remote sensing

MDPI

Article

Multi-Scale Context Aggregation for Semantic Segmentation of Remote Sensing Images

Jing Zhang [1,†], Shaofu Lin [1,2], Lei Ding [3,*] and Lorenzo Bruzzone [3]

[1] Faculty of Information Technology, Beijing University of Technology, Chaoyang District, Beijing 100022, China; zhangjing@emails.bjut.edu.cn (J.Z.); linshaofu@bjut.edu.cn (S.L.)

[2] Beijing Institute of Smart City, Beijing University of Technology, Chaoyang District, Beijing 100022, China

[3] Remote Sensing Laboratory, Department of Information Engineering and Computer Science, University of Trento, Via Sommarive 5, Trento 38122, Italy; lorenzo.bruzzone@unitn.it

* Correspondence: lei.ding@unitn.it; Tel.: +39-3318163585

† Current address: NO.100 Pingle Garden, Chaoyang District, Beijing 100022, China.

Received: 15 January 2020; Accepted: 17 February 2020; Published: 24 February 2020

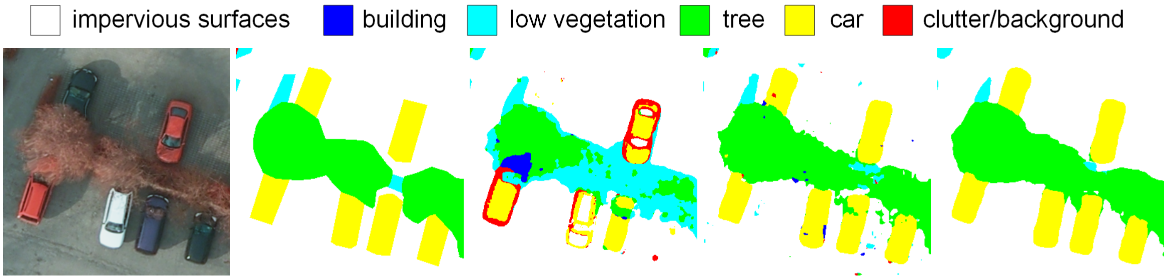
Abstract: The semantic segmentation of remote sensing images (RSIs) is important in a variety of applications. Conventional encoder-decoder-based convolutional neural networks (CNNs) use cascade pooling operations to aggregate the semantic information, which results in a loss of localization accuracy and in the preservation of spatial details. To overcome these limitations, we introduce the use of the high-resolution network (HRNet) to produce high-resolution features without the decoding stage. Moreover, we enhance the low-to-high features extracted from different branches separately to strengthen the embedding of scale-related contextual information. The low-resolution features contain more semantic information and have a small spatial size; thus, they are utilized to model the long-term spatial correlations. The high-resolution branches are enhanced by introducing an adaptive spatial pooling (ASP) module to aggregate more local contexts. By combining these context aggregation designs across different levels, the resulting architecture is capable of exploiting spatial context at both global and local levels. The experimental results obtained on two RSI datasets show that our approach significantly improves the accuracy with respect to the commonly used CNNs and achieves state-of-the-art performance.

Keywords: semantic segmentation; convolutional neural network; deep learning; image analysis; remote sensing

1. Introduction

Images collected from aerial and satellite platforms are widely used in a variety of applications, such as land-use mapping, urban resources management, and disaster monitoring. Semantic segmentation, namely the pixel-wise classification of images, is a crucial step for the automatic analysis and exploitation in applications of these data. The rise of convolutional neural networks (CNNs) and the emergence of fully convolutional networks (FCNs) [1] have led to a breakthrough in the semantic segmentation of remote sensing images (RSIs) [2]. Typical CNN architectures used in visual recognition tasks employ cascade spatial-reduction operations to force the networks to learn intrinsic representations of the observed objects [3]. However, this so-called "encoding" design has the side effect of losing spatial information. The classification maps produced by encoding networks usually suffer a loss of localization accuracy (e.g., the boundaries of classified objects are blurred, and some small targets may be neglected). Although there are "decoding" designs to recover spatial information by using features extracted from early layers of the CNNs [4,5], their effectiveness is limited due to the gap between the high-level and low-level features in both semantic information and spatial distribution [6].

