International Journal of
Geo-Information

Article

HA-MPPNet: Height Aware-Multi Path Parallel Network for High Spatial Resolution Remote Sensing Image Semantic Seg-Mentation

Suting Chen [1,2,*], Chaoqun Wu [1], Mithun Mukherjee [3] and Yujie Zheng [4]

[1] Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET), Nanjing University of Information Science & Technology, Nanjing 210044, China; 20191218011@nuist.edu.cn

[2] Wuxi Institute of Technology (NUIST-WIT), Nanjing University of Information Science & Technology, Wuxi 214100, China

[3] School of Artificial Intelligence, Nanjing University of Information Science & Technology, Nanjing 210044, China; mithun@nuist.edu.cn

[4] The 28th Research Institute of China Electronics Technology Group Corporation, Nanjing 210008, China; zhengyujie@cetc.com.cn

* Correspondence: sutingchen@nuist.edu.cn; Tel.: +86-139-1386-4015

check for updates

Citation: Chen, S.; Wu, C.; Mukherjee, M.; Zheng, Y. HA-MPPNet: Height Aware-Multi Path Parallel Network for High Spatial Resolution Remote Sensing Image Semantic Seg-Mentation. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 672. https://doi.org/10.3390/ijgi10100672

Academic Editor: Wolfgang Kainz

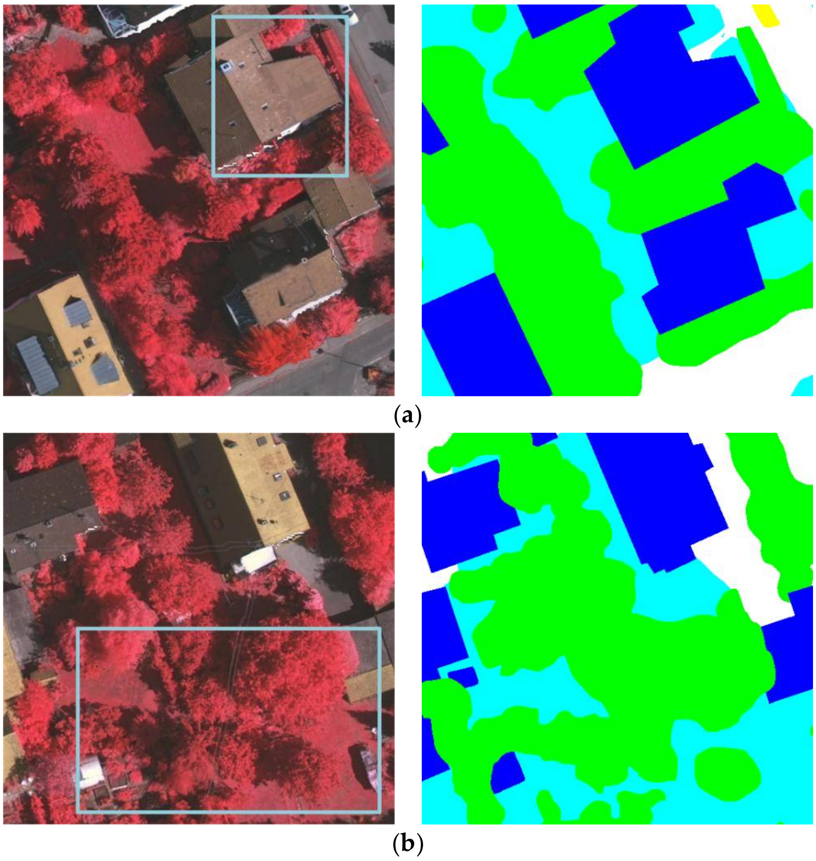
Received: 14 August 2021
Accepted: 1 October 2021
Published: 4 October 2021

Abstract: Semantic segmentation of remote sensing images (RSI) plays a significant role in urban management and land cover classification. Due to the richer spatial information in the RSI, existing convolutional neural network (CNN)-based methods cannot segment images accurately and lose some edge information of objects. In addition, recent studies have shown that leveraging additional 3D geometric data with 2D appearance is beneficial to distinguish the pixels' category. However, most of them require height maps as additional inputs, which severely limits their applications. To alleviate the above issues, we propose a height aware-multi path parallel network (HA-MPPNet). Our proposed MPPNet first obtains multi-level semantic features while maintaining the spatial resolution in each path for preserving detailed image information. Afterward, gated high-low level feature fusion is utilized to complement the lack of low-level semantics. Then, we designed the height feature decode branch to learn the height features under the supervision of digital surface model (DSM) images and used the learned embeddings to improve semantic context by height feature guide propagation. Note that our module does not need a DSM image as additional input after training and is end-to-end. Our method outperformed other state-of-the-art methods for semantic segmentation on publicly available remote sensing image datasets.

Keywords: remote sensing image; semantic segmentation; high spatial resolution; gated feature fusion; digital surface model (DSM); height features

1. Introduction

Attributed to the rapid development of satellite observation technology, a large number of high spatial resolution (HSR) remote sensing images can be easily acquired. Automatically extracting objects such as buildings, cars, and trees from remote sensing images is significant for land cover classification [1], urban management [2], and city planning [3]. Remote sensing image segmentation, as a crucial role in the field of handling remote sensing images, can predict the semantic category for every pixel in an input image. However, rich texture detail information, as shown in Figure 1a, and complicated scenes in the remote sensing image, cause difficulty in distinguishing the pixel category. Therefore, how to preserve the spatial information of remote sensing images and obtain the strengthened semantic context module during the process of segmentation is a challenging task.



(**a**)

(**b**)

Figure 1. Examples of rich spatial information and complicated senses in high resolution remote sensing image: (**a**) objects (such as buildings) have rich edge information and details and (**b**) low vegetation and trees are similar in 2D appearance, but belong to two different labels.

In recent years, convolutional neural networks (CNNs) have been successfully applied in the field of remote sensing images due to their excellent performance such as building extraction [4], object detection [5], image classification [6], and so on [7]. Evolving from CNN, the methods based on fully convolutional networks (FCNs) [8] have made great progress in semantic segmentation. For the extraction of semantic and recovery of spatial information, the structures [9,10] based on an encode–decode-network obtain semantic features by the downsampling operation in the encode stage and fuse shallow high-resolution features through the skip connections to recover spatial information in the decode stage. However, frequent downsampling operations lead to the loss of spatial information, which becomes the main obstacle in accurately extracting spatial information from remote sensing images. To solve this problem, DeepLabv1 [11] uses conditional random fields (CRF) for post-processing to optimize the segmentation of edges and Hierarchical [12] tries to enlarge the scale of the input image to obtain a high-resolution result, however, both of them increase the number of network calculations. Recently, HRNet [13] has proposed a high-resolution CNN to obtain semantic features while maintaining a high-resolution representation, but its high-level semantics are not rich.

In addition, context modeling is beneficial to improving the discrimination of pixels. In the field of natural imaging, PSPNet [14], ASPP [15], and DenseASPP [16] introduce a multi-scale context. At the same time, the self-attention-based methods [17,18] calculate a pixel-wise similarity map to capture long-range global context. Using these methods, the network can only learn 2D context appearance features, but for remote sensing images with more complex scenes, 3D geometric information is also essential [19,20]. Geometric features can discriminate those with similar 2D appearance, but significantly different 3D features objects such as trees and low vegetation, as shown in Figure 1b. Thus, a recent work [21] suggested that directly introducing height-related data (such as DSM) as an additional input to network and fuse the learned semantic features and height features to enhance the semantic performance. However, collecting the DSM images in a real-life application is not convenient and the obtained height maps usually do not align with remote sensing images, which limits their application.

In this paper, a height aware-multi path parallel network (HA-MPPNet) is proposed to alleviate the above problems. Instead of fusing shallow mappings to recover spatial information, we introduced the multi path parallel network (MPPNet) to learn multi-level features while fixing the spatial resolution in each path for preserving RSI detail and edge information. To enhance low-level semantic features, we designed a gated high-low-level feature fusion to fuse the selected features from both levels through a gate mechanism.

To utilize the DSM image to strengthen semantic context, we note that height features can be learned from a single remote sensing image [22]. Hence, a height aware context module was introduced by joining a new height feature decode branch to learn height features under the supervision of a ground-truth height map. Some recent works [23,24] have also introduced multi-task learning to simultaneously undertake semantic map prediction and height estimation, which is similar to our method. Unlike these methods, which decouple two tasks in the top layers of decoder networks, our method designs two specific decoder branches. More importantly, we introduce a height feature guide propagation module to use the learned high feature as an affinity guide to effectively fuse with semantic context to improve the performance. Finally, our context module does not require a DSM image as additional supervision after training, and can directly generate the segmentation result for the test image in an end-to-end fashion.

