DGPRNet. The Potsdam dataset contains 38 patches of 6000 × 6000 pixels, in our experiments, we considered 17 patches for training (2_10, 3_10, 3_11, 3_12, 4_11, 4_12, 5_10, 5_12, 6_8, 6_9, 6_10, 6_11, 6_12, 7_7, 7_9, 7_11, 7_12) and 7 patches for testing (2_11, 2_12, 4_10, 5_11, 6_7, 7_8, 7_10). Furthermore, the Vaihingen dataset comprised 33 images with pixels of 2494 × 2064. We split the 16 patches for training (1, 3, 5, 7, 13, 17, 21, 23, 26, 32, 37) and 5 patches for testing (11, 15, 28, 30, 34). For evaluation, we used the average pixel accuracy of each class and the average intersection over union as performance indicators. Moreover, the intersection over union was applied between the prediction and target regions to obtain the optimal segmentation weight. Therefore, the intersection over union was the main indicator for training and evaluating different methods on the two RSI datasets.

3.2. Implementation Details

The experimental platform was implemented in Ubuntu 20.04 using the PyTorch 1.9.1 environment, and the network was trained on a computer equipped with an NVIDIA Titan V graphics card and 12 GB of memory. Owing to the large size of the original RSIs, the patches of input RGB and DSM images were scaled to 256 × 256 pixels. Finally, we obtained 35,972 slices for training, 4032 slices for validation, and 252 slices for testing on Potsdam. Moreover, we obtained 17,656 slices for training, 412 slices for validation, and 111 slices for testing on Veihingen. The pretrained ResNet-50 was used as the backbone for feature extraction. Considering the dataset characteristics and training time, training proceeded over 300 epochs on the Vaihingen dataset and over 100 epochs on the Potsdam dataset. We used stochastic gradient descent with a momentum of 0.9, weight decay of 0.9, batch size of 10, and learning rate of 5×10^{-4} for optimization. Moreover, we used a poly strategy [42] to adjust the learning rate during training. The training process of the model on the two datasets took approximately 32 h, and the test time was 73 min and 6 min, respectively, on Potsdam and Vaihingen, including the model inference time and the concatenation from slices into high-resolution remote sensing images.

3.3. Comparison with State-of-the-Art Methods

3.3.1. Quantitative Evaluation

In particular, Tables 1 and 2 list the performance indicators obtained by applying various methods on the two RSI semantic segmentation datasets. Table 3 summarizes the comparison results of all models in terms of flops and parameters, including the method based on Transformer [43]. The proposed DGPRNet outperformed the comparison methods on the Potsdam and Vaihingen datasets, and the indicators verified the high detection performance of the proposed method. Moreover, the DGPRNet detection of the class car on the two datasets was remarkable, confirming correct object detection in challenging scenes. Compared with existing methods, DGPRNet showed outstanding results in three classes, namely impervious surfaces, buildings, and cars, on the Vaihingen dataset. In addition, both mAcc and mIoU outperformed the best indicators in the comparisons. In particular, the IoU indicator of DGPRNet in the impervious surface and building outperformed the

participating methods by 0.52% (SA-Gate) and 0.96% (ACNet). Especially for the class car, DGPRNet reached 92.30% and 84.84% in Acc and IoU indicators, which exceeded 8.03% and 6.77% compared with SA-Gate. In addition, the overall indicators mAcc and mIoU reached 90.43% and 82.36%, respectively, increasing by 1.81% and 1.69% compared to SA-Gate. Furthermore, on the Potsdam dataset, the DGPRNet outperformed the comparison methods in almost all classes except for low vegetation, tree, and clutter on the classification accuracy. Similarly, IoU in car and cluster reached 92.46% and 47.02%, respectively, compared with more than 2.03% and 3.48% for ACNet and Deeplabv3+. The accuracy on the class car exceeded HRCNet by 2.09% and reached 96.03%. In terms of overall performance, the proposed DGPRNet achieved mAcc of 85.69% and mIoU of 77.69%, increasing by 1.27% and 1.79% compared to the SA-Gate and RedNet, respectively. The improvement in the overall indicators of the proposed method in the small category can be explained as follows: by complementing each other at different resolutions, the aggregate representation information of a specific category at multiple scales can be obtained, greatly improving the accuracy and IoU on the small objects. Therefore, the improvement in this category is particularly significant.