To solve this problem, we introduce the use of the high-resolution network (HRNet) [7] to improve the embedding of high-resolution features. Instead of encoding spatially reduced features yielded by serial convolutions, the HRNet employs multi-branch parallel convolutions to produce low-to-high resolution feature maps. The multi-scale branches are fully connected so that the information flows smoothly between different branches. Therefore, the high-resolution branches have powerful semantic representations without losing spatial details. Figure 1 shows examples of classified features yielded from the early layer of ResNet50 [3] and from the high-resolution branch of HRNet. The compared features have the same spatial resolution; however, those yielded by the HRNet have better semantic representations.



(**a**) Input image (IRRG); (**b**) Ground truth; (**c**) ResNet50; (**d**) HRNet (before using ASP); (**e**) HRNet (after using ASP).

Figure 1. Examples of classified high-resolution features.

Moreover, building on top of the HRNet, we propose a novel architecture that further enhances the aggregation of the context information. Contextual information is known to be crucial for the semantic segmentation of remote sensing images [8,9], since it offers important clues to the recognition of objects and regions. In the original design of the HRNet, the encoded features from four low-to-high resolution branches are concatenated together to produce the results. Although this brings a significant improvement over the traditional encoding networks (such as VGGNet [10] and ResNet), the maximum valid receptive field (RF) of the network remains the same and the contextual information is not fully exploited. To improve the aggregation of contextual information, two scale-related processing modules are introduced into the HRNet architecture. Specifically, the low-resolution branches in the network are incorporated in a spatial-reasoning module to learn the long-range spatial correlations, while the high-resolution branches are enhanced by introducing an adaptive spatial pooling (ASP) module to aggregate local contexts. Figure 1b shows an example of the ASP module used on high-resolution features. Therefore, the resulting architecture is capable of aggregating local-to-global level contextual information from small-to-large scales.

In summary, the main contributions in this work are as follows:

- Building on top of the HRNet architecture, we propose a multi-scale context aggregation network to aggregate multi-level spatial correlations. In this network, scale-related information aggregation modules are designed to enlarge the RF of each branch, followed by a low-to-high fusion of multi-scale features to generate the classification results.
- We propose an ASP module that is able to incorporate resolution-related context information to enhance the semantic representation of CNN feature maps.
- We test the proposed module and architecture through both an ablation study and comparative experiments in relation to other methods. Experimental results show that the proposed approach achieves the state-of-the-art accuracy on the two considered RSI datasets (the Potsdam and the Vaihingen datasets).

The remainder of this paper is organized as follows. Section 2 introduces the background of semantic segmentation with CNNs. Section 3 describes in detail the proposed architecture. Section 4 illustrates the experimental settings. Section 5 reports the experimental results and discusses the performance of the tested methods. In Section 6, we draw conclusions from the study.

2. Related Work

In this section, first, we briefly review the development of a CNN for the task of semantic segmentation, including both the design of the architectures and some state-of-the-art modules for aggregating context information. We then review the use of CNNs in the semantic segmentation of RSIs and discuss the limitations in existing studies.

2.1. Encoder-Decoder Architectures

Encoder-decoder-based architectures have been commonly used since the design of the fully convolutional network (FCN) [1], the first CNN designed for dense image-labeling tasks. The "encoder" networks are usually backbone CNNs that use cascade pooling and convolution layers to learn semantic information about the objects. By contrast, the "decoder" parts are usually upsampling or deconvolution operations to recover the lost spatial resolution of the encoded features. In an FCN, the encoded features have only 1/32 of their original resolutions; thus, they suffer from a loss of spatial accuracy. As an alternative, a symmetrical "encoder-decoder" design has been introduced in UNet [5]. This network is designed for the semantic segmentation of medical images, and its minimum scaling ratio is 1/8. The multi-level encoding features are directly concatenated in the decoding stage to aggregate more spatial information. SegNet [4] has a similar design but uses pooling indexes to record and recover spatial information. RefineNet [11] strengthens the decoder with a multi-path fusion of features from different levels. The fusion of multi-level features is further enhanced in Exfuse [6] by using both pixel-wise sum and concatenation operations. The connection between high-level and low-level features is also introduced in DeepLabv3+ [12]. DenseASPP [13] and UNet++ [14] use dense skip-connections to improve the transition and re-use of features. One of the limitations of these encoder-decoder designs is that there is a significant loss of spatial details during the encoding stage, and the decoders are still not powerful enough to recover all the lost information.