In this paper, we aimed to preserve the spatial information of remote sensing images and utilize DSM images to strengthen the semantic context during the process of segmentation.

2. Related Work

2.1. Semantic Segmentation

A fully convolution network (FCN) [8] directly outputs the pixel-wise prediction from the input image with an arbitrary size via the network upsampling layer, making a breakthrough in the field of semantic segmentation. Then, encode–decode structures [9,10] extract semantic features by downsampling operations in the encode stage and restore the image resolution via upsampling layers in the decode stage. DeepLabv1 [11] introduces dilated convolution to obtain a larger respective field of feature maps and uses conditional random fields (CRF) as a post-process to refine the segmentation results. To capture multi-scale feature maps, ASPP [15] uses multi parallel atrous convolutions with different dilation rates while PSPNet [14] generates pyramid feature maps by the pyramid pooling module (PPM). PSANet [18] and CCNet [25] use a non-local [17] style to calculate a pixel-wise similarity map on the whole image to obtain a long-range global context. However, this context is only obtained by spatial attention, thus CBAM [26] and DANet [27] introduce channel attention to capture the channel dependencies between any two channel feature maps. Recent works on these aspects [12,13] have focused on paying attention to propose a high resolution CNN to obtain a high resolution result with precise detail and edge information of the image.

Semantic segmentation in high resolution remote sensing images also significantly benefits from the improvement of deep learning methods. Guo et al. [28] introduced FCN with atrous convolutions to segment remote sensing images and use CRF as a post-process to smooth the prediction results. Diakogiannis et al. [29] proposed a network by integrating residual connections based on U-Net and refining dice loss function to obtain accurate results. Wang et al. [30] utilized the ResNet-101 [31] as the backbone to extract high-level semantic feature maps and constructed a fully convolutional network that adaptively fuses multi-scale features. Marmanis et al. [32] tried to capture the edge details of segmentation objects to further finetune the semantic object boundaries. Afterward, SCAttNet [33] proposes spatial and channel attention to capture the context of every pixel. HMANet [34] introduces a class channel attention to compute class based correlation and recalibrate the category level information. Some researchers [35] have tried to apply weak supervision techniques to the field of remote sensing image segmentation.

2.2. Multi-Level Feature Fusion

Feature fusion is frequently employed in semantic segmentation to utilize different level features. U-Net [10] adds a skip connection between the encode and the decode stage to reuse the shallow feature maps to enhance the high-level features with strong semantics in spatial detail. In contrast, ExFuse [36] embeds high-level features into low-level features to enhance low-level semantics. FPN [37] introduces the method of lateral connections to obtain multi-levels of prediction. However, these methods all adopt simple pixel addition or channel concatenation operations to directly fuse multi-level features without measuring the effectiveness of information in all feature maps, which limits the propagation of useful features. To alleviate this issue, recent work [38] on natural language processing (NLP) has proposed the idea of using gate mechanisms to control the information flow. Inspired by these, Gated-SCNN [39] proposes the use of gate mechanisms to define information flow between the regular semantic stream and another shape stream to capture precise boundary information. Note that GFF [40] uses gates to fully fuse features from every level feature map, but the fused feature maps of each level are mainly dependent on the value of the gates of the current level. In our work, we introduced a gated high–low-level feature fusion to adaptively select useful information from a high–low-level via a gate mechanism and gradually perform feature fusion by a bottom-up pathway.

2.3. Height Estimation for Remote Sensing Image

For semantic segmentation of remote sensing images, 3D geometric information can be used to enhance segmentation performance. Some works have shown how the estimated DSM data providing useful additional information for building detection [19] or semantic segmentation [20]. For example, V-FuseNet [21] uses the DSM image as additional input for the segmentation task, which simultaneously learns RGB and height features through a two-stream network and then fuses the learned features from the two encode networks for final prediction. The later works [41,42] focused on designing a stronger encode network structure or fusing two features at different locations of the network to improve segmentation performance. However, these approaches all need the ground-truth height map as additional input, which severely limits their applications. It is worth noting that Lichao et al. [22] proved height features were naturally preserved by the remote sensing image, where the authors trained an encode–decode network to predict height map from a single remote sensing image. On this basis, Srivastava et al. [23] proposed a multi-task CNN, one for semantic prediction, and the other for height estimation, where the key of this work is that DSM data are not required during testing time. Volpi et al. [24] considered that both semantic map prediction and DSM regression required specific model layers, so a more suitable strategy was designed to perform middle splitting the multi-task architecture to divide it into two specific task branches. For the task of RGB–depth semantic segmentation, which is similar to height estimation, Wang et al. [43] proposed a joint framework to predict both depth and semantic maps, followed by a hierarchical CRF for post-processing to optimize the pixel-level depth estimations, but did not research the information sharing between different tasks.

3. Methods

To preserve the spatial information lost in the process of CNN downsampling and utilize geometric information to improve the discrimination of the pixel, we proposed a height aware-multi path parallel network (HA-MPPNet), as shown in Figure 2. Our proposed network consisted of the multi path parallel network (MPP-Net) in Section 3.1; the gated high-low level feature fusion in Section 3.2; and height aware context module

(HAC) in Section 3.3. In Section 3.4, we introduce the multi-task loss function to train the network.

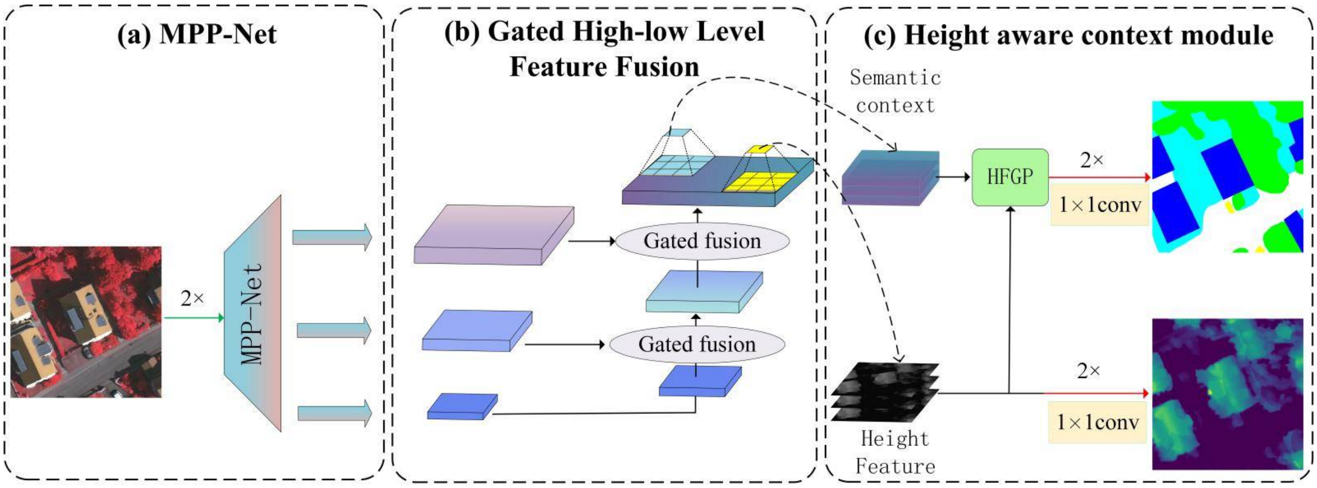


Figure 2. The pipeline of our network, the HSR remote sensing image, first sends two consecutive downsampling stems to extract feature maps and reduce its resolution, and feeds them into a MPP-Net to obtain multi-scale semantic feature maps x_1, x_2 and x_3. Adjacent high–low level feature maps adaptively select useful information from themselves via a gate mechanism to gradually perform feature fusion. Then, the fused lowest-level feature map, x_1 learns the separate semantic context and height feature, respectively, by two different 3 × 3 convolutions. Finally, the learned height features are used to improve the quality of semantic context by the height feature guide propagation (HFGP) module to realize height aware semantic context. 2× means two consecutive downsampling or upsampling operations.

3.1. Multi Path Parallel Network

Rather than recovering spatial information by the skip connections and fusing shallow mappings in traditional encoder–decoder-based networks, our multi path parallel network (MPPNet) can capture high-level semantics in the downsampling layer while preserving high-resolution remote sensing image edge and detail information. As shown in Figure 3, we first put the remote sensing image into two consecutive downsampling stems, each of which was composed of a 3 × 3 convolution followed by batch normalization (BN) and a rectified linear unit (ReLU) activation function to extract semantic feature maps with 64 channels (especially for two consecutive downsamping stems) and a 2 × 2 max pooling to downsample its resolution to 1/4 of the original remote sensing image. In particular, this operation aims to avoid issues related to the exhaustion of computing resources when we directly inputted the original 512 × 512 pixel image into the deep network.