Table 1. Quantitative results of the proposed DGPRNet and 11 state-of-the-art methods on the Vaihingen dataset. The values in bold indicate the best scores in the evaluation matrix.

		FCN-8S [13]	U-Net [14]	SegNet [15]	DeepLabv3+ [17]	BiseNetV2 [19]	HRCNet [20]	RedNet [16]	ACNet [23]	SA-Gate [18]	TSNet [21]	ESANet [22]	DCSwin [43]	Ours
Imp.surf	Acc	89.66	91.68	89.88	90.06	90.56	91.62	91.49	91.95	90.99	87.93	92.09	91.40	91.55
	IoU	79.71	80.90	80.93	81.11	80.97	81.60	84.62	85.34	85.70	78.98	85.18	84.45	86.22
Building	Acc	93.22	89.84	90.88	87.04	91.24	91.72	94.81	95.45	93.85	95.81	94.93	95.29	95.80
	IoU	86.80	86.50	86.54	82.70	86.69	88.01	91.07	91.82	91.72	91.47	91.16	91.30	92.78
Low veg.	Acc	75.83	77.97	78.66	76.65	74.68	79.24	78.67	78.64	84.95	71.62	75.72	79.02	81.27
	IoU	64.33	65.91	64.07	64.44	63.66	67.38	66.59	66.87	68.68	57.03	65.48	66.26	68.62
Tree	Acc	89.22	91.30	88.96	88.60	91.66	90.55	91.41	91.20	89.06	94.26	92.35	89.85	91.22
	IoU	75.58	77.86	75.96	76.64	76.54	78.58	78.27	78.55	79.15	81.26	77.65	77.54	79.34
Car	Acc	45.12	75.80	43.93	42.51	63.75	70.69	59.77	83.12	84.27	67.63	75.92	81.51	92.30
	IoU	40.16	71.22	43.16	43.10	61.80	68.73	56.06	76.81	78.07	66.86	70.11	73.47	84.84
mAcc		78.61	79.75	78.46	76.97	82.38	84.76	83.23	88.07	88.62	83.54	86.20	87.41	90.43
mIoU		69.32	71.34	70.13	69.49	73.93	76.86	75.32	79.88	80.67	75.12	77.92	78.60	82.36

Table 2. Quantitative results of the proposed DGPRNet and 11 state-of-the-art methods on the Potsdam dataset.