2.2. Aggregation of Context Information

The use of context information is essential to determine the object categories due to the common intra-class inconsistency and inter-class indistinction problems. Many approaches add modules/blocks at the top of encoding networks to enlarge their valid receptive fields (RFs) and integrate more context information. In [15], the importance of the RF is discussed, and the global pooling operation is introduced to learn the scene-level global context. PSPNet [16] extends the use of global pooling to image sub-regions and proposes a parallel spatial pooling design to aggregate multi-scale context information. Dilated convolution is another design that can enlarge the RF of CNNs without significantly increasing the calculations [17,18]. Combining dilated convolutions and the multi-level pooling design in the PSPNet, the atrous spatial pyramid pooling (ASPP) module was proposed in [19] and improved in [12,13,20]. The attention mechanism, another kind of context aggregation design, employs a sigmoid function after the global pooling operations to generate "attention" descriptors [21–23]. In this way, CNNs give biased focus on the image content based on the global-level information. These global and local context aggregation modules are studied and improved in this work to better serve the semantic segmentation of RSIs.

2.3. Semantic Segmentation of RSIs

The semantic segmentation on RSIs has attracted great research interests after the rising of CNNs and the release of several open datasets/contests such as the ISPRS Benchmarks, the DeepGlobe contest, and the SpaceNet competition. Several studies incorporate multiple models to increase the prediction certainty [24–26]. The prediction of object contours is an issue of concern. The detection of edges is explicitly added in [27], while an edge loss to enhance the preservation of object boundaries is introduced in [28]. The utilization of other types of data (e.g., Lidar data, digital surface models, and OpenStreetMap) is also widely studied [29–32]. However, there are limited studies focused on the special characteristics of RSIs (e.g., a large spatial size, a fixed imaging angle, and a small number of classes). In this work, we consider these characteristics when designing the proposed specific modules.

3. The Proposed Approach

In this section, we illustrate in detail the proposed multi-level context aggregation architecture for the semantic segmentation of RSIs. First, we give an illustration of the HRNet (our baseline network) architecture. After that, we present an overview of the proposed architecture, including the motivation of its design and the strategy used to incorporate together different branches to collect the complementary information. Finally, we describe the specific processing modules for each branch, including the proposed ASP module and the introduced spatial reasoning (SR) module.

3.1. Baseline: The High-Resolution Network (HRNet)

As discussed in Section 2, the use of cascade pooling operations in encoder networks results in a loss of spatial accuracy, which can hardly be recovered by the decoders. To overcome this limitation, the HRNet [7,33] introduces a multi-scale parallel design. As illustrated in Figure 2, there are four parallel branches in the HRNet architecture, which correspond to the four high-to-low spatial resolutions. The upper branch of the HRNet remains high-resolution, so the spatial details are kept through the convolutions. Meanwhile, strided convolutions are applied in the lower branches to reduce the spatial size of feature maps and enlarge the RFs. The scale-relevant features are aggregated in each branch with a set of convolution blocks. Connections through different branches are established in a fully connected fashion after each convolution block. Finally, the network produces four groups of feature maps at different resolutions. They are first spatially enlarged to the same spatial size (as the same size of Branch 1) and then concatenated together along the channel dimension. The fused features can be used to produce the segmentation result. To reduce the computational load, the early layers of the HRNet down-sample the inputs to 1/4 of their original size. Therefore, the four branches of HRNet correspond to 1/4, 1/8, 1/16, and 1/32 of the original input size, respectively.

