remote sensing

MDPI

Article

BES-Net: Boundary Enhancing Semantic Context Network for High-Resolution Image Semantic Segmentation

Fenglei Chen , Haijun Liu *, Zhihong Zeng, Xichuan Zhou and Xiaoheng Tan

School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China; flyc@cqu.edu.cn (F.C.); azhihong@cqu.edu.cn (Z.Z.); zxc@cqu.edu.cn (X.Z.); txh@cqu.edu.cn (X.T.)
* Correspondence: haijun_liu@cqu.edu.cn

Abstract: This paper focuses on the high-resolution (HR) remote sensing images semantic segmentation task, whose goal is to predict semantic labels in a pixel-wise manner. Due to the rich complexity and heterogeneity of information in HR remote sensing images, the ability to extract spatial details (boundary information) and semantic context information dominates the performance in segmentation. In this paper, based on the frequently used fully convolutional network framework, we propose a boundary enhancing semantic context network (BES-Net) to explicitly use the boundary to enhance semantic context extraction. BES-Net mainly consists of three modules: (1) a boundary extraction module for extracting the semantic boundary information, (2) a multi-scale semantic context fusion module for fusing semantic features containing objects with multiple scales, and (3) a boundary enhancing semantic context module for explicitly enhancing the fused semantic features with the extracted boundary information to improve the intra-class semantic consistency, especially in those pixels containing boundaries. Extensive experimental evaluations and comprehensive ablation studies on the ISPRS Vaihingen and Potsdam datasets demonstrate the effectiveness of BES-Net, yielding an overall improvement of 1.28/2.36/0.72 percent in mF1/mIoU/OA over FCN_8s when the BE and MSF modules are combined by the BES module. In particular, our BES-Net achieves a state-of-the-art performance of 91.4% OA on the ISPRS Vaihingen dataset and 92.9%/91.5% mF1/OA on the ISPRS Potsdam dataset.

Keywords: remote sensing images; semantic segmentation; boundary enhancing semantic context; fully convolutional network

Citation: Chen, F.; Liu, H.; Zeng, Z.; Zhou, X.; Tan, X. BES-Net: Boundary Enhancing Semantic Context Network for High-Resolution Image Semantic Segmentation. *Remote Sens.* **2022**, *14*, 1638. https://doi.org/10.3390/rs14071638

Academic Editor: Giuseppe Scarpa

Received: 28 February 2022
Accepted: 26 March 2022
Published: 29 March 2022

1. Introduction

Semantic segmentation of remote sensing images [1], also known as land-cover classification [2], aims at locating objects at the pixel level and predicting the semantic categorical label for each pixel in a remote sensing image. It plays an important role in many remote sensing applications [3,4] such as environmental change monitoring, precision agriculture, environmental protection, and urban planning and 3D modeling.

Driven by the rapid development of aeronautics and astronautics technology, together with Earth observation and remote-sensing technology, massive numbers of high-quality and high-resolution remote sensing images have been captured. These high-resolution (HR) remote sensing images are rich in information and contain substantial spatial detail, which can provide the data support for land-cover classification and segmentation. Although semantic segmentation of natural images has achieved considerable progress, the semantic segmentation of HR remote sensing images is still a challenging task, since larger scenes always contain more complex ground information with heterogeneous objects. HR remote sensing images often exhibit large intra-class variations and small inter-class variations at the semantic object level, due to the diversity and complexity of ground objects. For example, the structure, color, and size of buildings show significant variation even in a single scene, as shown by the two buildings outlined with blue rectangles in Figure 1, while

trees and low vegetation are always indistinguishable due to their similar colors and fuzzy boundaries, as shown by the areas in yellow circles in Figure 1. This makes it a difficult task to classify high-resolution remote sensing images pixel by pixel.

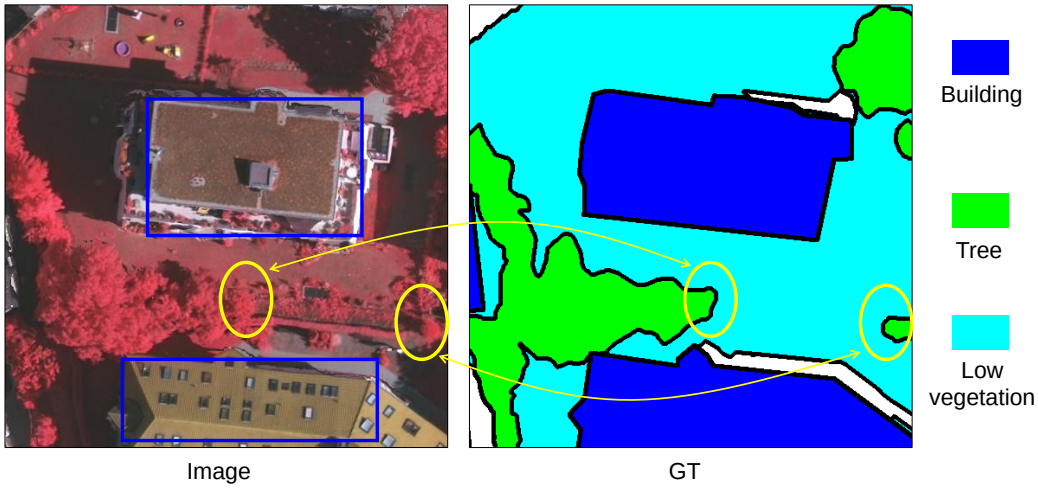


Figure 1. A high-resolution remote sensing sample image and the corresponding ground truth (GT) from ISPRS Vaihingen dataset. One can see the high intra-class variance (blue rectangles) and the low inter-class variance (yellow circles).

Extensive studies have been presented on the challenging HR remote sensing images semantic segmentation task, including studies using traditional methods and deep-learning-based methods. The earlier traditional methods [5,6] mainly consisted of two parts: first extracting features based on the color, shape, texture and spatial position relations of a potential semantic object and then adopting clustering or classification methods to segment the image. They depended heavily on hand-crafted features, always achieving unsatisfactory performance. Recently, with the advent of deep learning, deep convolutional neural networks (DCNNs) have made great progress in semantic segmentation, due to the ability of DCNNs to automatically extract nonlinear and hierarchical features at different semantic levels. Most current semantic segmentation methods are based on the fully convolutional network (FCN) [7] framework, which replaces the fully connected layers with convolutional ones to output spatial feature maps, then utilizes an upsampling operation to generate the predicted maps. Generally, FCN-based architectures consist of a contracting path (also known as an encoder), which extracts information from the input image and obtains high-level feature maps, and an expanding path (also known as a decoder), where high-level feature maps are utilized to generate the mask for pixel-wise segmentation using single-level (e.g., FCN [7], DeepLab [8]) or multilevel (e.g., UNet [9]) upsampling procedures.

The **semantic context** information is a key factor in HR remote sensing images semantic segmentation. Although DCNNs can automatically extract hierarchical semantic features, they simultaneously also reduce the spatial resolution and degrade the spatial detail information in high-level feature maps. Based on the high-level semantic features, single-level upsampling methods (e.g., FCN [7], DeepLab [8]) directly adopt upsampling and (or) dilated (atrous) convolution to generate the high-resolution segmentation output. They may fail in some cases, especially when there are many relatively small objects (e.g., cars) in the fine-spatial-resolution remote sensing images. Alternatively, to address this issue, the utilization of multi-scale contextual feature fusion is a feasible solution for differentiating semantic categories at different spatial scales. The most commonly utilized techniques for aggregating multi-scale contextual features include pyramid pooling modules [10], atrous spatial pyramid pooling [8,11], and context encoding modules [12]. These strategies are always incorporated into the decoder part of the UNet framework. However, in the UNet framework, the low-level and fine-grained detailed features extracted by the

encoder are simply copied and then concatenated with the high-level and coarse-grained semantic features extracted by the decoder, leading to insufficient exploitation of the feature discrimination ability. Therefore, in this study, we investigated how to fuse the multi-scale semantic contexts for parsing HR remote sensing images.

In addition, the **boundary** information of a semantic object may determine the final performance of the HR remote sensing images semantic segmentation. It is well known that in the semantic segmentation task, the pixels at the boundary are more likely to be misclassified. However, HR remote sensing images always contain large and complex scenes with heterogeneous objects with various shapes, scales, textures, and colors. The boundaries of objects are often ambiguous and blurry [13,14] due to the lighting conditions, imaging angles, occlusions, and shadows, as shown by the areas marked by yellow circles in Figure 1. Although DCNNs have the ability to learn robust and discriminative features, the features extracted by a DCNN always fail to distinguish the adjacent objects, since in HR remote sensing images the semantic objects are adjacent and often have a similar appearance (color). The reason may be that the high-level features of DCNNs tend to extract the local semantic information while ignoring the global geometric prior and over-smoothing the boundaries of objects, which is important for object localization. There are generally two ways to improve the accuracy of the boundary: by adding the boundary loss [15–17] and by adding extra boundary detection sub-networks [18–20]. These methods mainly focus on the detection of the boundary, ignoring the relationships between the boundary and the semantic context. That is, they pay little attention to utilizing the boundary information to guide the semantic context, in order to improve the final performance of the semantic segmentation at the object level. Therefore, we also investigated how to explicitly adopt the extracted boundary information to enhance the semantic context for parsing HR remote sensing images.