		FCN-8S [13]	U-Net [14]	SegNet [15]	DeepLabv3+ [17]	BiseNetV2 [19]	HRCNet [20]	RedNet [16]	ACNet [23]	SA-Gate [18]	TSNet [21]	ESANet [22]	DCSwin [43]	Ours
Imp.surf	Acc	89.47	90.03	90.18	91.57	90.12	90.03	92.19	91.32	85.84	85.22	91.38	91.66	92.76
	IoU	79.77	80.27	80.46	82.49	80.58	81.68	82.83	82.74	80.64	76.85	82.92	82.28	83.33
Building	Acc	90.69	88.71	90.21	91.78	88.88	90.87	93.61	93.83	93.65	91.85	93.69	92.92	93.94
	IoU	83.60	82.92	84.18	87.59	83.70	85.75	90.13	90.06	88.51	86.65	89.82	89.12	91.26
Low veg.	Acc	85.13	85.82	85.88	87.36	87.68	88.17	87.00	86.16	86.46	88.52	87.10	87.31	87.12
	IoU	71.12	71.60	71.63	73.63	71.55	73.18	73.22	73.53	72.71	67.98	73.16	74.48	74.46
Tree	Acc	82.86	84.06	82.49	85.45	81.11	82.02	83.00	86.03	85.70	78.75	82.48	84.46	85.84
	IoU	71.23	72.05	70.68	73.32	71.55	71.32	71.77	72.87	72.89	67.49	70.81	73.23	73.80
Car	Acc	91.02	93.89	93.15	93.89	93.14	93.94	93.36	93.79	92.18	78.22	93.08	96.31	96.03
	IoU	81.53	90.24	89.72	90.04	89.17	89.82	90.08	90.43	89.39	76.85	88.53	90.12	92.46
Clutter	Acc	49.05	50.30	51.76	53.80	50.66	56.72	56.74	54.51	62.70	37.49	55.68	56.01	58.48
	IoU	36.49	36.26	37.21	43.54	36.35	40.03	43.51	41.65	40.59	30.85	43.38	43.37	47.02
mAcc		78.61	82.13	82.28	83.97	81.93	83.63	84.32	84.27	84.42	76.68	83.90	84.61	85.69
mIoU		69.32	72.22	72.31	75.10	71.86	73.63	75.26	75.21	74.12	67.78	74.77	75.43	77.05

3.3.2. Qualitative Evaluation

Figure 3 shows the segmentation results obtained using DGPRNet and 11 state-of-the-art methods. Examples of multiple scenes were included, such as scenes with objects of different scales (clutter), small objects (car), large objects (building), low contrast with the background, and blurred boundaries. In general, the qualitative results show that DGPRNet has improved scene adaptability and reconstruction accuracy compared to similar methods.

In the visual contrast result, the area marked by the red rectangle shows the place that differs the most. Some methods could not precisely locate the objects of clutter because the object was usually located in a complex scene with different sizes. As shown in the first to fifth lines of Figure 3, the clutter can be accurately located in many complex scenes compared with other models. As the classes with the largest proportion in the dataset, the detection results of the proposed method for buildings are accurate and the edges are smooth, as shown in the second, third, sixth, and seventh rows in Figure 3. Compared with other methods, there are fewer cases of incomplete detection. The key problem to be solved in this study is to accurately segment the small objects (car) in the dataset. As seen from the fifth to the ninth rows of Figure 3, the segmentation result of the class car is more precise than that of other models. The qualitative results showed that DGPRNet better adapted to different scenes and reconstructed features with higher accuracy than similar methods. Moreover, DGPRNet performed highly in various complex scenes and detected small objects and the object's edges better than the other evaluated methods.

Table 3. The comparison on flops and parameters in all methods.

	Flops (GMac)	Params (M)
FCN8s	74.55	134.29
UNet	55.93	26.36
SegNet	18.3	53.56
DeepLabv3+	32.45	59.33
BiseNet	3.23	3.63
HRCNet	30.28	62.71
RedNet	21.17	81.95
ACNet	26.41	116.6
SA-Gate	41.23	110.85
TSNet	34.27	41.8
ESANet	10.15	45.42
DCSwin	34.4	118.39
Ours	55.39	142.82