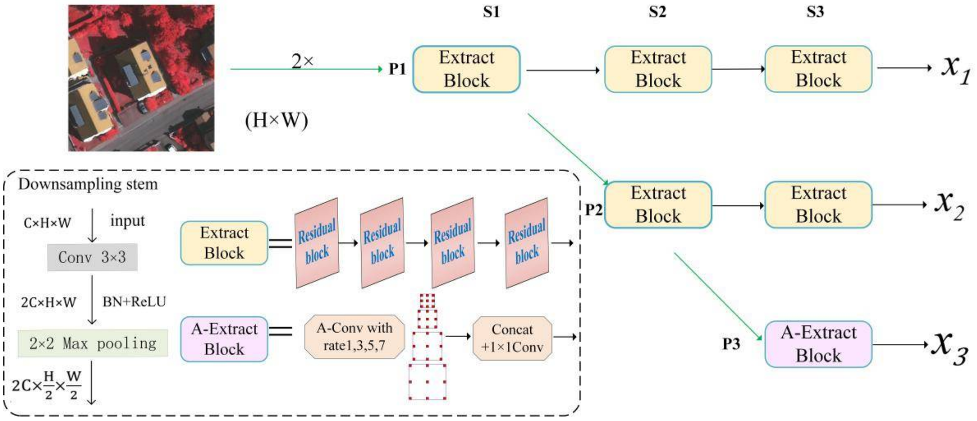


Figure 3. Illustration of our multi path parallel network (MPP-Net), the green arrow is the operation of downsampling stem and the dilation rates of 3 × 3 atrous convolutions are 1, 3, 5, and 7.

We divided MPP-Net into three stages (S1, S2, and S3) and three parallel paths (P1, P2, and P3) to consider the complexity and efficiency of the network as analyzed in the experimental setup described in Section 4.2.2. In each stage, we sent the feature maps obtained from the previous sage into a feature extraction block to extract richer semantics with the condition of a fixed spatial resolution. At the end of the stage, a new parallel path is generated through a downsampling stem to capture high-level semantic features with a double downsampling resolution and double channels, as shown in the dashed box of Figure 3. The feature maps extracted by each path maintain the resolution, thereby retaining more edge and detail information of the object in the remote sensing image. In each path, channels and resolutions of feature maps are both fixed during the entire process of feature extraction. The feature extraction block is composed of a series of cascade residual blocks [31]. The impact of a different number of residual blocks on MPPNet performance is explained in the experimental setup in Section 4.2.2.

We note that the extract block of P3 was replaced with an A-extract block, which consisted of four parallel atrous convolutions with different dilation rates, as shown in the dashed box of Figure 3. This operation aims to integrate a much more receptive-field of the high-level and enhance the high-level feature semantics. Next, the extracted feature maps were channel-wise concatenated to reuse multi-receptive-field feature maps and sent into a 1 × 1 convolution to restore input channel dimension. Finally, our MPP-Net obtained three scale feature maps x_1, x_2 and x_3, where the spatial resolutions were 1/4, 1/8, and 1/16 of the original image, and the corresponding numbers of channels were 64, 128, and 256, respectively.

3.2. Gated High-Low Level Feature Fusion

Low-level feature maps near the input image had a high resolution remote sensing image, but their semantic information was much fewer than the high-level feature maps. The latter obtained richer semantic information and a much more receptive-field from the extraction of deep network and these semantics can assist in predicting most pixels of larger objects in remote sensing images, but are limited by resolution. Therefore, it is natural to consider complementing the advantages of multi-levels by fusing high-level feature maps to low-level features, in order to realize a fused feature map with both high resolution and rich semantic information. In most cases of high-low level feature fusion, bilinear interpolation and 1 × 1 convolution are first used to restore the high-level feature maps to the same resolution and channel number of the low-level feature maps, and then adopt addition or concatenation operation to fuse high-low level feature. Among them, the method of addition adds their features at each pixel position and the method of concatenation concatenates two feature maps along their channel dimension. However, the problem of both fusion methods is that useful information is mixed with massive amounts of useless information without selection during the fusion process, which increases the amount of calculation and reduces the segmentation performance.

To address this issue, we proposed a gated high–low level feature fusion to adaptively fuse useful information from each layer, as shown in Figure 4. In particular, we first reshaped the high-level feature map $x_{l+1} \in R^{2C\times\frac{H}{2}\times\frac{W}{2}}$ with the same spatial resolution size and channel number as a low-level feature map $x_l \in R^{C\times H\times W}$ through bilinear interpolation, followed by 1 × 1 convolution, and l is the l-th level feature map. Then, x_l and reshaped $\hat{x}_{l+1}$ are concatenated (| |), and followed by a 1 × 1 convolutional layer $C_{1\times1}$ with the output channels reduced to C. Then, we can obtain a gate map $G \in R^{C\times H\times W}$ by the ψ project function as follows:

$$G = \psi\left(C_{1\times1}\left(\hat{x}_{l+1}\middle|\middle|x_l\right)\right) = \sigma\left(C_{3\times3}\left(\mathrm{SE}\left(C_{1\times1}\left(\hat{x}_{l+1}\middle|\middle|x_l\right)\right)\right)\right) \tag{1}$$

where SE is a squeeze-and-excitation block [44] implementing channel attention and σ denotes the sigmoid operation. Intuitively, G can be seen as a selective map, which decides the fusing weight between high-level features and low-level features. Then, the gated fusion of adjacent high–low level based on pixel-wise addition can be calculated as shown in:

$$\widetilde{x_l} = G \otimes x_l + (1 - G)\hat{x}_{l+1} \tag{2}$$

where $\widetilde{x_l}$ is the fused feature map; and $\otimes$ denotes the Hadamard product in the channel dimension. The fused $\widetilde{x_l}$ continues to re-fuse with lower level feature maps and gradually obtains the final fused feature map from bottom to top. After using the selecting function of a gate mechanism, the final fused feature map $\widetilde{x_1}$ obtains both high resolution and rich semantics with the less computational cost.

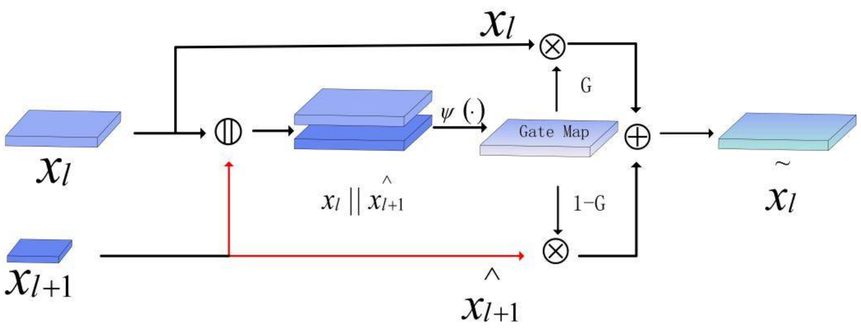

x_l
x_l
$\psi(\cdot)$
G
Gate Map
$x_l \| \hat{x}_{l+1}$
1−G
$\widetilde{x_l}$
x_{l+1}
$\hat{x}_{l+1}$

Figure 4. Computation details of our proposed gated high–low-level feature fusion. The red arrow denotes the operations of bilinear interpolation followed by a 1 × 1 convolution.

3.3. Height Aware Context Module

In the remote sensing image, the scene is much more complicated than the natural image. This means that there is larger intra-class variance and smaller inter-class variance, which causes the problem of false alarms. The introduction of semantic context alone cannot effectively distinguish objects with similar 2D semantic appearance, but different 3D height features in remote sensing images (such as trees and low vegetation). To alleviate this problem, a height aware context module is proposed to improve discrimination of the pixel by learning height features as affinity guidance of semantic context. As shown in Figure 2c, we designed a new height feature decode branch, where the ground truth of the height map was used as a label to guide the learning of height features during the training process. Then, the learned height features were used as an affinity guide to fuse with the semantic context by height feature guide propagation (HFGP), as shown in Figure 5.