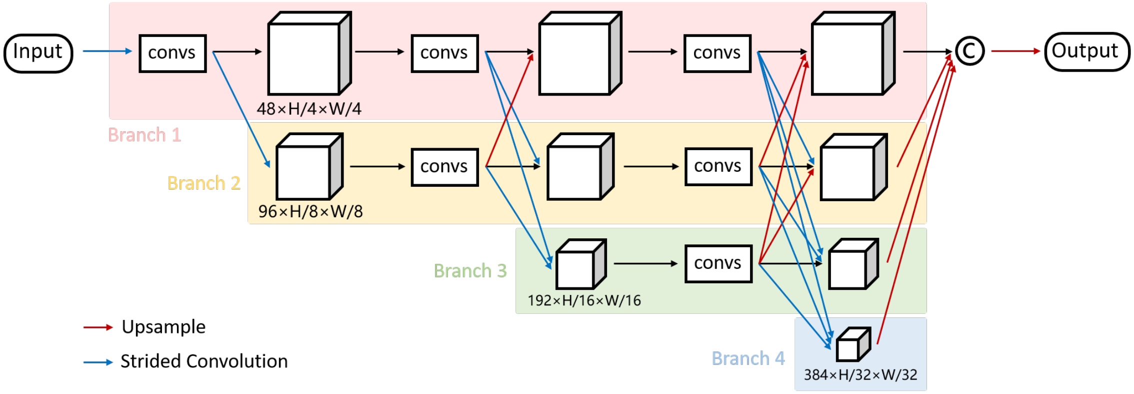


Figure 2. The baseline high-resolution network. Each 'convs' box refers to a convolution block consisting of cascaded convolutional layers.

3.2. Overview of the Proposed Architecture

In the baseline HRNet, features from the four parallel branches are concatenated together to generate the output. Although this architecture improves the preservation of spatial information, it fails to make full use of the contextual information. Thus, we improve this architecture by introducing scale-related context aggregation modules into different branches.

As shown in Figure 3, the proposed multi-scale context aggregation network is built on top of the four parallel branches of the HRNet. Note that, here, HRNet serves only as a feature extractor instead of directly producing the segmentation results. The feature maps extracted from the HRNet have different spatial size and channel numbers. The high-resolution features are related to small RFs; thus, they are more or less fragmented. We propose an ASP module to embed local context information from different levels into these features. Meanwhile, the low-resolution features (extracted from the third and fourth branches) are considered rich in semantic information even if having small spatial size. Therefore, we introduce the SR module to model their long-range spatial correlations. This module is calculation-intensive and can only be used to process small-size features. As a result, the proposed network is able to learn both the long-range and the local context information through the four low-to-high resolution branches. Finally, a top-down feature fusion across the branches is performed with the upsampling and channel-wise concatenation operations.

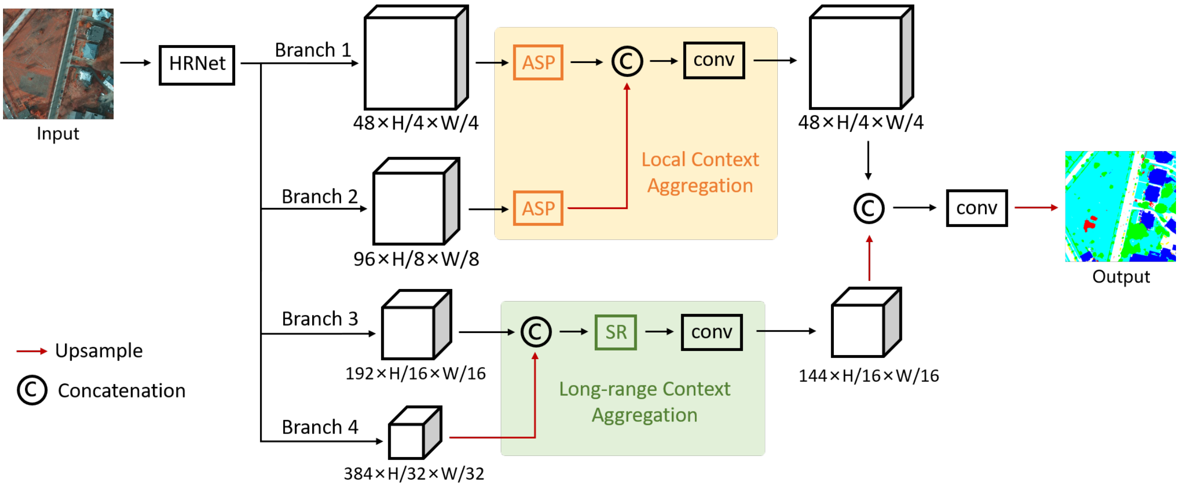


Figure 3. Overview of the proposed multi-scale context aggregation network.