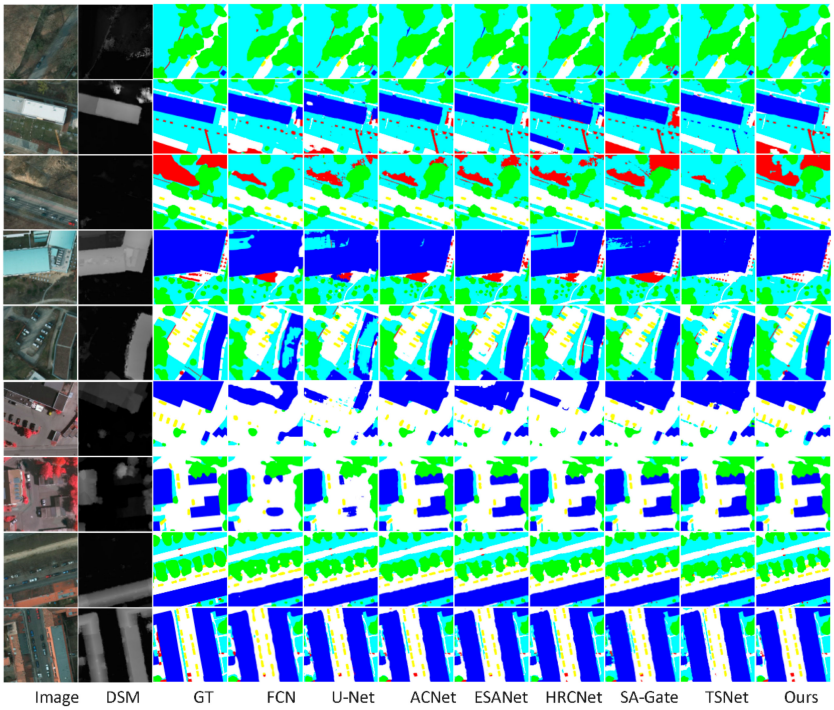

Figure 3. Comparison of segmentation results from different methods.

3.4. Ablation Study

To verify the effectiveness of the adopted modules, we conducted a comparative experiment on two datasets. Table 4 lists the comparison of the ablation indicators.

Table 4. Ablation study on the Vaihingen and Potsdam datasets.

	Vaihingen		**Potsdam**	
	mAcc	**mIoU**	**mAcc**	**mIoU**
Baseline	84.49	76.66	83.72	74.79
W/o DSM	89.67	81.07	84.89	75.88
W/o PRB	88.50	80.73	84.54	74.82
W/o DSEM	87.14	78.84	80.81	69.85
Ours	**90.43**	**82.36**	**85.69**	**77.05**

3.4.1. Effect of Modal DSM

The effectiveness of multimodal data was verified. In this regard, the DSM data were removed and represented as w/o DSM in Table 4. The ablation results show that the scores of the model in the single modal are slightly lower than those in the multimodal data. The results indicate that DSM modal data can indeed improve the performance of the model from another perspective.

3.4.2. Effects of Module PRB

The w/o PRB indicates the scheme implemented without the PRB module. In the decoding part, we replaced the PRB module with the convolution block combined with 3×3 convolutional layers + BN + ReLU to verify the effectiveness of the PRB module. Figure 4 shows the prediction map. The scheme without the PRB module performed lower than the full model. For example, the detection area of the building in the first, second, and fourth rows is discontinuous. Furthermore, a clutterer was present with incorrect classification in the third and fourth rows. The above comparison diagram also verifies that the PRB module reduces the feature reconstruction error in the process of network decoding and plays a crucial role in network inference. Compared with the full model, the ablation indicators mAcc and mIoU of PRB decreased by 1.93% and 1.63% on the Vaihingen dataset and 1.15% and 2.23% on the Potsdam dataset, respectively. The above results demonstrate the importance of the PRB module from both qualitative and quantitative perspectives.

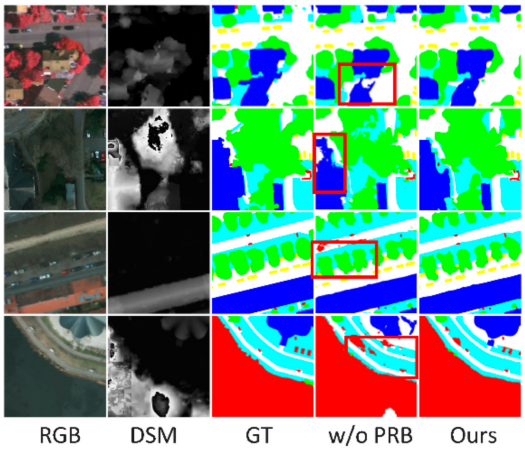

Figure 4. Ablation performance comparisons with the effect of PRB.