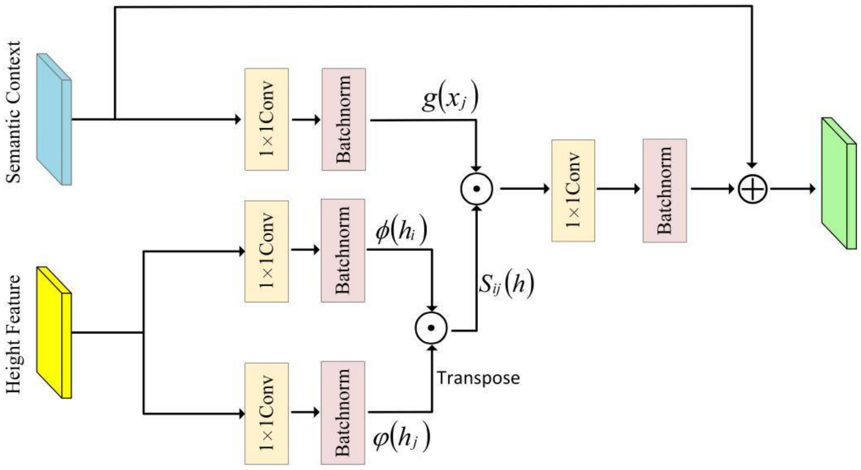


Figure 5. Illustration of our height feature guide propagation.

Concretely, for the fused feature map, $\widetilde{x_1}$ is sent into two different 3 × 3 convolutions followed with BN and ReLU to learn the independent semantic context x and height features h, respectively. We processed the learned height embeddings into two sub-embeddings with ϕ and φ project functions. Then, we calculated the height similarity matrix S_{ij} of two height sub-embeddings by dot-product as follows:

$$S_{ij}(h) = \phi(h_i)^T \cdot \varphi(h_j) \tag{3}$$

where i and j are the location of pixel; T is the operation of matrix transpose, and respectively, ϕ and φ are just implemented by a 1 × 1 convolutional layer followed by batch normalization. Then, the produced height similarity S_{ij} is used as an affinity guide to fuse height features and semantic context by another dot-product. Finally, the original semantic context is added to the obtained result to avoid interruption during the whole propagation. Note that the whole propagation process maintains the size and dimension of the semantic features. The height aware semantic context propagation output y_i at location i can be calculated as:

$$y_i = \frac{1}{R}\sum_{j}^{N}\left(\left(S_{ij}(h) \cdot g(x_j)\right)\right) + x_i \tag{4}$$

where N is the number of pixels; g is the implementation of a 1 × 1 convolutional layer followed by batch normalization and to project the semantic feature to deal with the dimension variation during the propagation. The normalization factor R is set as $R = \sum_j^N S_{ij}$.

Compared to other work [21,41,42], by using both HSR remote sensing images and height maps as inputs for a two-stream network, our introduced height aware context module does not require DSM images as additional inputs after training and can directly generate the segmentation maps for the test images in an end-to-end fashion.

3.4. Multi-Task Loss Function

In our work, we introduce both semantic segmentation training and height feature training, hence the joint loss function is defined as:

$$L = L_{seg} + L_h \tag{5}$$

where L_{seg} denotes semantic segmentation loss and L_h denotes height feature loss.

For semantic segmentation learning, most existing methods adopt cross-entropy loss function to measure the difference between the predicted semantic map and ground truth label. However, the problem exists in high resolution remote sensing images that some classes (buildings, surface, and so on) have more samples than the car class, which causes the predicted accuracy of few samples to have low accuracy. Motivated by the proposed focal loss [45] in the field of image object detection, we applied it and set the focusing parameter γ = 2 in our segmentation learning, so our segmentation loss function can be formulated as:

$$L_{seg} = -\sum_i\sum_c (1 - p_{ic})^2 \times l_i \times \log(p_{ic}) \tag{6}$$

where i is the pixel location; c is the pixel class index; p_{ic} denotes the predicted probability of the i-th pixel belonging to class c; and l_i corresponds to its ground truth. This loss function can promote the loss contribution of hard samples and suppress the loss of easy training samples. For example, if a pixel is predicted correctly with p = 0.8, then the weight of the pixel loss value is 0.04; if a pixel is predicted incorrectly with p = 0.1, then the weight of the pixel loss value is 0.81.

For height features learning, we adopted Smooth L1 loss for our height feature supervision as follows:

$$L_h = \begin{cases} \sum_i 0.5(h_i - h_i{}^*)^2 & \text{if} |h_i - h_i{}^*| < 1 \\ \sum_i |h_i - h_i{}^*| - 0.5 & \text{otherwise} \end{cases} \tag{7}$$

where h_i and $h_i{}^*$ denote the predicted and ground truth height values at pixel i. This loss function can avoid the issue of gradient explosion when the value of $|h_i - h_i{}^*|$ is large and strengthen the robustness of the model. At the same time, a low $|h_i - h_i{}^*|$ value will cause the L_h to be small, which can accelerate the convergence model.

4. Experiment

4.1. Dataset and Metric

To evaluate our method, we conducted experiments on two publicly available remote sensing image datasets from two cities in Germany: Vaihingen and Potsdam. In both datasets, each ground truth label provides six classes for corresponding images: impervious surfaces (e.g., roads), cars, trees, low vegetation, buildings, and clutter.

The ISPRS Vaihingen dataset has a spatial resolution of 9 cm, with an average size of 2500 × 2100 pixels. There were 33 high resolution remote sensing images with three spectral bands of infrared, red, green (IRRG) and related DSM images. Following the standard split, we used 16 titles providing ground truth for model training and the remaining 17 tiles were used for model testing.

The ISPRS Potsdam dataset has a spatial resolution of 5 cm, with an average size of 6000 × 6000 pixels. There were 38 high resolution remote sensing images with four spectral bands infrared, red, green, blue (IRRGB) and related DSM images. Following the standard split, we used 24 titles providing ground truth for model training and the remaining 14 titles were used for model testing.

We employed the commonly used metrics including overall accuracy (*OA*) and intersection over union (*IoU*) to evaluate the performance of our method, defined as follows:

$$OA = \frac{TP + FN}{N} \tag{8}$$

$$IoU = \frac{TP}{FP + TP + FN} \tag{9}$$

where TP, FP, and FN represent the number of true positive pixels, false positive pixels, false negative pixels, respectively, and N is the number of pixels.

4.2. Implement Details and Experimental Setup

4.2.1. Implement Details

Our research was implemented on the TensorFlow platform, and run by an NVIDA Titan-V GPU with 12 Gigabyte of memory. We selected the Adam solver, with beta1 and beta2 set to default as recommended, in order to optimize the network. The initial learning rate was set to 0.001 for all datasets and we trained the model for 80 epochs. Considering the limit of the GPU memory, the batch size was set to 4.

As the titles of both datasets were very high resolution, we could not directly process them in our network for training and testing. As a result, we used random cropping to crop each large image to a size of 512 × 512 pixels to fit the GPU memory. At the same time, the number of training samples was not adequate for training, so we used rotation, mirroring, adding noise, and other data augmentation methods to expand the dataset to better train the model. Note that we used DSM images as additional labels for our network during the training time and did not require them during the testing time.

4.2.2. Experimental Setup

In general, the accuracy and the parameters of the network are mainly affected by the structure and depth of the network. In this section, we conducted an experiment on the Vaihingen dataset to research the influence of the number of paths and residual blocks on the performance of MPPNet. We set the number of paths from two to four, and the number of residual blocks from three to six according to the experience. The metrics mIoU and params were used to measure the accuracy and complexity of the network respectively.

The experimental results are shown in Figure 6. On one hand, as the number of residual blocks increased, the mIoU score reached the highest when the number was 5 and then decreased. This may be due to the fact that the generalization ability of the network decreased with the increase in depth and parameters and the network params maintained linear growth. On the other hand, when the number of paths increased, the mIoU score also increased, while the params of the network increased exponentially. We note that when there were three paths, the mIoU score was slightly less than the highest score of four, but the cost params were much lower than the latter.

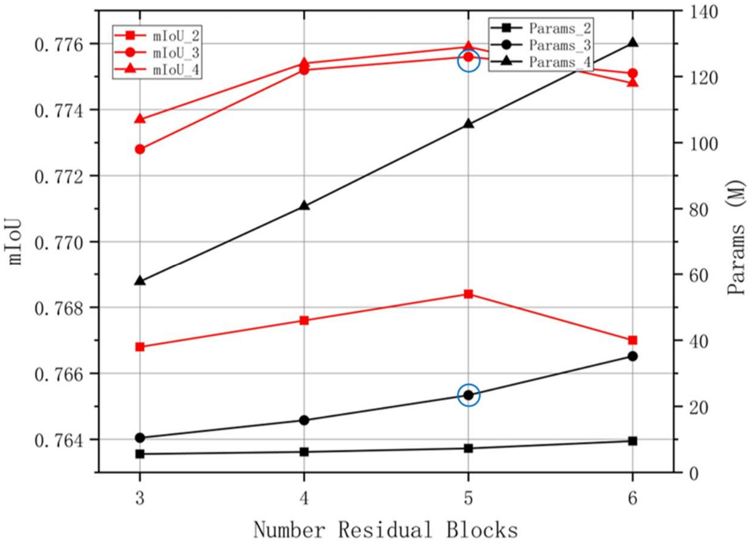


Figure 6. Performance of our MPPNet with different numbers of paths and residual blocks. Red and black lines represent the mIoU result and model params. The rectangle, circle, and triangle patterns on the line represent the number of network paths 2, 3, and 4, respectively. The horizontal axis represents the different number of residual blocks in each extracted block.