Let $\mathbf{B}_1$, $\mathbf{B}_2$, $\mathbf{B}_3$, and $\mathbf{B}_4$ be the four branches of features produced by the HRNet. A high-resolution feature $\mathbf{F}_{high}$ can be generated by concatenating $\mathbf{B}_1$ and $\mathbf{B}_2$ after the processing of ASP modules:

$$\mathbf{F}_{high} = \{ASP(\mathbf{B}_1), f_u[ASP(\mathbf{B}_2)]\} \tag{1}$$

where f_u is an upsampling operation. Meanwhile, the low-resolution branches are concatenated and processed by the SR module, denoted as

$$\mathbf{F}_{low} = SR[\mathbf{B}_3, f_u(\mathbf{B}_4)] \tag{2}$$

Finally, a fusion between $\mathbf{F}_{high}$ and $\mathbf{F}_{low}$ is performed to obtain $\mathbf{F}_{fuse}$, i.e.

$$\mathbf{F}_{fuse} = \{R_1(\mathbf{F}_{high}), f_u[R_2(\mathbf{F}_{low})]\} \tag{3}$$

where R_1 and R_2 are both 1×1 convolutions that reduce the channel dimensions. $\mathbf{F}_{fuse}$ goes through another convolution block to be embedded into the outputs.

3.3. Adaptive Spatial Pooling Module

The pooling operation refers to the calculation of generating a descriptor based on the mean/extreme value of a given region. This descriptor offers important clues of the local contexts; thus, it can be used to learn a biased focus on certain categories. Most existing works perform pooling operations at a global level (whole-image-wise) regardless of the size of the input images; thus, they are less meaningful. We address this problem by proposing an ASP module, which limits the calculation of descriptors into local regions and adjusts adaptively according to the scale of the inputs.

The ASP module is an extension of the pyramid spatial pooling (PSP) module [16]. The original PSP module is designed for the semantic segmentation of natural images (which usually have a relatively small spatial size). It contains average pooling layers with different pooling windows to aggregate information from various scales. However, in the PSP module, the size of the pooling windows is set according to a set of predefined values (e.g., [1–3,6]). As a result, the spatial sizes of the generated local descriptors are fixed and do not change with the input size. In the semantic segmentation of RSIs, the training stage is usually performed on small image batches, whereas the testing stage can be performed on large-scale images. Since there could be a tremendous difference in input size during the training and the testing stage, a direct use of the original PSP module may result in inconsistent feature representations and thus affect the performance. Moreover, in the PSPNet, the PSP module is added on top of the ResNet. Since the low-resolution features yielded by the late layers of ResNet already have relatively large RFs, the improvement of using the PSP module is limited.

Considering both the characteristics of RSIs and the multi-scale design of the HRNet, in the ASP, the sizes of pooling windows are calculated adaptively according to the down-sampling rate of the features. First, we define a set of expected RFs of the generated local descriptors S_{RF} (empirically set to [40, 80, 160, 320]). The size of pooling windows S_{pool} can then be calculated according to the down-sampling rate r_d of the input features. Finally, the size of generated feature maps S_d can be calculated according to S_{pool} and the input size S_{in} as follows:

$$S_d = S_{in}/S_{pool} = S_{in}/\left(S_{RF}/r_d\right) \tag{4}$$

The detailed design of the ASP is shown in Figure 4. Given an input feature $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$, a 1×1 convolution is applied to reduce the channels of $\mathbf{X}$ to $c = C/4$. After that, four parallel pooling operations are built adaptively to generate the descriptors $\mathbf{d} \in \mathbb{R}^{c \times s \times s}, s \in S_d$. The descriptors are further upsampled to the same spatial size as $\mathbf{X}$ and concatenated together with $\mathbf{X}$, denoted as $\mathbf{D} \in \mathbb{R}^{2C \times H \times W}$. Finally, a channel-reduction convolution is performed on $\mathbf{D}$ to recover the size to $C \times H \times W$.