3.4.3. Effects of Module DSEM

Similarly, we verified the effectiveness of DSEM and the multi-supervision strategy. We removed the DSEM and multi-supervision strategy of the last three layers of semantic

features and used the scheme w/o DSEM to represent it. The performance of the scheme w/o DSEM was low. As shown in Table 4, compared with the full model, the DSEM scheme decreased the two indicators by 3.29% and 3.52% in the Vaihingen dataset and 4.88% and 7.2% in the Potsdam dataset, respectively. Moreover, from the perspective of visualization, if no further exploration of the deep feature exists between classes, the first and second lines of Figure 5 are confused with building and clutter. Similarly, a category misclassification will be present in the third and fourth lines. Hence, we concluded that this module considerably facilitated the reconstruction ability of the network decoding layer. The module improved the specification effect on the deep semantic features and helped the decoding module enhance reconstruction under different resolutions.

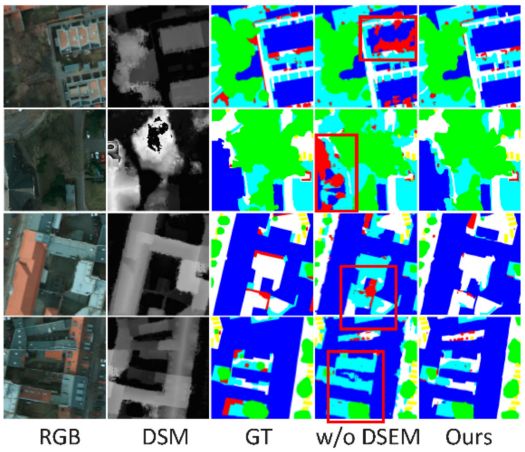

Figure 5. Ablation performance comparisons with the effect of DSEM.

4. Conclusions

This study proposed a novel network framework called DGPRNet for semantic segmentation of remote sensing images by exploring inter and intraclass relationships in deep features and decreasing feature reconstruction loss in the decoder. First, adjacent intermediate features were complemented before decoding to improve the expression of multi-scale features. Second, PRB was developed and deployed at five stages in the decoder to capture detailed features obtained from different receiving fields at multiple resolutions, reducing error and maintaining accuracy during reconstruction. Finally, the proposed DSEM distinguished and aggregated interclass and intraclass features based on semantic features to leverage deep features in detecting objects with different scales. Experimental results on two RSI datasets showed that DGPRNet outperformed 11 state-of-the-art methods.

Author Contributions: Conceptualization, J.M. and W.Z.; methodology, J.M., X.Q.; software, X.Q.; validation, L.Y.; writing—review and editing, W.Z., L.Y.; supervision, W.Z.; project administration, W.Z.; funding acquisition, X.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by National Natural Science Foundation of China (61502429, 61672337, 61972357); the Zhejiang Provincial Natural Science Foundation of China (LY18F020012, LY17F020011), and Zhejiang Key R & D Program (2019C03135).

Data Availability Statement: The code used and the datasets generated during the different steps of the analysis are available from the corresponding author on reasonable request.

Acknowledgments: The authors sincerely appreciate the helpful comments and constructive suggestions given by the academic editors and reviewers.