In order to balance the accuracy and complexity of the network, we chose three parallel paths and five residual blocks of each extract block to construct our MPPNet as shown by circles marked with blue, which saves a large amount of network params at the expense of losing a smaller accuracy cost.

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

4.3.1. Results on Vaihingen Dataset

We compared our proposed method with state-of-the-art methods through the experiment on the public Vaihingen dataset and the numerical comparison results are shown in Table 1, with the best performances marked in bold. As indicated in Table 1, our method outperformed all of the compared methods with the highest OA of 91.54% and mean intersection over union (mIoU) of 82.81%. In addition, when analyzing the results on some categories in the IoU form as shown in Table 1, we can see that our method performed better than the other methods in most categories. In particular, in the hard category (such as car), our method still achieved a higher IoU due to the newly introduced multi-task loss function. It can be seen that HRNet [13] achieved the highest result on buildings, due to the high resolution result the network obtained. At the same time, V-FuseNet [21] used the DSM image as the additional input of the network and was obviously better than the traditional IRRG input method, which further verifies the effectiveness of the height features. However, unlike this one, we used the DSM image as additional supervision of the model training and did not require it during the test time, which is convenient for real-life applications.

Table 1. Comparisons with state-of-the-art methods on the Vaihingen dataset. The accuracy of each category is presented in the IoU form. DSM(s) is the model using the DSM image as additional supervision.

Model	Input	Imp. Surf	Building	Low. Veg	Tree	Car	OA (%)	mIoU(%)
U-Net	IRRG	76.59	78.46	71.82	72.94	62.69	84.35	72.50
PSPNet	IRRG	77.45	79.51	73.04	75.01	64.58	86.07	73.92
HMANet	IRRG	78.34	80.21	78.18	78.57	64.12	87.19	75.88
HRNet	IRRG	79.23	**84.35**	75.86	77.42	67.62	87.96	76.90
V-FuseNet	IRRG+DSM	82.57	82.86	**83.94**	82.45	73.32	90.12	81.02
HA-MPPNet	IRRG+DSM(S)	**83.46**	83.85	83.27	**84.68**	**78.79**	**91.54**	**82.81**

We also selected some samples from the experimental results to show the comparison of the semantic results of our model with other methods on the Vaihingen dataset in Figure 7. It is clear that our method achieved more satisfactory results than the other methods. For objects (such as buildings) with richer details and edges, our method could well preserve this information during the segmentation process. In particular, we used a red dashed box to mark challenging areas that are easy to misclassify. It can be seen from the marked areas in Figure 7 that our method was superior to other methods in distinguishing between trees and low vegetation. This is because our height aware context module uses height features to improve the distinction between objects with a 2D appearance, but different geometric characteristics.

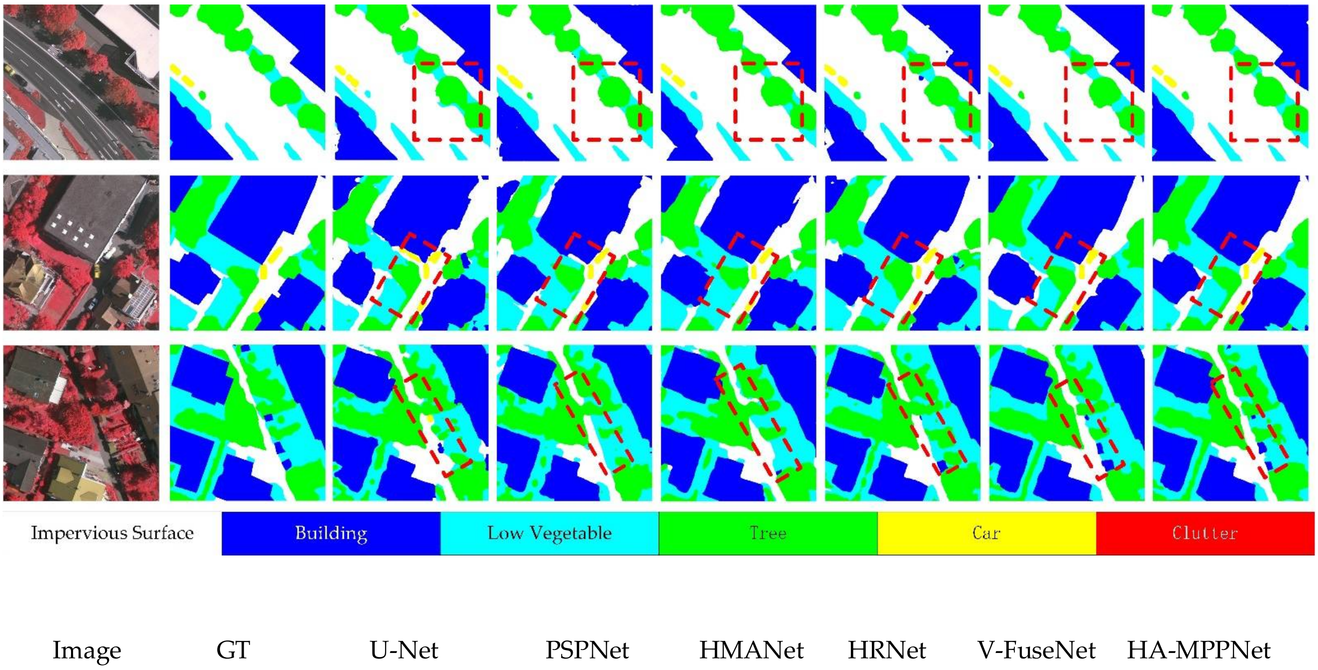


Figure 7. Selected samples of the segmentation maps on the Vaihingen dataset.

4.3.2. Results on the Postam Dataset

We also conducted the experiment on the Postam dataset to further evaluate our proposed method and adopted the same training and testing settings on the Potsdam dataset. We report the results of the numerical comparisons of our method and state-of-the-art methods as shown in Table 2, and the best performances aree marked in bold. It is noteworthy that our model obtained the highest OA of 90.21% and mIoU of 80.51%. At the same time, in the prediction results of each category, buildings with richer spatial information, trees, and low vegetation with similar 2D appearance, our method also achieved a higher mIoU. In addition, selected samples of the segmentation result maps are shown in Figure 8. We also used the red dashed box to mark challenging areas that are easy to misclassify. These areas better verified the predictive ability of our model and shows that our proposed method produced better segmentation maps than the other methods.

Table 2. Comparisons with state-of-the-art methods on the Postam dataset. The accuracy of each category is presented in the IoU form. DSM(s) is the model using DSM images as additional supervision.

Model	Input	Imp. Surf	Building	Low. Veg	Tree	Car	OA (%)	mIoU(%)
U-Net	IRRGB	75.68	77.59	70.76	72.49	61.93	83.82	71.69
PSPNet	IRRGB	77.05	78.32	72.83	74.41	64.05	85.24	73.33
HMANet	IRRGB	78.04	79.28	76.56	78.85	65.39	86.56	75.62
HRNet	IRRGB	78.75	**83.89**	75.20	76.54	66.23	87.43	76.12
V-FuseNet	IRRGB+DSM	81.38	82.96	**81.96**	81.25	69.56	89.79	79.42
HA-MPPNet	IRRGB+DSM(S)	**82.85**	83.43	81.63	**83.13**	**71.52**	**90.21**	**80.51**