To address the above-mentioned two research points, in this paper we propose the boundary enhancing semantic context network (BES-Net) for high-resolution remote sensing images semantic segmentation. Based on the FCN framework, a ResNet [21] model is adopted as the backbone to extract the hierarchical semantic features. Additionally, BES-Net builds three modules to emphasize the boundary and semantic context information, including the boundary extraction (BE) module, the multi-scale semantic context fusion (MSF) module, and the boundary enhancing semantic context (BES) module. The BE module is introduced to predict the binary boundary of the objects, simultaneously adopting low-level detailed features and the highest-level semantic features from the backbone as the input. It is supervised by the binary boundary labels generated from the segmentation ground truth using the Laplacian operation [22]. The MSF module adopts the high-level semantic features from the backbone as the input and fuses them with the attention mechanism in a hierarchical manner, to obtain the fused semantic features containing objects with multiple scales. Finally, the BES module is designed to explicitly adopt the extracted boundary information to enhance the fused semantic context using simple addition and multiplication operations. By aggregating the semantic context information along with the boundaries, pixels from the same semantic category can receive a similar response, enhancing the semantic consistency.

The main contributions can be summarized as follows:

- We present a simple yet effective semantic segmentation framework, BES-Net, for HR remote sensing images semantic segmentation.
- We *explicitly*, not implicitly, adopt the well-extracted boundary to enhance the semantic context for semantic segmentation. Accordingly, three modules are designed to enhance the semantic consistency in the complex HR remote sensing images.
- Experimental results on two HR remote sensing images semantic segmentation datasets demonstrate the effectiveness of our proposed approach compared with state-of-the-art methods.

2. Related Work

This section briefly reviews related deep-learning-based semantic segmentation methods, mainly including the following two aspects: multi-scale feature learning for semantic segmentation and boundary improved semantic segmentation.

2.1. Multi-Scale Feature Learning for Semantic Segmentation

Recently, deep convolutional neural network (DCNN)-based methods have dominated the field of semantic segmentation. Fully convolutional networks (FCNs) [7] laid the foundation for the application of CNNs in semantic segmentation by replacing the fully connected layers with convolutional ones, to output spatial maps. Based on the FCN framework, various semantic segmentation approaches have arisen. However, the feature maps learned by the FCN backbone reduce the spatial resolution and lose the spatial location information, which results in the inaccurate prediction of small objects and the boundaries of objects.

To address this issue, a number of studies focus on multi-scale feature learning for capturing more context information to enhance the feature representation. For example, Zhao et al. [10] first proposed a pyramid pooling module to aggregate the multi-scale semantic context information. Chen et al. [8] introduced an atrous spatial pyramid pooling (ASPP) module with dilated convolution to expand the receptive fields at different scales.

Additionally, some research [23,24] is based on the UNet [9] framework, which includes an encoder–decoder structure with skip connections between the encoder and decoder layers to bridge the high-level semantic information and low-level spatial information. Bai et al. [24] presented a hierarchical context aggregation network, incorporating the proposed compact ASPP module into the UNet framework to replace the copy-and-crop operation for extracting the multi-scale context information. Diakogiannis et al. [25] extended the UNet framework by incorporating residual connections, atrous convolutions, and pyramid scene parsing pooling, to perform multi-tasking learning.

Moreover, some research focused on enhancing feature learning by introducing attention mechanisms [14,23,26,27]. Chen et al. [26] presented a pioneering study utilizing the attention to re-weight the multi-scale features. Then, Wang et al. [28] proposed a non-local attention module to capture the global dependencies for pixels at all positions and to utilize these to refine the feature representations. Yang et al. [29] proposed a multipath encoder framework, consisting of a multipath attention-fused block module to fuse multipath features and a refinement attention-fused block module to fuse high-level abstract features and low-level spatial features. Li et al. [23] proposed a multi-attention network to extract contextual dependencies through multiple efficient attention modules with linear complexity. They also proposed an attentive bilateral contextual network, i.e., a lightweight convolutional neural network with a spatial path and a contextual path [27].

The above methods could refine the semantic features with some detailed spatial information; however, they still could not handle the pixels at the boundary well, i.e., the pixels that are easily misclassified.

2.2. Boundary Improved Semantic Segmentation

As an essential element of the image, boundaries play a vital role in improving the semantic segmentation performance. Several studies have made great progress in improving the accuracy of boundaries. In the early studies on DCNN-based semantic segmentation performance, most modules adopted boundary information as a post-processing step to refine the segmentation results, using methods such as boundary neural fields [30], an affinity field [31], and a random walk [32].

Recently, some approaches have tended to explicitly construct a boundary detection sub-network in parallel with the semantic segmentation network to distinguish these confusing pixels. Takikawa et al. [17] specially designed a boundary detection stream and combined the two tasks of boundary and semantic modeling for boundary enhancement. Li et al. [16] pointed out that in an image, the object boundary and body area correspond

to high- and low-frequency information, respectively, and they proposed solving semantic segmentation by explicitly modeling the body consistency and edge preservation at the feature level and then jointly optimizing them in a unified framework. Ding et al. [33] proposed learning the boundary as an additional semantic class, to enable the network to be aware of the boundary layout, using a boundary-aware feature propagation module to harvest and propagate the local features within the regions isolated by the learned boundaries. Ma et al. [18] also exploited the boundary information for context aggregation. Sun et al. [19] proposed a boundary-aware semi-supervised semantic segmentation network with a focus on the object boundaries in complex scenes, including a channel-weighted multi-scale feature module that fuses the semantic and spatial information and a boundary attention module that weights the feature map containing rich boundary information.

Moreover, some research [15,34,35] has additionally focused on developing novel loss methods to address the boundary information. Zheng et al. [15] proposed a dice-based edge-aware loss function to guide the networks to refine both the pixel-level and image-level edge information directly from semantic segmentation prediction. EdgeNet [35] addresses segmentation tasks from the perspective of efficiency. It contains a class-aware edge loss module and a channel-wise attention mechanism. It aims to preserve the segmentation performance with no drop in inference speed.

In summary, these methods always separately focus on learning semantic features or extracting the boundary, ignoring their combination. They assume that the well-extracted boundary can *implicitly* improve the semantic features for semantic segmentation. In our study, we take one further step and *explicitly* design a boundary enhancing semantic context module to refine the semantic features for improving the performance of HR remote sensing images semantic segmentation.

3. Methodology

In this section, we introduce the framework of our proposed boundary enhancing semantic network (BES-Net) for parsing high-resolution remote sensing images.

3.1. The Framework of BES-Net

As depicted in Figure 2, *BES-Net explicitly adopts boundary information to enhance the semantic context*, ensuring that those pixels within one object achieve similar responses after the semantic feature aggregation. Specifically, our method employs the ResNet [21] model as its backbone to extract the hierarchical features, where the vanilla convolutions are replaced by dilated ones to enlarge the receptive field (also known as FCN_8s [7]). The backbone outputs five feature maps: F_1 and F_2 at $1/4$ of the size of the input resolution, with low-level detail information, and F_3, F_4, and F_5 at $1/8$ of the size of the input resolution, with high-level semantic information. Moreover, BES-Net contains three modules for aggregating the boundary and semantic context information, including the boundary extraction (BE) module, the multi-scale semantic context fusion (MSF) module, and the boundary enhancing semantic context (BES) module.

The BE module focuses on extracting the boundary information by regarding it as an independent sub-task along with the mainstream semantic segmentation. The low-level detail features are concatenated with the highest-level semantic features to capture the semantic boundaries. The boundary information is supervised by the binary boundary labels generated from the segmentation ground truth using a Laplacian operation. Furthermore, regarding the three high-level semantic features which have different semantic scales, the MSF module fuses them using the attention mechanism in a hierarchical manner to obtain the refined semantic features. Finally, the BES module is carefully designed to enhance the semantic context with the extracted semantic boundaries from the BE module. By aggregating the semantic information along the boundaries, pixels from the same semantic category can receive a similar response, thus enhancing the semantic consistency. We give details of each of the modules in the following subsections.

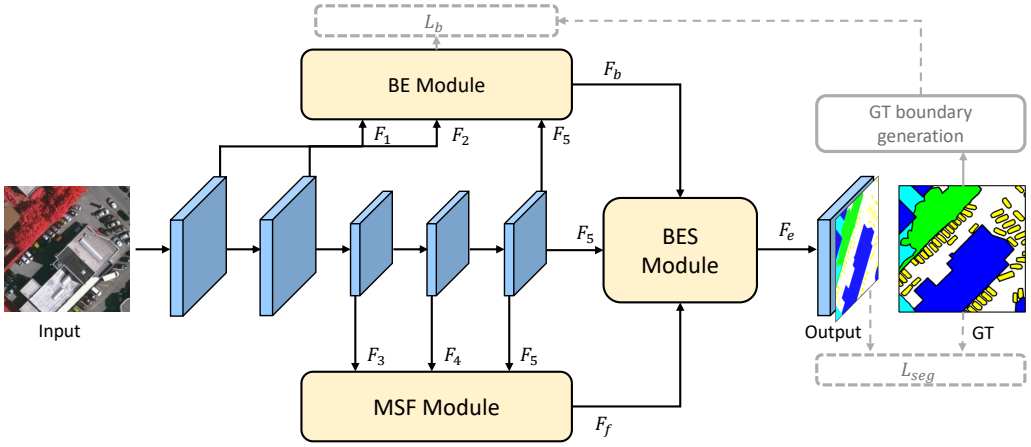


Figure 2. The pipeline of our proposed BES-Net framework for parsing HR remote sensing images. In addition to the backbone, BES-Net also contains three other main components: the boundary extraction (BE) module, the multi-scale semantic context fusion (MSF) module, and the boundary enhancing semantic context (BES) module. The BES module can enhance the semantic information extracted from MSF module with the boundary information extracted from the BE module, to improve the semantic segmentation performance, giving more intra-class segmentation consistency.