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

References

1. Neupane, B.; Horanont, T.; Aryal, J. Deep learning-based semantic segmentation of urban features in satellite images: A review and meta-analysis. *Remote Sens.* **2021**, *13*, 808. [CrossRef]
2. Zhou, W.; Liu, J.; Lei, J.; Hwang, J.-N.; Yu, L. GMNet: Graded-feature multilabel-Learning network for RGB-Thermal urban scene semantic segmentation. *IEEE Trans. Image Process.* **2021**, *30*, 7790–7802. [CrossRef]
3. Chen, Z.; Li, D.; Fan, W.; Guan, H.; Wang, C.; Li, J. Self-attention in reconstruction bias U-Net for semantic segmentation of building rooftops in optical remote sensing images. *Remote Sens.* **2021**, *13*, 2524. [CrossRef]
4. Wang, L.; Li, R.; Wang, D.; Duan, C.; Wang, T.; Meng, X. Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images. *Remote Sens.* **2021**, *13*, 3065. [CrossRef]
5. Zhou, W.; Yang, E.; Lei, J.; Wan, J.; Yu, L. PGDENet: Progressive Guided Fusion and Depth Enhancement Network for RGB-D Indoor Scene Parsing. *IEEE Trans. Multimed.* **2022**. [CrossRef]
6. Zhou, W.; Liu, W.; Lei, J.; Luo, T.; Yu, L. Deep binocular fixation prediction using hierarchical multimodal fusion network. *IEEE Trans. Cogn. Dev. Syst.* **2021**. [CrossRef]
7. Wu, J.; Zhou, W.; Luo, T.; Yu, L.; Lei, J. Multiscale multilevel context and multimodal fusion for RGB-D salient object detection. *Signal Process.* **2021**, *178*, 107766. [CrossRef]
8. Zhou, W.; Jin, J.; Lei, J.; Yu, L. CIMFNet: Cross-Layer Interaction and Multiscale Fusion Network for Semantic Segmentation of High-Resolution Remote Sensing Images. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 666–676. [CrossRef]
9. Liu, X.; Jiao, L.; Zhao, J.; Zhao, J.; Zhang, D.; Liu, F.; Tang, X. Deep multiple instance learning-based spatial–spectral classification for PAN and MS imagery. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 461–473. [CrossRef]
10. Mou, L.; Hua, Y.; Zhu, X.X. A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 12416–12425.
11. Zhou, W.; Dong, S.; Lei, J.; Yu, L. MTANet: Multitask-Aware Network with Hierarchical Multimodal Fusion for RGB-T Urban Scene Understanding. *IEEE Trans. Intell. Veh.* **2022**. [CrossRef]
12. Zhou, W.; Guo, Q.; Lei, J.; Yu, L.; Hwang, J.-N. IRFR-Net: Interactive recursive feature-reshaping network for detecting salient objects in RGB-D images. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**. [CrossRef]
13. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
14. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
15. Vijay, B.; Alex, K.; Roberto, C. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495.
16. Jiang, J.; Zheng, L.; Luo, F.; Zhang, Z. Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation. *arXiv* **2018**, arXiv:1806.01054.
17. 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 European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 801–818.
18. Chen, X.; Lin, K.Y.; Wang, J.; Wu, W.; Qian, C.; Li, H.; Zeng, G. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 561–577.
19. Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; Sang, N. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *Int. J. Comput. Vis.* **2021**, *129*, 3051–3068. [CrossRef]
20. Xu, Z.; Zhang, W.; Zhang, T.; Li, J. HRCNet: High-resolution context extraction network for semantic segmentation of remote sensing images. *Remote Sens.* **2020**, *13*, 71. [CrossRef]
21. Zhou, W.; Yuan, J.; Lei, J.; Luo, T. TSNet: Three-stream self-attention network for RGB-D indoor semantic segmentation. *IEEE Intell. Syst.* **2020**, *36*, 73–78. [CrossRef]
22. Seichter, D.; Köhler, M.; Lewandowski, B.; Wengefeld, T.; Gross, H.M. Efficient rgb-d semantic segmentation for indoor scene analysis. In Proceedings of the IEEE International Conference on Robotics and Automation, Xi'an, China, 30 May–5 June 2021; pp. 13525–13531.
23. Hu, X.; Yang, K.; Fei, L.; Wang, K. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In Proceedings of the IEEE International Conference on Image Processing, Taipei, Taiwan, 22–25 September 2019; pp. 1440–1444.
24. Zhou, W.; Yu, L.; Zhou, Y.; Qiu, W.; Wu, M.; Luo, T. Local and global feature learning for blind quality evaluation of screen content and natural scene images. *IEEE Trans. Image Process.* **2018**, *27*, 2086–2095. [CrossRef]
25. Zhou, W.; Yang, E.; Lei, J.; Yu, L. FRNet: Feature Reconstruction Network for RGB-D Indoor Scene Parsing. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 677–687. [CrossRef]
26. Zhou, W.; Lin, X.; Lei, J.; Yu, L.; Hwang, J.-N. MFFENet: Multiscale feature fusion and enhancement network for RGB–Thermal urban road scene parsing. *IEEE Trans. Multimed.* **2022**, *24*, 2526–2538. [CrossRef]