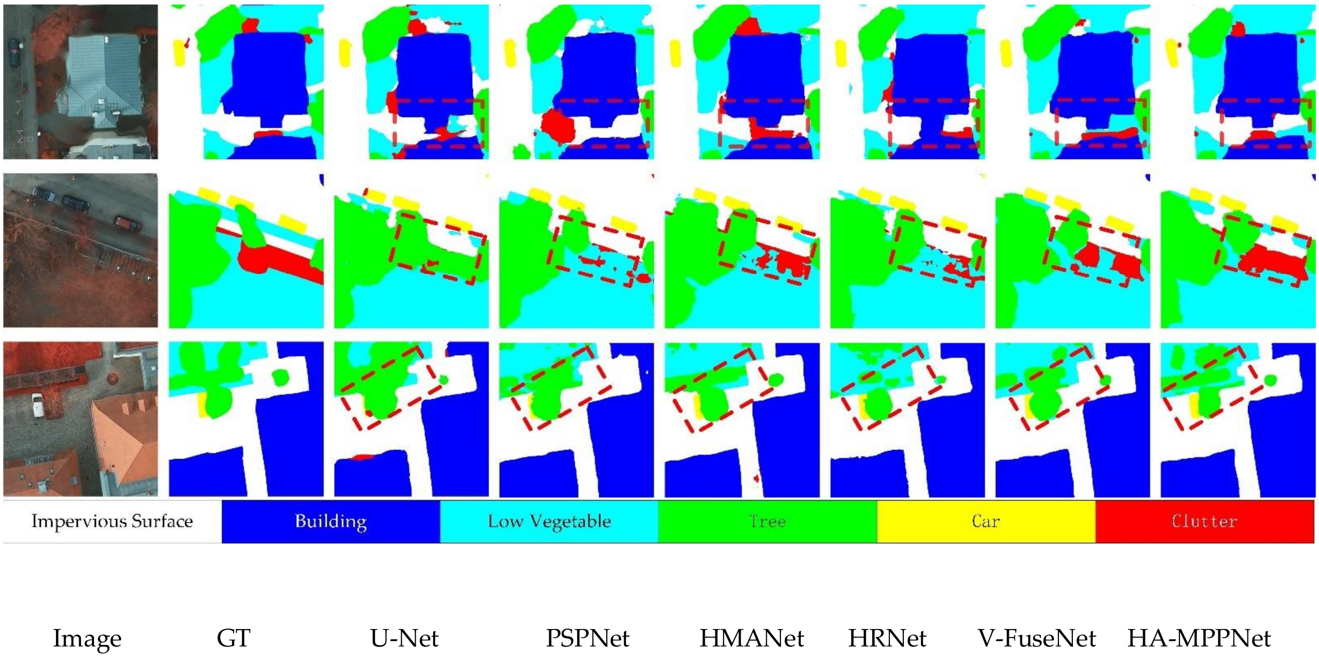


Figure 8. Selected samples of the segmentation maps on the Postam dataset.

4.4. Ablation Study

In order to verify the effectiveness of each module in our proposed network, we carried out an ablation study on the Vaihingen dataset. All the training and testing environments of each ablative experiment were kept the same. In our ablation study, we designed different variants of our HA-MPPNet by replacing or removing four key modules of the network. We constructed our baseline as follows: (1) we replaced the MPPNet with ResNet-101 [31] as the feature extraction for the network; (2) we replaced the GHLF module with direct channel concatenation followed by a 1 × 1 convolution for feature fusion; (3) we removed the height features decode (HFD) branch and used traditional IRRG as the network input; and (4) we replaced the HFGP module with direct concatenation for fusing learned height features and semantic context. Then, we gradually replaced or added the four key modules to the baseline, and provide the visualization and quantitative results in Figure 8 and Table 3.

Table 3. Quantitative evaluation of ablation studies on the Vaihingen dataset.

Module					OA (%)	mIoU(%)
Baseline	**MPPNet**	**GHLF**	**HFD**	**HFGP**		
√					85.73	73.64
√	√				87.86	77.52
√	√	√			88.95	78.85
√	√	√	√		90.28	81.67
√	√	√	√	√	91.54	82.81

As shown in Figure 9, we can see that the baseline can roughly segment the image but loses partial information of the object. By using MPPNet for feature extraction, which preserves spatial information of images, buildings and others with complex texture objects are well predicted while increasing the percentage gains of OA 2.13% and mIoU 3.88% on the Vaihingen dataset. Then, we introduced the GHLF module, which selectively fuses features from high and low levels by using a gate mechanism, and contributes to gains of 1.09% and 1.33% in terms of OA and mIoU. Next, utilizing the HFD module, which incorporates height features, enabled a boost in the performance considerably, which further revealed the effectiveness of height information. It was obviously seen that objects with similar 2D appearance but completely different height features such as trees and low vegetation were well distinguished, and this module improved the percentage gains of OA 1.33% and mIoU 2.82%. Finally, we introduced the HFGP module to use the learned height feature as an affinity guide for semantic features, which further improved the whole HA-MPPNet in two metrics.

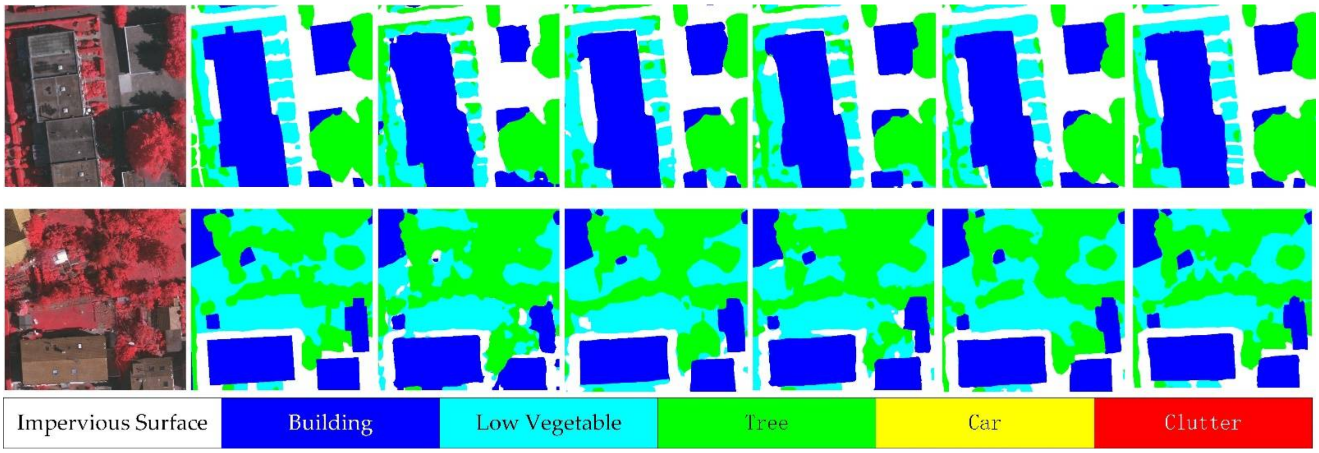


Figure 9. Visualization of ablative results on the Vaihingen dataset. From left to right: input IRRG, ground truth, segmentation maps predicted by baseline, Baseline + MPPNet, Baseline + MPPNet + GHLF, Baseline + MPPNet + GHLF + HED, and the full HA-MPPNet.

In addition, we also counted the computational cost of MPPNet with different feature fusion operations to further verify the effectiveness of our GHLF, and the quantitative results are shown in Table 4. It can be seen that compared with MPPNet + Contact/Addition using the traditional channel contact or pixel addition feature fusion method, our MPPNET + GHLF module required less GPU memory usage and had less computational cost with a few parameters. Meanwhile, our MPPNET + GHLF module still achieved a higher value of mIoU than the other fusion operations and reached 78.85% in the experiment of the Vaihingen dataset, which we attributed to the GHLF module that adaptively selects effective information from each layer and then performs feature fusion.

Table 4. Computational cost comparison of MPPNet with different feature fusion operations.

Module	mIo U(%)	Memory (MB)	Params (M)
MPPNet + Addition	77.68	256	33.1
MPPNet + Contact	77.52	238	32.8
MPPNet + GHLF	78.85	185	29.3

4.5. Visualization Results

To explore the feasibility of our proposed method for large-scale images, we used the large-scale titles of the Vaihingen and Postam datasets as inputs for the HA-MPPNet. We also selected samples of segmentation results, as shown in Figures 10 and 11. It can clearly be seen that our module could also predict excellent segmentation maps for large-scale images, which confirms the generalization of our network.