3.2. Boundary Extraction Module

The boundary extraction module is designed to extract the boundaries of semantic objects. Since deep convolutional neural networks can learn features containing both low-level detail information and high-level semantic information in a hierarchical manner, the BE module directly borrows the intermediate features from the backbone. Although the boundary information exists in the low-level detail feature maps, most of them lack the semantic information. Therefore, to extract the semantic boundary, as well as the two low-level detail features (F_1 and F_2), the BE module also utilizes the highest-level semantic feature (F_5) as an input, as shown in Figure 3.

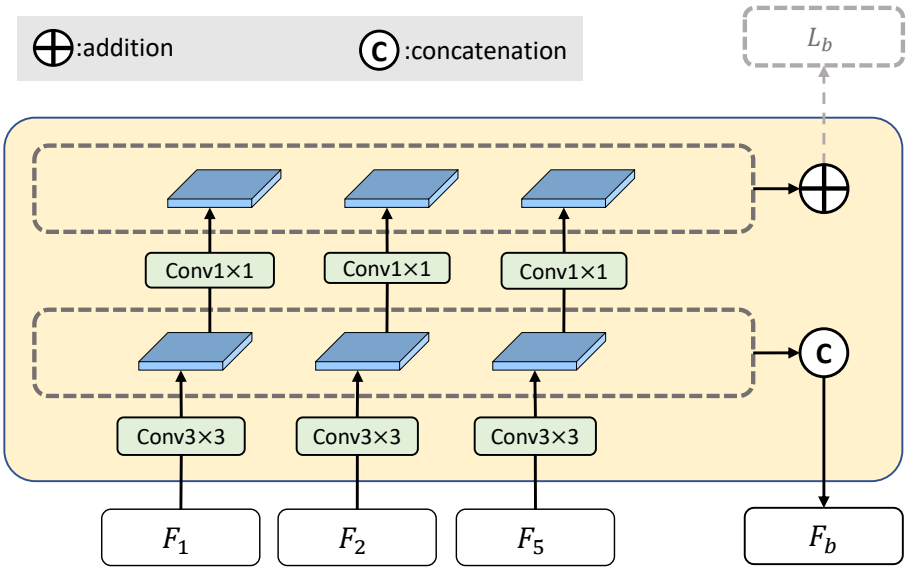


Figure 3. The boundary extraction (BE) module.

Based on the inputs, the BE module firstly unifies their channels to d_b (e.g., = 64) using 3×3 convolutional layers. The extracted boundary feature map of F_5 is upsampled to the size of F_1 (i.e., 1/4 of the size of the input) and then they are concatenated together as the boundary features F_b.

Furthermore, to ensure the boundary features truly contain the boundary information of a semantic object, the extracted multi-scale boundary feature maps are supervised by the binary boundary labels generated from the segmentation ground truth. Specifically, the multi-scale boundary feature maps are mapped to the 1-channel boundary maps by 1×1 convolutional layers and the sigmoid function, then merged by element-wise addition.

The binary boundary labels are generated from the semantic segmentation ground truth using the Laplacian operation, as shown in Figure 4. Since the semantic objects in a scene always have various sizes, inspired by ASPP [8] we performed the boundary extraction at different scales. Given the ground truth (GT), we can obtain the corresponding boundary (GT_b) as follows:

$$GT_b = \text{Convs}\Big(\text{Cat}_{i=1,2,4}\big(\text{Up}(\text{LConv}_i(GT))\big)\Big), \tag{1}$$

where LConv_i is the convolution which adopts the Laplacian operator $\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$ as the convolutional kernel to perform 2D convolutions with strides of i ($i = 1$, 2 and 4, respectively). This can produce soft, thin-detail feature maps with multi-scale semantic boundaries. Then, the feature maps are bilinearly upsampled (Up) to the original size, and channels are concatenated (Cat) together and mapped to the 1-channel boundary maps by 1×1 convolutional layers (Convs).

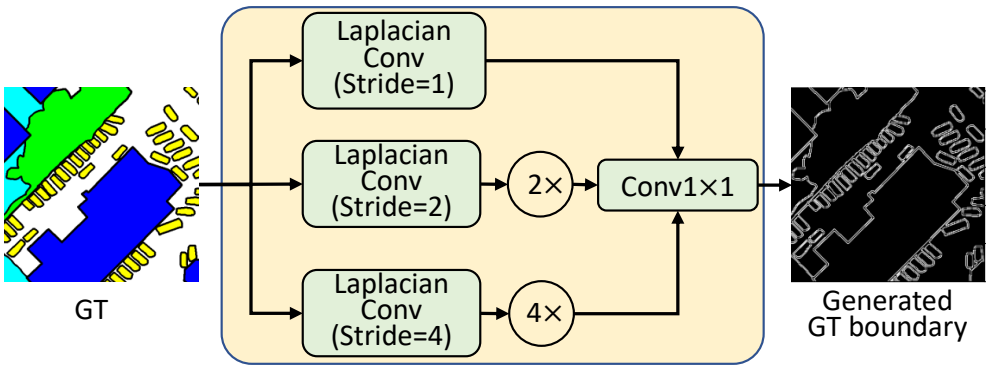


Figure 4. The procedure for GT boundary generation.

3.3. Multi-Scale Semantic Context Fusion Module

The three high-level features (F_3, F_4, and F_5) have different semantic scales, due to the different receptive fields. The multi-scale semantic context fusion module is designed to fuse the multi-scale semantic information into one feature map. As shown in Figure 5, the three high-level features (F_3, F_4, and F_5) are firstly converted to f_3, f_4, and f_5 with the same channel d_f (e.g., = 128 for ResNet18, = 512 for ResNet50 and ResNet101) using 1×1 convolutional layers. Then, they are fused by the fusion blocks in a hierarchical manner to obtain the fused semantic features.

The fusion block is based on the channel attention mechanism, since different channels of features may correspond to different semantic classes [36]. Compared to features from different spatial positions, the features from different channels may have higher class discriminability. Considering this, the fusion block aims to effectively exploit the cross-scale complementary information by re-weighting the importance of single-scale features in a channel-dependent way.

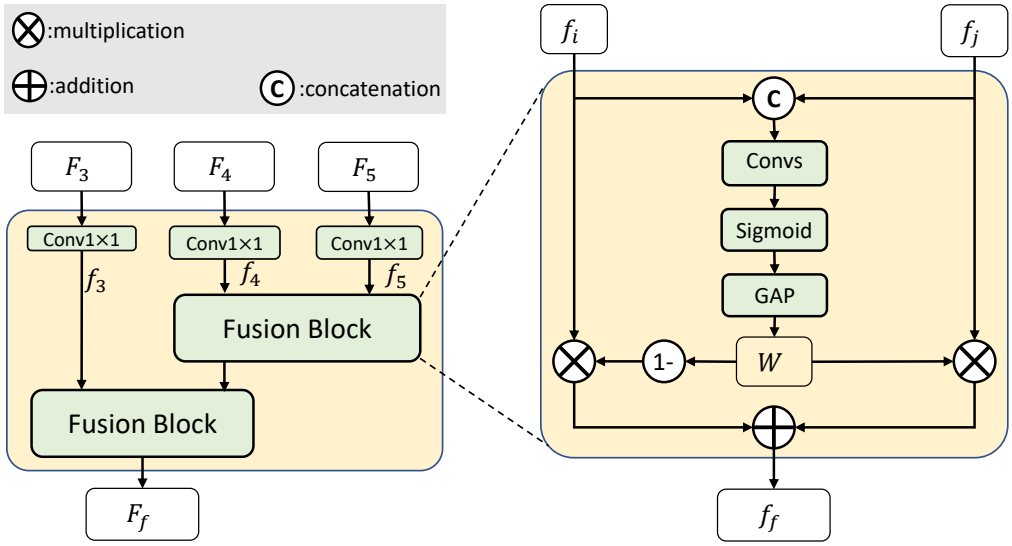


Figure 5. The multi-scale semantic context fusion (MSF) module.

As shown on the right of Figure 5, given the two-scale features (f_i and f_j, $i < j$), they are concatenated and then fed into two convolutional layers to obtain the relative importance of the paired features from different scales but in the same channels. The channel importance (or attention) weights W can be calculated as follows:

$$W = \mathrm{GAP}\Big(\mathrm{Sigmoid}\big(\mathrm{Convs}\big(\mathrm{Cat}(f_i, f_j)\big)\big)\Big), \tag{2}$$

where $\mathrm{Cat}(*)$ is the channel concatenation operation, and $\mathrm{Convs}(*)$ is a convolutional block with a 1×1 convolutional layer (for channel reduction) and a 3×3 convolutional layer (for feature refining). $\mathrm{Sigmoid}(*)$ is the sigmoid function, and $\mathrm{GAP}(*)$ denotes the global average pooling operation.

Higher values of W indicate that the corresponding channels of features at the jth scale are more likely to be important than the corresponding channels of features at the ith scale, and vice versa. As a result, the relative importances of the channels of features from different scales are obtained. Therefore, we can adopt the gate fusion method [37] to fuse the two scale features, where the channel importance weights correspond to the gate. Based on the channel importance weights, the fused features can be computed as follows:

$$\begin{aligned} f_f &= \mathrm{fusion}(f_i, f_j) \\ &= f_i \bullet (1 - W) + f_j \bullet W, \end{aligned} \tag{3}$$

where $\mathrm{fusion}(*)$ represents the whole fusion block, and $\bullet$ denotes channel-wise multiplication.