27. Zheng, X.; Wu, X.; Huan, L.; He, W.; Zhang, H. A Gather-to-Guide Network for Remote Sensing Semantic Segmentation of RGB and Auxiliary Image. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–15. [CrossRef]
28. Ma, X.; Zhang, X.; Pun, M.O. A Crossmodal Multiscale Fusion Network for Semantic Segmentation of Remote Sensing Data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 3463–3474. [CrossRef]
29. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Las Vegas, NE, USA, 27–30 June 2016; pp. 770–778.
31. Zhou, W.; Liu, C.; Lei, J.; Yu, L.; Luo, T. HFNet: Hierarchical feedback network with multilevel atrous spatial pyramid pooling for RGB-D saliency detection. *Neurocomputing* **2022**, *490*, 347–357. [CrossRef]
32. Zhou, W.; Lv, Y.; Lei, J.; Yu, L. Global and Local-Contrast Guides Content-Aware Fusion for RGB-D Saliency Prediction. *IEEE Trans. Syst. Man Cybern. Syst.* **2021**, *51*, 3641–3649. [CrossRef]
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
34. Zhou, W.; Guo, Q.; Lei, J.; Yu, L.; Hwang, J.-N. ECFFNet: Effective and consistent feature fusion network for RGB-T salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 1224–1235. [CrossRef]
35. Li, G.; Liu, Z.; Zeng, D.; Lin, W.; Ling, H. Adjacent context coordination network for salient object detection in optical remote sensing images. *IEEE Trans. Cybern.* **2022**. [CrossRef]
36. Zhou, W.; Liu, C.; Lei, J.; Yu, L. RLLNet: A lightweight remaking learning network for saliency redetection on RGB-D images. *Sci. China Inf. Sci.* **2022**, *65*, 160107. [CrossRef]
37. Gong, T.; Zhou, W.; Qian, X.; Lei, J.; Yu, L. Global contextually guided lightweight network for RGB-thermal urban scene understanding. *Eng. Appl. Artif. Intell.* **2023**, *117*, 105510. [CrossRef]
38. Zhou, W.; Wu, J.; Lei, J.; Hwang, J.-N.; Yu, L. Salient object detection in stereoscopic 3D images using a deep convolutional residual autoencoder. *IEEE Trans. Multimed.* **2021**, *23*, 3388–3399. [CrossRef]
39. Zhou, W.; Zhu, Y.; Lei, J.; Wan, J.; Yu, L. CCAFNet: Crossflow and cross-scale adaptive fusion network for detecting salient objects in RGB-D images. *IEEE Trans. Multimed.* **2022**, *24*, 2192–2204. [CrossRef]
40. International Society for Photogrammetry and Remote Sensing. 2D Semantic Labeling Contest-Potsdam. Available online: https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx (accessed on 1 January 2020).
41. International Society for Photogrammetry and Remote Sensing. 2D Semantic Labeling Contest-Vaihingen. Available online: https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx (accessed on 1 January 2020).
42. 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]
43. Wang, L.; Li, R.; Duan, C.; Zhang, C.; Meng, X.; Fang, S. A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]