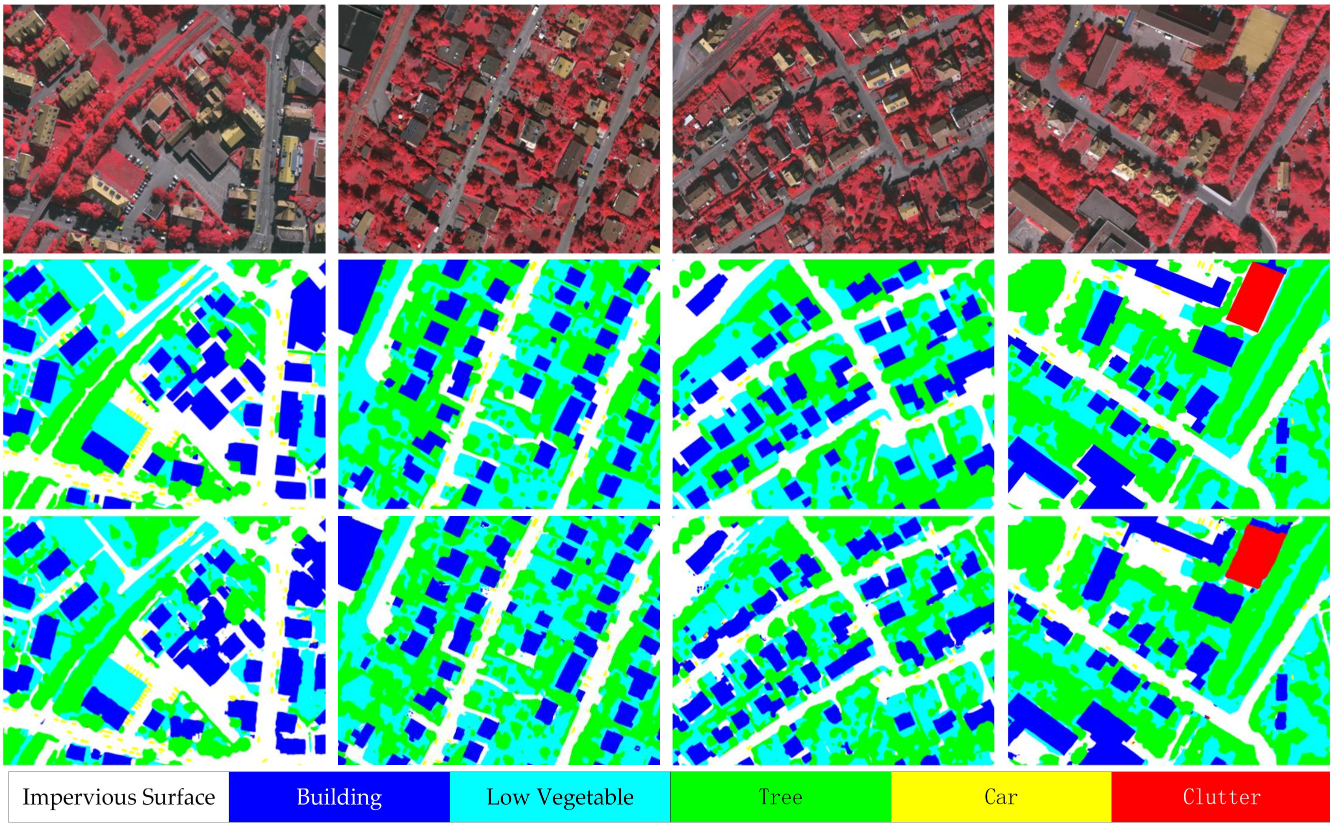


Figure 10. Results for large-scale images on the Vaihingen dataset. From top to bottom: large-scale image, ground truth, and segmentation maps predicted by HA-MPPNet.

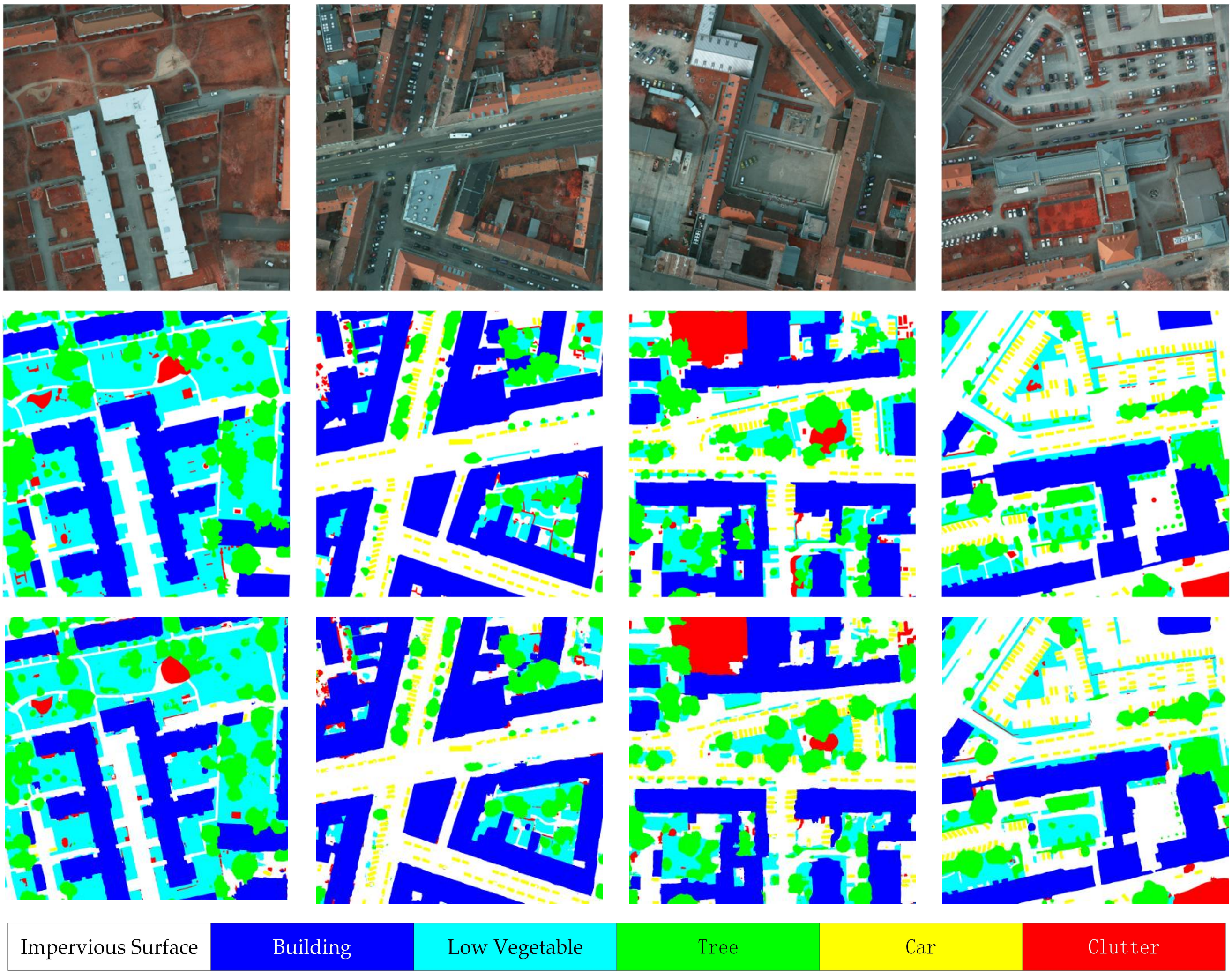


Figure 11. Results for large-scale images on the Postam dataset. From top to bottom: large-scale image, ground truth, and segmentation maps predicted by HA-MPPNet.

4.6. Discussion

As shown in the results, our method could well preserve the buildings with richer details and edges as the MPPNet maintains the resolution during the progress of the extraction of the feature. Our method was superior to the other methods in distinguishing between trees and low vegetation with a similar 2D appearance. This is because our height aware context module uses height features to improve the distinction between objects with 2D appearance, but different geometric characteristics. More importantly, our network does not require DSM images as an additional input after training and can directly generate the segmentation map for the test image. In addition, the ablation study evaluates the effectiveness of the proposed module. Gated feature fusion module adaptively selects effective information from each layer to obtain a higher accuracy with less computational cost than other feature fusion operations. Experiments have demonstrated the potentials of applying HA-MPPNet on high-resolution remote sensing image semantic segmentation.

5. Conclusions

In this paper, we proposed a height aware-multi path parallel network to alleviate the problem of the loss of rich detail information and indistinguishable pixel categories in semantic segmentation of high spatial resolution remote sensing images. The first contribution to our work was to introduce a multi path parallel network (MPPNet) to learn the multi-level semantic features while fixing the spatial resolution in each path to preserve the detail and edge information of images. Second, the gated high-low-level feature fusion (GHLF) module was used to fuse the selected features from both high–low levels by a gate mechanism and gradually enhanced low-level feature semantics. Then, we designed a height feature decode branch (HED) to learn the height features under the supervision of the DSM image. Followed by a height feature guide propagation (HFGP) architecture, the learned height embeddings were used as guidance to improve the semantic context. Our work only uses the DSM image as a side supervision for the network during the training stage, does not require it during the inference stage, and is an end-to-end fashion. Finally, the experiments on both the Vaihingen dataset and Postam dataset demonstrated that HA-MPPNet outperformed other state-of-the-art methods with a higher mIoU. In addition, we conducted an ablation study to further evaluate the effectiveness of the proposed module.