Finally, the three high-level features (F_3, F_4, and F_5) are fused by the MSF module in a hierarchical manner with a fusion block, to obtain the final fused feature F_f.

$$F_f = \mathrm{fusion}\big(F_3, \mathrm{fusion}(F_4, F_5)\big). \tag{4}$$

3.4. Boundary Enhancing Semantic Context Module

Since the semantic boundary has intrinsic partitioning capability for an object, the goal of the BES module is to enhance the fused semantic features F_f, to give them more intra-class consistency using the extracted boundary features F_b. The key point is the method of aggregating the two features. F_f has sufficient semantic information (focusing on the body area of an object without the boundary information), while F_b has salient boundary information. They complement each other in describing an object.

Like the fusion block of the MSF module in the previous subsection, the BES module is designed based on two fundamental mathematical operations (element-wise addition, $+$, and element-wise multiplication, $\times$) to aggregate the two complementary features. Generally, the multiplication operation can filter out the boundary-related information to emphasize it, while the addition operation can complement two features. With the two simple, parameter-free mathematical operations, two complementary features can be effectively fused to describe the complete information of objects. Moreover, to obtain more robust performance, we also utilize the extracted boundary features F_b to enhance the highest-level semantic feature F_5.

As shown in Figure 6, given three features, i.e., the fused semantic features F_f, the extracted boundary features F_b, and the highest-level semantic features F_5, the BES module firstly unifies the channel numbers of the three features to d_e (e.g., = 128) using 1×1 convolutional layers. Taking the highest-level semantic features F_5 as an example, this is updated as follows:

$$F_5 \Leftarrow \mathrm{ASPP}\big(\mathrm{Conv}(F_5)\big), \tag{5}$$

where $\mathrm{ASPP}(*)$ is the atrous spatial pyramid pooling module [8], and $\mathrm{Conv}(*)$ denotes the 1×1 convolutional layer for unifying the channel numbers.

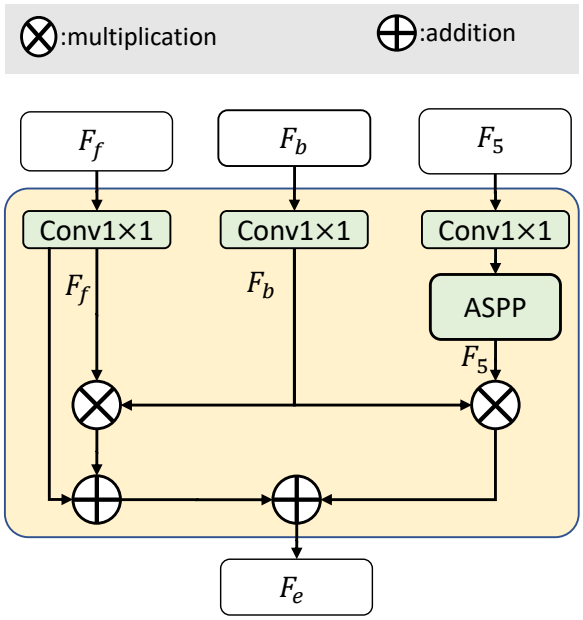


Figure 6. The boundary enhancing semantic context (BES) module.

All the features are then upsampled to the size of F_b (i.e., 1/4 of the size of the input).

Finally, the fused semantic features F_f are enhanced by the extracted boundary features F_b using the addition and multiplication operations. Simultaneously, in order to make the boundary-related information more salient, features F_5 are also enhanced by F_b using the multiplication operation. The final enhanced features F_e are calculated as follows:

$$F_e = F_f + F_b \times F_f + F_b \times F_5. \tag{6}$$

3.5. Loss Function

Based on the final enhanced features F_e, with sufficient semantic information and boundary information, we can predict the segmentation results with a convolution block (sequentially including a 3×3 convolutional layer and a 1×1 convolutional layer), in the same way as for FCN [7] and DeepLab [8].

In our framework, BES-Net outputs two main results that aim to generate the segmentation masks and boundaries, respectively. For semantic segmentation, we adopt the standard cross-entropy L_{seg} to measure the difference between the predicted masks P and the ground truth G:

$$L_{seg} = -\frac{1}{K}\sum_{k=1}^{K}\sum_{i}\left(G_{i,k}\log P_{i,k} + (1 - G_{i,k})\log(1 - P_{i,k})\right), \tag{7}$$

where K is the number of semantic categories corresponding to K predicted feature maps, and the subscript $_{i,k}$ denotes the ith pixel in the kth predicted feature map.

Similarly, for boundary prediction, we also adopt the binary cross-entropy to measure the difference between the predicted boundary P^b and the ground truth boundary G^b:

$$L_b = -\sum_{i}\left(G_i^b \log P_i^b + (1 - G_i^b)\log(1 - P_i^b)\right). \tag{8}$$

Additionally, following the settings in DeepLab [8], to accelerate model convergence we also apply an auxiliary cross-entropy loss L_{aux} to the intermediate feature representations of the backbone. Therefore, the overall training loss is

$$L = L_{seg} + \lambda_1 L_b + \lambda_2 L_{aux}, \tag{9}$$

where λ_1 and λ_2 are the hyperparameters. We empirically set $\lambda_1 = 1$ and $\lambda_2 = 0.4$.

4. Experiments and Results

In this section, we evaluate the effectiveness of our proposed methods for HR remote sensing images semantic segmentation on two public 2D semantic labeling datasets provided by the International Society for Photogrammetry and Remote Sensing (ISPRS) (https://www2.isprs.org/commissions/comm2/wg4/benchmark/semantic-labeling/, accessed on 26 February 2022): Vaihingen and Potsdam. Both datasets cover urban scenes. The reference data are labeled according to six land-cover classes: impervious surfaces, buildings, low vegetation, trees, cars, and clutter/background.

4.1. Experimental Settings

4.1.1. Datasets and Settings

The ISPRS Vaihingen dataset recorded a relatively small village with many detached buildings and small multi-story buildings. It contains 33 orthophoto image patches. The images have an average size of 2494 × 2064 pixels and a resolution of 9 cm. The near-infrared (IR), red (R), and green (G) channels, together with corresponding digital surface models (DSMs) and normalized DSMs (NDSMs), are provided in the dataset. We only utilized the IR-R-G images; the DSMs were not used in the experiments. Following the official data split, we employed 16 images for training and 17 images for testing.

The ISPRS Potsdam dataset recorded a typical historic city with large building blocks, narrow streets, and a dense settlement structure. It contains 38 orthophoto image patches. The images have a size of 6000 × 6000 pixels and a resolution of 5 cm. The dataset provides the near-infrared, red, green, and blue channels, as well as DSMs and NDSMs. Again, we only utilized the IR-R-G images in the experiments. Following the official data split, we employed 24 images for training and 14 images for testing.

4.1.2. Evaluation Metrics

Following existing studies, the overall accuracy (OA), mean intersection over union (mIoU), and mean F1-score (mF1) were adopted as evaluation metrics. Based on the accumulated confusion matrix, the OA, mIoU, and mF1 were calculated as.

$$\mathrm{OA} = \frac{\sum_{k=1}^{K} \mathrm{TP}_k}{\sum_{k=1}^{K} \mathrm{TP}_k + \mathrm{FP}_k + \mathrm{TN}_k + \mathrm{FN}_k}, \tag{10}$$

$$\mathrm{mIoU} = \frac{1}{K} \sum_{k=1}^{K} \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k + \mathrm{FN}_k}, \tag{11}$$

$$\begin{aligned} \mathrm{mF1} &= \frac{1}{K} \sum_{k=1}^{K} \mathrm{F1}_k \\ &= \frac{1}{K} \sum_{k=1}^{K} 2 \times \frac{\mathrm{precision}_k \times \mathrm{recall}_k}{\mathrm{precision}_k + \mathrm{recall}_k}, \end{aligned} \tag{12}$$

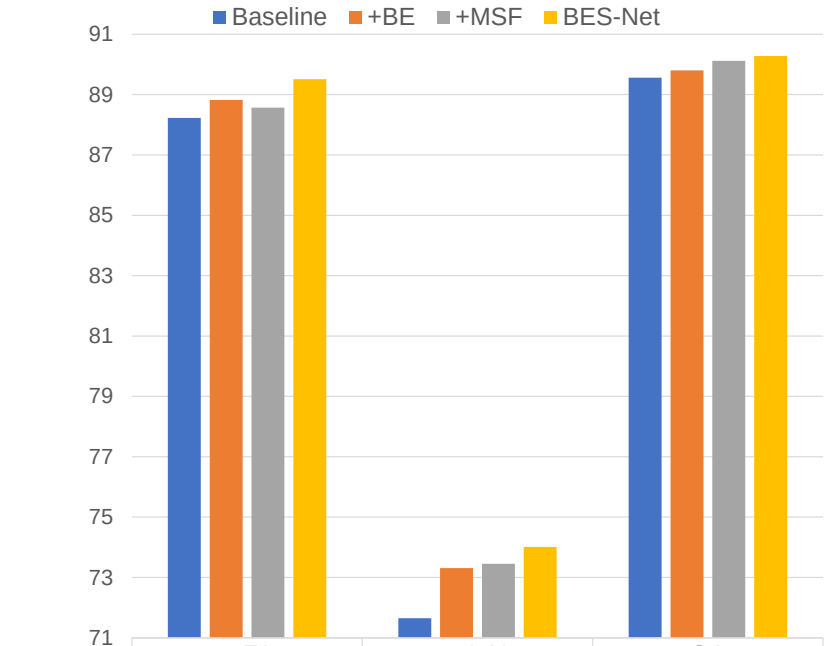
where TP_k, FP_k, TN_k, and FN_k denote the true positive, false positive, true negative, and false negative values for a class k, respectively. In addition, $\mathrm{precision}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k}$ and $\mathrm{recall}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k}$ are the precision and recall indicators for class k, respectively.

4.1.3. Implementation Details

The implementation (https://github.com/FlyC235/BESNet, accessed on 26 February 2022) of our method uses the PyTorch framework. Following existing HR remote sensing images semantic segmentation studies, the ResNet model was adopted as the backbone network for a fair comparison, and the pre-trained ImageNet parameters were adopted for the network initialization. We used the dilated FCN_8s as the baseline. In the training phase, considering the limited GPU memory, we cut the training images, as well as the corresponding labels, into patches with a size of 512×512, using a sliding window with an overlap of 171 ($\approx 512 \times \frac{1}{3}$) pixels. To avoid overfitting, some common data augmentation methods were adopted, including random flipping and rotating at 30-degree intervals. We adopted the mini-batch stochastic gradient descent (SGD) optimizer for optimization, with the momentum and weight decay parameters set to 0.9 and 0.0005, respectively. The maximum training epoch number $iter_{max}$ was set to 120 for the Vaihingen dataset and 200 for the Potsdam dataset. We set the initial learning rate as 0.005. The "poly" learning rate strategy [38] was adopted to update the learning rate, where at each iteration ($iter$) the learning rate is multiplied by $(1 - \frac{iter}{iter_{max}})^{0.9}$. Following the settings used in other studies [19,24,27], in the testing phase we also adopted data augmentation strategies, including horizontal and vertical flipping, multiple scales $[0.75\times, 1\times, 1.25\times]$, and overlay fusion on the full test files, which is also known as test-time augmentation (TTA).

4.2. Ablation Experiments

In this section, we evaluate the effectiveness of our proposed BES-Net method, including three components: the boundary extraction (BE) module, the multi-scale semantic context fusion (MSF) module, and the boundary enhancing semantic context (BES) module. The ablation experiments were only conducted on the ISPRS Vaihingen dataset (note that to simply show the effectiveness of different components during the ablation experiments, we only reported the results without the test-time augmentation (TTA) during testing). The dilated FCN_8s with the ResNet50 backbone was adopted as the baseline. On this basis, we tested the effectiveness of each module by incorporating them separately or simultaneously. The results are shown in Figure 7.



	mF1	mIoU	OA
Baseline	88.23	71.65	89.56
+BE	88.82	73.31	89.80
+MSF	88.57	73.45	90.12
BES-Net	89.51	74.01	90.28

Figure 7. Ablation study (%) on ISPRS Vaihingen dataset.

From the figure, we can see that the three modules all have a positive effect on improving the baseline performance,since:

- Compared to the baseline, using only the BE module (+BE) improved the mF1, mIoU, and OA by 0.59%, 1.66%, and 0.24%, respectively.
- Compared to the baseline, using only the MSF module (+MSF) improved the mF1, mIoU, and OA by 0.34%, 1.80%, and 0.56%, respectively.
- When combining the BE and MSF modules using the BES module (our BES-Net), the mF1, mIoU, and OA were improved by 1.28%, 2.36%, and 0.72%, respectively, compared to the baseline.
- Compared to +BE and +MSF methods, our BES-Net performed much better. This demonstrates the effectiveness of explicitly adopting boundary information to enhance the semantic context.
- The ablation experiments demonstrated the effectiveness of our proposed three modules, BE, MSF, and BES, for HR remote sensing images semantic segmentation.

Furthermore, based on the framework of BES-Net, we visualized the intermediate feature maps generated by our proposed three modules. As shown in Figure 8, the feature map of F_b has rich boundary information at the semantic object level, while the feature map of F_f has sufficient high-level semantic information (focusing on the body area of an object without the boundary information) of multi-scale objects, including those at a fine scale (the small cars) and at a coarse scale (the large buildings). F_b and F_f complement each other in describing an object. After aggregating the fused semantic features F_f using the extracted boundary features F_b to obtain the final enhanced features F_e, the feature map of F_e has abundant boundary and semantic context information simultaneously. Finally, the predicted results from F_e matched the ground truth more accurately, with more intra-class semantic consistency.

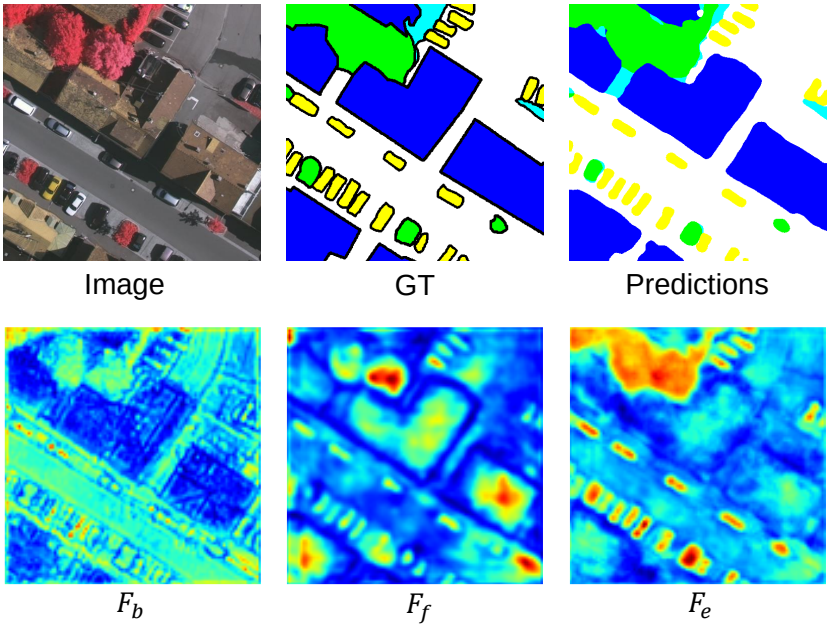


Figure 8. Visualization results of intermediate feature maps of one sample image from ISPRS Vaihingen dataset. The top row shows the sample image, corresponding ground truth (GT), and the segmentation results predicted by our proposed BES-Net method. The bottom row shows the feature maps of the extracted boundary features F_b, the fused multi-scale semantic features F_f, and the final enhanced features F_e, respectively.

4.2.1. Boundary Extraction

In our proposed boundary extraction module, we adopted two low-level detail features (F_1 and F_2) and only the highest-level semantic feature (F_5) as the inputs. To verify why only F_5 was utilized and not the other two features F_3 and F_4, we conducted the following experiments (Baseline+BE) with different input combinations. Since the goal of the BE module is to extract the semantic boundaries, we used a fixed input for the two low-level detail features and gradually added the high-level semantic features from F_5 to F_3. The results are listed in Table 1. We can see that only combining the highest-level semantic features F_5 achieved the best performance. When adding more features, the performance was degraded, especially regarding the mIoU metric. This is reasonable, because F_5 can provide sufficient semantic information to guide the boundary learning, and adding F_3 and F_4 introduces more convolution parameters.

Table 1. The performance (%) of BE module with different input combinations on ISPRS Vaihingen dataset.

Index	Inputs					mF1	mIoU	OA
	F_1	F_2	F_3	F_4	F_5			
ine ①	✓	✓	-	-	✓	**88.82**	**73.31**	**89.80**
②	✓	✓	-	✓	✓	88.85	71.56	89.73
③	✓	✓	✓	✓	✓	88.59	71.90	**89.80**

4.2.2. Multi-Scale Semantic Context Fusion

In our proposed multi-scale semantic context fusion module, we fuse the three high-level semantic features F_3, F_4, and F_5 in a hierarchical manner (Equation (4)), firstly fusing two features and then fusing the last one with the fused result. To determine the fusion strategy, we conducted the following experiments (Baseline+MSF) with different fusion orders. F_3, F_4, and F_5 have sequentially higher semantic scales. According to their semantic

scales, we set three kinds of orders, corresponding to indexes ④, ⑤, and ⑥, as listed in Table 2. We can see that method ④ with the fusion fusion$\big(F_3, \text{fusion}(F_4, F_5)\big)$ order, in a coarse-to-fine scale manner, achieved the best performance.

Table 2. The performance (%) of MSF module with different fusion orders on ISPRS Vaihingen dataset.

Index	Fusion Orders	mF1	mIoU	OA
④	fusion$\big(F_3, \text{fusion}(F_4, F_5)\big)$	88.57	**73.45**	**90.12**
⑤	fusion$\big(F_4, \text{fusion}(F_3, F_5)\big)$	88.53	72.95	89.95
⑥	fusion$\big(\text{fusion}(F_3, F_4), F_5\big)$	88.35	72.31	90.09
⑦	Cat(④,⑤,⑥)	**88.89**	72.66	89.97

Moreover, we also conducted more experiments by concatenating the results of the above three methods, corresponding to index ⑦. Compared to our method (index ④), this introduced twice as many convolution parameters, achieving a slightly better mF1 result (+0.32%) but a much worse mIoU result (-0.79%).