Author Contributions: Conceptualization, Suting Chen and Chaoqun Wu.; Methodology, Chaoqun Wu; Data curation, Chaoqun Wu.; Writing—original draft preparation, Chaoqun Wu; Writing—review and editing, Suting Chen, Mithun Mukherjee, and Yujie Zheng. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Nos. 61906097, 41875184) and a project funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://www2.isprs.org/commissions/comm2/wg4/benchmark/semantic-labeling/, (accessed on 28 October 2020).

Acknowledgments: The authors thank all the reviewers for their comments, which improved the quality of the paper.

Conflicts of Interest: The authors declare that there are no conflict of interest regarding the publication of this article.

References

1. Hu, B.; Xu, Y.; Huang, X.; Cheng, Q.; Ding, Q.; Bai, L.; Li, Y. Improving Urban Land Cover Classification with Combined Use of Sentinel-2 and Sentinel-1 Imagery. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 533. [CrossRef]
2. Kampffmeyer, M.C.; Salberg, A.; Jenssen, R. Semantic Segmentation of Small Objects and Modeling of Uncertainty in Urban Remote Sensing Images Using Deep Convolutional Neural Networks. Paper presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 26 June–1 July 2016. [CrossRef]
3. Damos, M.A.; Zhu, J.; Li, W.; Hassan, A.; Khalifa, E. A Novel Urban Tourism Path Planning Approach Based on a Multiobjective Genetic Algorithm. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 530. [CrossRef]
4. Ding, C.; Weng, L.; Xia, M.; Lin, H. Non-Local Feature Search Network for Building and Road Segmentation of Remote Sensing Image. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 245. [CrossRef]
5. Liu, S.; Tang, J. Modified Deep Reinforcement Learning with Efficient Convolution Feature for Small Target Detection in VHR Remote Sensing Imagery. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 170. [CrossRef]
6. Xie, F.; Hu, D.; Li, F.; Yang, J.; Liu, D. Semi-Supervised Classification for Hyperspectral Images Based on Multiple Classifiers and Relaxation Strategy. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 284. [CrossRef]
7. Jiang, J.; Lyu, C.; Liu, S.; He, Y.; Hao, X. RWSNet: A semantic segmentation network based on SegNet combined with random walk for remote sensing. *Remote Sens.* **2019**, *41*, 487–505. [CrossRef]
8. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *39*, 640–651. [CrossRef]
9. 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]
10. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Paper presented at the International Conference on Medical image computing and computer-assisted intervention, Munich, Germany, 5–9 October 2015. [CrossRef]
11. Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *arXiv* **2014**, arXiv:1412.7062.

12. Tao, A.; Sapra, K.; Catanzaro, B. Hierarchical Multi-Scale Attention for Semantic Segmentation. *arXiv* **2020**, arXiv:2005.10821.
13. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *99*, 3349–3364. [CrossRef]
14. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. Paper presented at the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, 21–26 July 2017. [CrossRef]
15. 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]
16. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. DenseASPP for Semantic Segmentation in Street Scenes. Paper presented at the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018. [CrossRef]
17. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local Neural Networks. Paper presented at the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018. [CrossRef]
18. Zhao, H.; Zhang, Y.; Liu, S.; Shi, J.; Loy, C.; Lin, D.; Jia, J. PSANet: Point-wise Spatial Attention Network for Scene Parsing. Paper presented at the 2018 European Conference on Computer Vision, Munich, Germany, 8–14 September 2018. [CrossRef]
19. Wurm, M.; Droin, A.; Stark, T.; Geiß, C.; Sulzer, W.; Taubenböck, H. Deep Learning-Based Generation of Building Stock Data from Remote Sensing for Urban Heat Demand Modeling. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 23. [CrossRef]
20. Ghamisi, P.; Yokoya, N. IMG2DSM: Height simulation from single imagery using conditional generative adversarial net. *Remote Sens.* **2018**, *15*, 794–798. [CrossRef]
21. Audebert, N.; Le Saux, B.; Lefèvre, S. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS-J. Photogramm. Remote Sens* **2018**, *140*, 20–32. [CrossRef]
22. Lichao, M.; Zhu, X.X. IM2HEIGHT: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network. *arXiv* **2018**, arXiv:1802.10249.
23. Srivastava, S.; Volpi, M.; Tuia, D. Joint height estimation and semantic labeling of monocular aerial images with CNNs. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; 2017; pp. 5173–5176. [CrossRef]
24. Volpi, M.; Tuia, D. Deep multi-task learning for a geographically-regularized semantic segmentation of aerial images. *ISPRS-J. Photogramm. Remote Sens.* **2018**, *144*, 48–60. [CrossRef]
25. Zilong, H.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. CCNet: Criss-Cross Attention for Semantic Segmentation. Paper presented at the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019. [CrossRef]
26. Woo, S.; Park, J.; Lee, J.; Kweon, I.S. CBAM: Convolutional Block Attention Module. Paper presented at the 2018 European Conference on Computer Vision, Munich, Germany, 8–14 September 2018. [CrossRef]
27. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. ”Paper presented at the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019. [CrossRef]
28. Guo, R.; Liu, J.; Li, N.; Liu, S.; Chen, F.; Cheng, B.; Ma, C. Pixel-wise classification method for high resolution remote sensing imagery using deep neural networks. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 110. [CrossRef]
29. Diakogiannis, F.I.; Waldner, F.; Caccetta, P.; Wu, C. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 94–114. [CrossRef]
30. Wang, J.; Shen, L.; Qiao, W.; Dai, Y.; Li, Z. Deep Feature Fusion with Integration of Residual Connection and Attention Model for Classification of VHR Remote Sensing Images. *Remote Sens.* **2019**, *11*, 1617. [CrossRef]
31. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. Paper presented at the 2016 IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 27–30 June 2016. [CrossRef]
32. Marmanis, D.; Schindler, K.J.; Wegner, D.; Galliani, S.; Datcu, M.; Stilla, U. Classifiification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS J. Photogramm. Remote Sens.* **2018**, *135*, 158–172. [CrossRef]
33. Li, H.; Qiu, K.; Li, C.; Mei, X.; Hong, L.; Tao, C. SCAttNet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2020**.
34. Niu, R.; Sun, X.; Tian, Y.; Diao, W.; Chen, K.; Fu, K. Hybrid Multiple Attention Network for Semantic Segmentation in Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *99*, 1–8. [CrossRef]
35. Hong, D.; Yokoya, N.; Chanussot, J.; Xu, J. Joint and Progressive Subspace Analysis (JPSA) With Spatial–Spectral Manifold Alignment for Semisupervised Hyperspectral Dimensionality Reduction. *IEEE Trans. Cybern.* **2020**, *51*, 3602–3615. [CrossRef] [PubMed]
36. Zhang, Z.; Zhang, X.; Peng, C.; Xue, X.; Sun, J. Exfuse: Enhancing feature fusion for semantic segmentation. Paper presented at the 2018 European Conference on Computer Vision, Munich, Germany, 8–14 September 2018. [CrossRef]
37. Lin, T.; Dollár, P.; Girshick, R.B.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. Paper presented at the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, 21–26 July 2017. [CrossRef]
38. Yann, D.; Fan, A.; Auli, M.; Grangier, D. Language Modeling with Gated Convolutional Networks. Paper presented at the 34th International Conference on Machine Learning (ICML) 2017, Sydney, NSW, Australia, 6–11 August 2017.

39. Towaki, T.; Acuna, D.; Jampani, V.; Fidler, S. Gated-SCNN: Gated Shape CNNs for Semantic Segmentation. Paper presented at the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Kore, 27 October–2 November 2019. [CrossRef]
40. Xiangtai, L.; Zhao, H.; Han, L.; Tong, Y.; Tan, S.; Yang, K. Gated Fully Fusion for Semantic Segmentation. Paper presented at the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–20 February 2020. [CrossRef]
41. Peng, Y.; Sun, S.; Wang, Z.; Pan, Y.; Li, R. Robust Semantic Segmentation by Dense Fusion Network on Blurred VHR Remote Sensing Images. Paper presented at the 2020 6th International Conference on Big Data and Information Analytics (BigDIA), ShenZhen, China, 4–6 December 2020. [CrossRef]
42. Sun, S.; Yang, L.; Liu, W.; Li, R. Feature Fusion Through Multitask CNN for Large-scale Remote Sensing Image Segmentation. In Proceedings of the 2018 10th IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS), Beijing, China, 19–20 August 2018. [CrossRef]
43. Wang, P.; Shen, X.; Cohen, S.; Price, B.; Yuille, A. Towards unified depth and semantic prediction from a single image. Paper presented at the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015. [CrossRef]
44. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2011–2023. [CrossRef] [PubMed]
45. Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *42*, 318–327. [CrossRef] [PubMed]