4.2.3. Boundary Enhancing Semantic Context

In our proposed boundary enhancing semantic context module, we introduced a feature aggregation method containing two simple mathematical operations to utilize the extracted boundary features to enhance the fused multi-scale semantic features. To verify the effectiveness of the feature aggregation method, we conducted the following experiments (BES-Net) with different boundary enhancing semantic features. Table 3 shows that:

- When the highest-level backbone semantic feature F_5 is enhanced by the boundary feature F_b, corresponding to index ⑧, it achieves better performance compared to the baseline.
- When F_b enhances the fused multi-scale semantic features F_f, corresponding to index ⑨, it slightly outperforms method ⑧.
- Finally, F_b simultaneously enhancing F_f and F_5, corresponding to index ⑩, achieves the best performance.
- The experimental results demonstrate the effectiveness of our BES-Net in explicitly adopting boundary information to enhance the semantic context.

Table 3. The performance (%) of BES module with different boundary enhancing semantic features on ISPRS Vaihingen dataset.

Index	Inputs			mF1	mIoU	OA
	F_b	F_f	F_5			
Baseline	-	-	-	88.23	71.65	89.56
⑧	✓	-	✓	89.01	72.97	90.18
⑨	✓	✓	-	89.47	73.61	90.17
⑩	✓	✓	✓	**89.51**	**74.01**	**90.28**

4.2.4. Backbone

Furthermore, we conducted experiments based on our BES-Net with different ResNet backbones, including ResNet18, ResNet50, and Resnet101. As shown in Table 4, with increasing model parameters, the segmentation performances were improved according to the mF1 and OA metrics. This is reasonable, because the well-structured ResNet models with more parameters correspond to more powerful feature representation capabilities. However, from ResNet18 to ResNet101, the parameters increase by about five times, and the results on the Vaihingen dataset show improvements in mF1 of 0.82% and in OA of

0.65%, while the results on the Vaihingen dataset only show improvements in mF1 of 0.07% and in OA of 0.18%.

Moreover, The results of ResNet18 on the Potsdam dataset are better than those on Vaihingen dataset, which is reasonable since the Potsdam dataset has a fine resolution (5 cm), while the Vaihingen dataset has a slightly coarser resolution (9 cm). Fine-resolution images can provide more spatial detail about the objects.

Table 4. The performance (%) of BES-Net with different ResNet backbones on ISPRS Vaihingen and Potsdam datasets.

Backbone	Vaihingen			Potsdam		
	mF1	**mIoU**	**OA**	**mF1**	**mIoU**	**OA**
ine ReNet18	88.86	73.00	89.92	92.05	**78.21**	90.52
ResNet50	89.51	74.01	90.28	92.09	77.98	90.64
ResNet101	**89.68**	**75.04**	**90.57**	**92.26**	77.91	**90.71**

4.2.5. The Hyperparameters in the Loss Function

There are two hyperparameters in the loss function (Equation (9)): λ_1 for boundary loss L_b and λ_2 for auxiliary loss L_{aux}. As most of the related studies follow the settings of DeepLab [8] to set the hyperparameter $\lambda_2 = 0.4$, we also adopted this setting here. For the hyperparameter λ_1, we conducted experiments with different values, $\{0.1, 0.5, 1.0, 5.0, 10.0\}$.

As shown in Figure 9, we find that the mF1/OA of BES-Net improves when the weight range is from 0.1 to 0.5 to 1.0, then drops when the weight range is from 1.0 to 5.0 to 10.0. The highest performance is achieved when $\lambda_1 = 1.0$ (the boundary loss and segmentation loss have equal weight). Furthermore, the performance of BES-Net drops slightly when the weights are 0.1 and 0.5 but drops dramatically when the weights are 5.0 and 10.0. These results suggest that focusing too much on the boundary information may lead to ignoring the inter-class semantic information. Therefore, the boundary loss and segmentation loss should have equal weight.

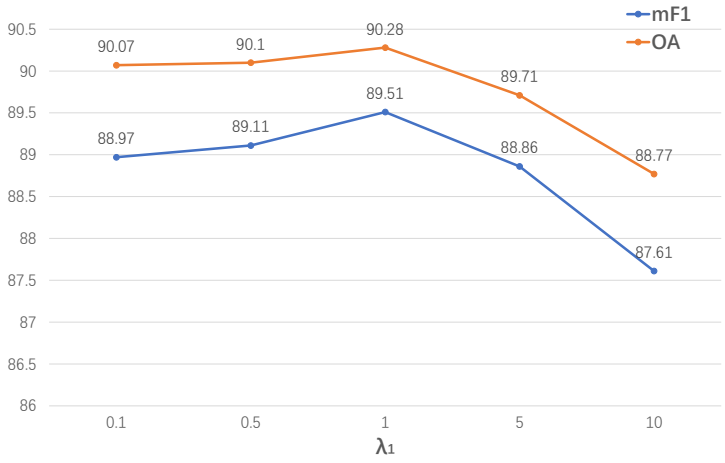


Figure 9. The effect (%) of different weights λ_1 of the boundary loss.

4.3. Comparison to the State of the Art

This section compares our BES-Net (with a ResNet18, ResNet50, and ResNet101 backbone) to state-of-the-art HR remote sensing images semantic segmentation methods. The results with test-time augmentation (TTA) on the ISPRS Vaihingen and Potsdam datasets are listed in Table 5.

The experiments on the Vaihingen dataset show that:

- Our proposed BES-Net method can achieve comparable performance to the current state-of-the-art results obtained by BAS^{4}Net [19], except on car segmentation, and

it outperforms all the other comparison methods. However, BAS^{4}Net has more parameters due to the additional discriminator network.
- Regarding the lightweight models, compared to ABCNet [27] our BES-Net method with the ResNset18 backbone can achieve a slightly better performance on all metrics.
- All the results demonstrate the effectiveness of our proposed BES-Net method for enhancing the semantic context using boundary information to improve the intra-class semantic consistency.

Table 5. Quantitative results (%) comparing state-of-the-art methods on ISPRS Vaihingen and Potsdam test datasets.

	Method	Backbone	Per-Class F1-Score					mF1	OA
			Impervious Surfaces	Building	Low Vegetation	Tree	Car		
Vaihingen	DeepLabV3+ [11]	ResNet101	92.4	95.2	84.3	89.5	86.5	89.6	90.6
	PSPNet [10]	ResNet101	92.8	95.5	84.5	89.9	88.6	90.3	90.9
	IPSPNet [39]	ResNet101	89.6	91.5	82.0	88.3	68.4	84.0	87.8
	CVEO [40]	SDFCN139	90.5	92.4	81.7	88.5	79.4	86.5	88.3
	LWN [1]	ResNet101	91.0	94.9	79.2	88.6	88.4	87.6	88.9
	DANet [41]	ResNet101	91.6	95.0	83.3	88.9	87.2	89.2	90.4
	DDCM-Net [42]	ResNet50	92.7	95.3	83.3	89.4	88.3	89.8	90.4
	CASIA2 [43]	ResNet101	93.2	**96.0**	84.7	89.9	86.7	90.1	91.1
	HCANet [24]	ResNet101	92.5	95.0	84.2	89.4	84.0	89.0	90.3
	BAS^{4}Net [19]	ResNet101	93.3	95.8	85.0	90.1	**90.1**	**90.9**	91.3
	ABCNet [27]	ResNet18	92.7	95.2	84.5	89.7	85.3	89.5	90.7
	BES-Net (ours)	ResNet18	92.8	95.5	84.8	90.0	85.8	89.8	90.9
	BES-Net (ours)	ResNet50	93.0	**96.0**	**85.4**	90.0	88.3	90.6	91.2
	BES-Net (ours)	ResNet101	**93.4**	95.9	85.2	**90.3**	87.8	90.5	**91.4**
Potsdam	DeepLabV3+ [11]	ResNet101	93.0	95.9	87.6	88.2	96.0	92.1	90.9
	PSPNet [10]	ResNet101	93.4	97.0	87.8	88.5	95.4	92.4	91.1
	CVEO [40]	SDFCN139	91.2	94.5	86.4	87.4	95.4	91.0	89.0
	DDCM-Net [42]	ResNet50	92.9	96.9	87.7	**89.4**	94.9	92.4	90.8
	CCNet [44]	ResNet101	93.6	96.8	86.9	88.6	96.2	92.4	**91.5**
	HCANet [24]	ResNet101	93.1	96.6	87.0	88.5	96.1	92.3	90.8
	ABCNet [27]	ResNet18	93.5	96.9	87.9	89.1	95.8	92.6	91.3
	BES-Net (ours)	ResNet18	93.8	97.0	**88.1**	88.9	96.4	**92.9**	**91.5**
	BES-Net (ours)	ResNet50	**93.9**	**97.3**	87.9	88.5	**96.5**	92.8	91.4
	BES-Net (ours)	ResNet101	93.7	97.2	87.9	88.9	96.3	92.8	91.3

The experiments on the Potsdam dataset show that our proposed BES-Net method with the ResNet18 backbone obtains the best performance with respect to the two metrics mF1 and mIoU. The experiments suggest that our proposed method can boost the semantic segmentation performance, with more intra-class segmentation consistency, at a holistic semantic object level. Regarding the backbones, ResNet18 can achieve comparable (or slightly better) performance on the two metrics mF1 and mIoU, compared to ResNet50 and ResNet101. However, ResNet18 has considerably fewer parameters.

The experimental results indicate that for the Potsdam dataset with a fine resolution (5 cm), ResNet18 is sufficient to extract the spatial details and semantic information for parsing the HR remote sensing images. It can provide the best balance between the computational burden of forward inference and richness of feature learning.

4.4. Qualitative Analysis

Figures 10 and 11 show the visualization results of our BES-Net (ResNet18) method and the corresponding baseline (FCN_8s) on the Vaihingen and Potsdam test datasets, respectively. It can be seen that compared to the baseline, our BES-Net method significantly improved the segmentation performance, especially in the regions marked with red dashed circles (or boxes). Benefiting from the utilization of the BE, MSF, and BES modules, our BES-Net method could obtain a coherent and accurately labeled result in these heterogenous

regions which are hard to distinguish. The three modules together learn the features from the entire semantic object level, resulting in segmentation results with more complete boundaries. Some representative samples can be found in Figures 10 and 11. We can observe that:

- Some pixels are easily misclassified by the baseline method. (1) in Figures 10a and 11a, some regions of the building with a red roof are misclassified as low vegetation. (2) In Figure 11b, some regions of low vegetation with complex textures are misclassified as impervious surfaces. (3) In Figure 11c, some regions of the building with a gray roof are misclassified as clutter/background. The baseline method could not process those pixels belonging to one object in a holistic fashion, while our BES-Net method considering the semantic boundary had the global concept of an entire semantic object to improve the segmentation performance.
- As shown in Figure 10b, our BES-Net can separate two adjacent cars while the baseline may link them. This is because our BES-Net with boundary enhancement can generate clear boundaries and regular shapes.
- The object boundaries generated by our BES-Net are remarkably more complete than those from baseline, especially for regular objects such as buildings, as shown in Figures 10c and 11c. BES-Net can draw out the complete shape of the building with a clear boundary, while the baseline yields an incomplete building due to interruptions caused by different textures.
- All these results demonstrate that our BES-Net is more robust to adjacent object confusion and can effectively capture fine-structured objects with both boundary and semantic information at an entire semantic object level.

Moreover, to show the effectiveness of our proposed method for handling challenging situations when there is cloud shadow, some sample visualization results of our BES-Net (ResNet18) method and the corresponding baseline (FCN_8s) method are shown in Figure 12. As the red circles illustrate, the baseline method may miss some cars in shadow, while our BES-Net is able to perceive them. Our network can achieve better performance in challenging situations with shadows.

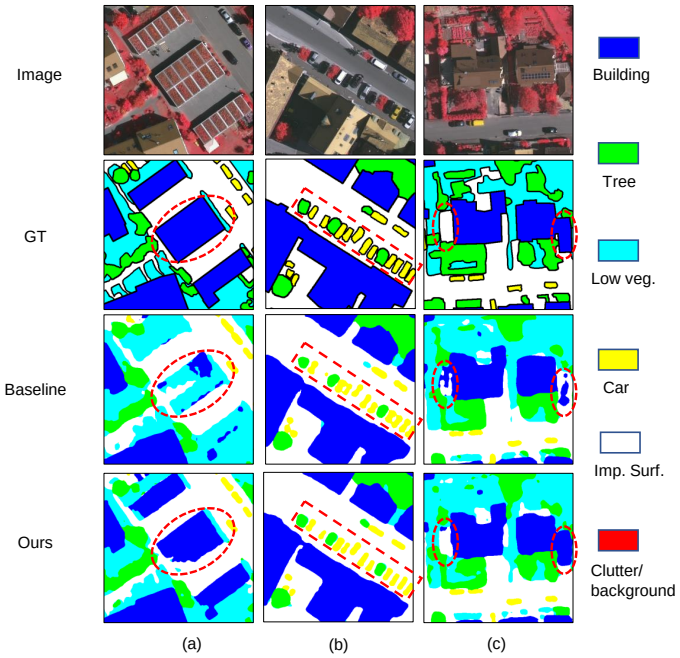


Figure 10. Visualization of semantic segmentation results on the Vaihingen test dataset. The red dashed circles (or boxes) are used to mark the regions which have been obviously improved by our method.

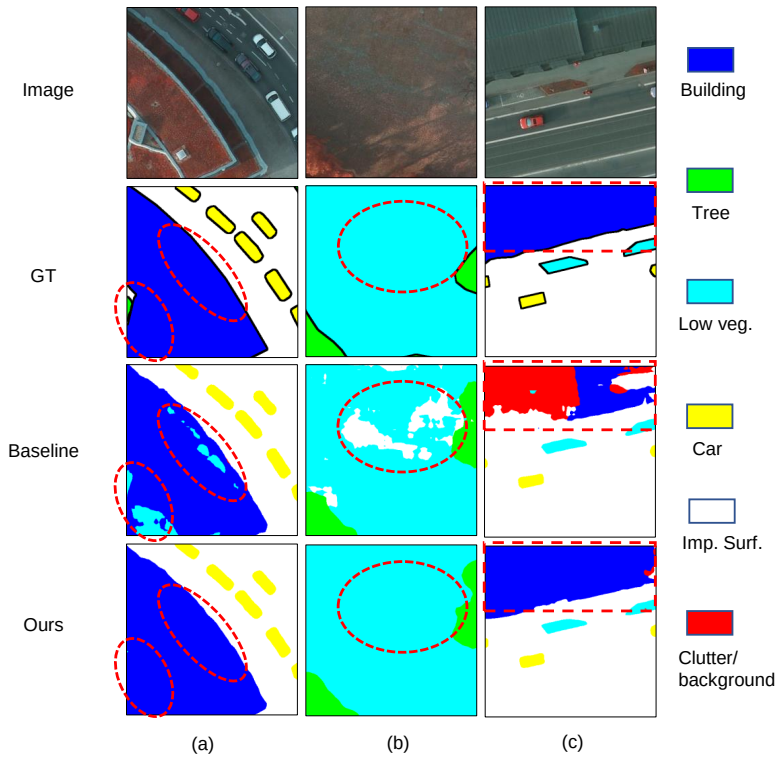


Figure 11. Visualization of semantic segmentation results on the Potsdam test dataset. The red dashed circles (or boxes) are used to mark the regions which have been obviously improved by our method.

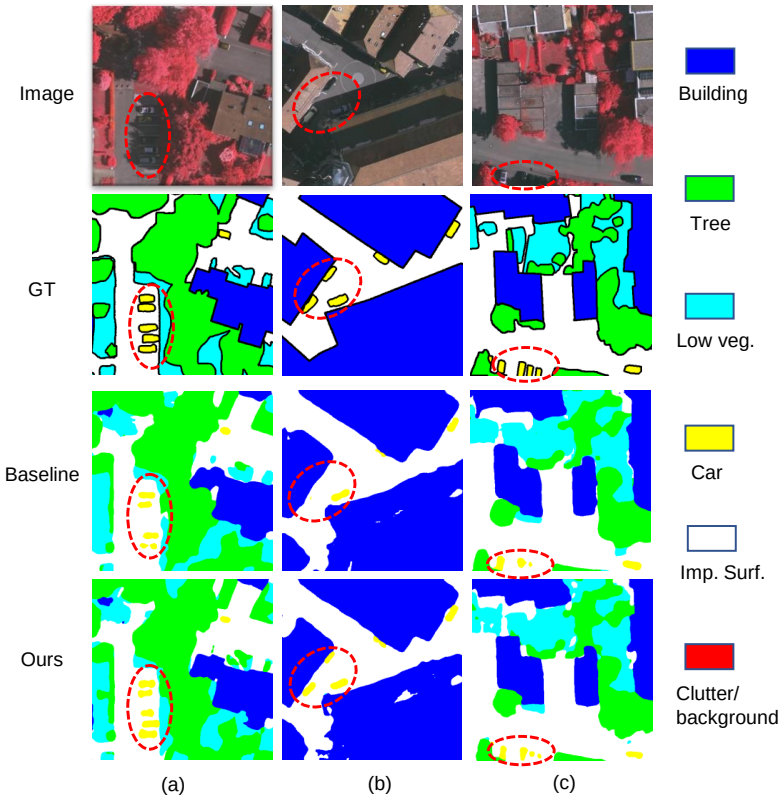


Figure 12. Visualization of semantic segmentation results on some sample images with shadow. The red dashed circles are used to mark the regions with shadow which have been obviously improved by our method.

4.5. Computational Complexity

We compared the computational complexity with state-of-the-art methods such as ABCNet [27], FANet [45], MAResU-Net [46], and SwiftNet [47]. Model parameters and computation FLOPs are also listed for comparison in Table 6. Note that for a fair comparison, we used the same backbone (ResNet18) network to evaluate the computational complexity. We can see that compared to the state-of-the-art lightweight method ABCNet, our BES-Net achieved better performance with fewer parameters and FLOPs, maintaining both low computational cost and high accuracy simultaneously.

Table 6. The computational complexity of our BES-Net method and other lightweight methods.

Model	Backbone	Params (M)	FLOPs (G)	mF1	OA
PSPNet [10]	ResNet18	24.0	12.6	79.0	87.7
DANet [41]	ResNet18	12.7	9.9	79.6	88.2
FANet [45]	ResNet18	21.7	13.8	85.4	88.9
MAResU-Net [46]	ResNet18	25.4	16.2	87.7	90.1
SwiftNet [47]	ResNet18	18.8	34.2	88.3	90.2
ABCNet [27]	ResNet18	14.1	18.7	89.5	90.7
BES-Net(ours)	ResNet18	13.6	15.8	89.8	90.9

5. Conclusions and Discussion

We presented a boundary enhancing semantic context network (BES-Net) in this paper that could improve the semantic segmentation performance for parsing high-resolution remote sensing images. The main idea was that we explicitly, not implicitly, used the well-extracted boundary to enhance the semantic context for semantic segmentation. BES-Net simultaneously takes the boundary and semantic context information into account using three designed modules. The BES module enhances the fused multi-scale semantic context extracted from the MSF module, using the boundary information extracted from the BE module to boost the semantic segmentation performance, giving more intra-class segmentation consistency at a holistic semantic object level. Experiments on the ISPRS Vaihingen and Potsdam datasets showed that when only the boundary information or multi-scale semantic context was incorporated, the segmentation performance was slightly improved, while adding our BES module to explicitly enhance the fused multi-scale semantic segmentation using boundary information improved the performance considerably. This demonstrates the progressiveness and superiority of our proposed BES-Net method, even compared to the current state-of-the-art methods.

Although it achieves a relatively fine combination of boundary information and semantic context, the proposed BES-Net still has room for improvement. (1) *Efficiency.* At present, due to the high dimensionality, our BE module still has a relatively high computational cost. (2) *Performance.* To reduce the computational cost, our BES module only utilizes two simple parameter-free mathematical operations(element-wise addition and element-wise multiplication), to fuse two complementary features. There should be more effective ways to design the BES module. Therefore, our future work will focus on further optimizing the BE and BES modules, reducing the complexity of the BE module while maintaining its performance and adopting more effective methods for using boundary information to enhance semantic context. Furthermore, focusing on more challenging situations in the actual scenarios, e.g., images with clouds and shadows, is also an interesting topic.

Author Contributions: Conceptualization, H.L. and X.Z.; methodology, F.C. and H.L.; validation, F.C. and Z.Z.; writing—original draft preparation, H.L. and F.C.; writing—review and editing, H.L., X.Z., and X.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under grants 62001063, U20A20157, 61971072, and U2133211, and in part by the China Postdoctoral Science Foundation under grant 2020M673135 and Chongqing Postdoctoral Research Program under grant XmT2020050.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: We utilized two public 2D semantic labeling datasets, Vaihingen and Potsdam, provided by the International Society for Photogrammetry and Remote Sensing (ISPRS), https://www2.isprs.org/commissions/comm2/wg4/benchmark/semantic-labeling/, accessed on 26 February 2022.

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

References

1. Liu, S.; Cheng, J.; Liang, L.; Bai, H.; Dang, W. Light-Weight Semantic Segmentation Network for UAV Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 8287–8296. [CrossRef]
2. Tong, X.Y.; Xia, G.; Lu, Q.; Shen, H.; Li, S.; You, S.; Zhang, L. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sens. Environ.* **2020**, *237*, 111322. [CrossRef]
3. Zhu, X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
4. Ma, L.; Liu, Y.; liang Zhang, X.; Ye, Y.; Yin, G.; Johnson, B. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *152*, 166–177. [CrossRef]
5. Cheng, J.; Ji, Y.; Liu, H. Segmentation-Based PolSAR Image Classification Using Visual Features: RHLBP and Color Features. *Remote Sens.* **2015**, *7*, 6079–6106. [CrossRef]
6. Yang, Y.; Hallman, S.; Ramanan, D.; Fowlkes, C.C. Layered Object Models for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 1731–1743. [CrossRef]
7. Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef]
8. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
9. 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; Springer: Cham, Switzerland, 2015; pp. 234–241.
10. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6230–6239.
11. 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.
12. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context encoding for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7151–7160.
13. Liu, Y.; Xie, Y.; Yang, J.; Zuo, X.; Zhou, B. Target Classification and Recognition for High-Resolution Remote Sensing Images: Using the Parallel Cross-Model Neural Cognitive Computing Algorithm. *IEEE Geosci. Remote Sens. Mag.* **2020**, *8*, 50–62. [CrossRef]
14. Li, X.; Xu, F.; Xia, R.; Lyu, X.; Gao, H.; Tong, Y. Hybridizing Cross-Level Contextual and Attentive Representations for Remote Sensing Imagery Semantic Segmentation. *Remote Sens.* **2021**, *13*, 2986. [CrossRef]
15. Zheng, X.; Huan, L.; Xia, G.; Gong, J. Parsing very high resolution urban scene images by learning deep ConvNets with edge-aware loss. *ISPRS J. Photogramm. Remote Sens.* **2020**, *170*, 15–28. [CrossRef]
16. Li, X.; Li, X.; Zhang, L.; Cheng, G.; Shi, J.; Lin, Z.; Tan, S.; Tong, Y. Improving semantic segmentation via decoupled body and edge supervision. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 435–452.
17. Takikawa, T.; Acuna, D.; Jampani, V.; Fidler, S. Gated-scnn: Gated shape cnns for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 5229–5238.
18. Ma, H.; Yang, H.; Huang, D. Boundary Guided Context Aggregation for Semantic Segmentation. *arXiv* **2021**, arXiv:2110.14587.
19. Sun, X.; Shi, A.; Huang, H.; Mayer, H. BAS4Net: Boundary-Aware Semi-Supervised Semantic Segmentation Network for Very High Resolution Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 5398–5413. [CrossRef]
20. Marmanis, D.; Schindler, K.; Wegner, J.D.; Galliani, S.; Datcu, M.; Stilla, U. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS J. Photogramm. Remote Sens.* **2018**, *135*, 158–172. [CrossRef]

21. 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, NV, USA, 27–30 June 2016; pp. 770–778.
22. Fan, M.; Lai, S.; Huang, J.; Wei, X.; Chai, Z.; Luo, J.; Wei, X. Rethinking BiSeNet For Real-time Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9716–9725.
23. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Su, J.; Wang, L.; Atkinson, P.M. Multiattention Network for Semantic Segmentation of Fine-Resolution Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–13. [CrossRef]
24. Bai, H.; Cheng, J.; Huang, X.; Liu, S.; Deng, C. HCANet: A Hierarchical Context Aggregation Network for Semantic Segmentation of High-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
25. 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]
26. Chen, L.C.; Yang, Y.; Wang, J.; Xu, W.; Yuille, A.L. Attention to scale: Scale-aware semantic image segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3640–3649.
27. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Wang, L.; Atkinson, P.M. ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of Fine-Resolution remotely sensed imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *181*, 84–98. [CrossRef]
28. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
29. Yang, X.S.; Li, S.; Chen, Z.; Chanussot, J.; Jia, X.; Zhang, B.; Li, B.; Chen, P. An Attention-Fused Network for Semantic Segmentation of Very-High-Resolution Remote Sensing Imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *177*, 238–262. [CrossRef]
30. Bertasius, G.; Shi, J.; Torresani, L. Semantic segmentation with boundary neural fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3602–3610.
31. Ke, T.W.; Hwang, J.J.; Liu, Z.; Yu, S.X. Adaptive affinity fields for semantic segmentation. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 587–602.
32. Bertasius, G.; Torresani, L.; Yu, S.X.; Shi, J. Convolutional random walk networks for semantic image segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 858–866.
33. Ding, H.; Jiang, X.; Liu, A.Q.; Thalmann, N.M.; Wang, G. Boundary-aware feature propagation for scene segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 6819–6829.
34. Zhang, C.; Jiang, W.; Zhao, Q. Semantic Segmentation of Aerial Imagery via Split-Attention Networks with Disentangled Nonlocal and Edge Supervision. *Remote Sens.* **2021**, *13*, 1176. [CrossRef]
35. Han, H.Y.; Chen, Y.C.; Hsiao, P.Y.; Fu, L.C. Using Channel-Wise Attention for Deep CNN Based Real-Time Semantic Segmentation With Class-Aware Edge Information. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 1041–1051. [CrossRef]
36. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014.
37. Li, X.; Zhao, H.; Han, L.; Tong, Y.; Tan, S.; Yang, K. Gated fully fusion for semantic segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11418–11425.
38. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
39. Yu, B.; Yang, L.; Chen, F. Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 3252–3261. [CrossRef]
40. Chen, G.; Zhang, X.; Wang, Q.; Dai, F.; Gong, Y.; Zhu, K. Symmetrical dense-shortcut deep fully convolutional networks for semantic segmentation of very-high-resolution remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 1633–1644. [CrossRef]
41. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
42. Liu, Q.; Kampffmeyer, M.; Jenssen, R.; Salberg, A.B. Dense dilated convolutions' merging network for land cover classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 6309–6320. [CrossRef]
43. Liu, Y.; Fan, B.; Wang, L.; Bai, J.; Xiang, S.; Pan, C. Semantic labeling in very high resolution images via a self-cascaded convolutional neural network. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 78–95. [CrossRef]
44. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 603–612.
45. Hu, P.; Perazzi, F.; Heilbron, F.C.; Wang, O.; Lin, Z.L.; Saenko, K.; Sclaroff, S. Real-Time Semantic Segmentation with Fast Attention. *IEEE Robot. Autom. Lett.* **2021**, *6*, 263–270. [CrossRef]
46. Li, R.; Su, J.; Duan, C.; Zheng, S. Multistage Attention ResU-Net for Semantic Segmentation of Fine-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
47. Orsic, M.; Segvic, S. Efficient semantic segmentation with pyramidal fusion. *Pattern Recognit.* **2021**, *110*, 107611. [CrossRef]