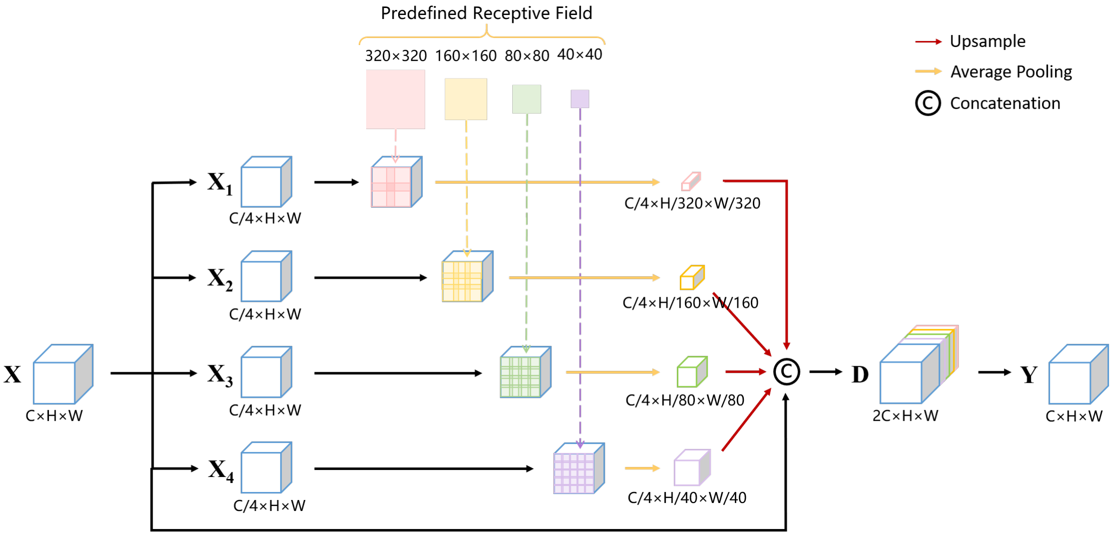


Figure 4. Detailed design of the adaptive spatial pooling module.

In the proposed architecture, the ASP modules are added to the lower branches of the HRNet. These branches are related to high-resolution features with small RFs; thus, the use of ASP modules can greatly enlarge the RFs and improve the aggregation of local context information. The effect of using ASP modules on each branch is further discussed in Section 5.

3.4. Spatial Reasoning Module

This module is inspired by the idea of non-local reasoning presented in [8,34,35]. The motivation is to model the long-range spatial correlations across different image sub-regions. Figure 5 shows the spatial reasoning module. Given an input feature map $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, the intuition is to generate a positional attention map $\mathbf{A} \in \mathbb{R}^{HW\times HW}$ to highlight the importance between pixel-pairs, and, combining $\mathbf{X}$ and $\mathbf{A}$, to generate the enhanced output $\mathbf{E}$. The calculations from $\mathbf{X}$ to $\mathbf{E}$ can be summarized as follows:

1. Apply 1×1 convolutions to $\mathbf{X}$ to generate three projected features, denoted as $\mathbf{X'} \in \mathbb{R}^{C\times H\times W}$ and $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{C'\times H\times W}$. C' is C multiplied by a reduction ratio r.
2. Reshape $\mathbf{X'}$ to $\mathbb{R}^{C\times N}$ and $\mathbf{U}, \mathbf{V}$ to $\mathbb{R}^{C'\times N}$, where $N = H \times W$.
3. Calculate the positional attention map $\mathbf{A} \in \mathbb{R}^{N\times N}$ using $\mathbf{V}$, the transpose of $\mathbf{U}$, and a softmax function σ:
$$\mathbf{A} = \sigma\left(\mathbf{U}^T\mathbf{V}\right) \tag{5}$$
4. Generate the enhanced feature $\mathbf{E}$ with $\mathbf{X}$, $\mathbf{X'}$, and $\mathbf{A}$:
$$\mathbf{E} = f_r\left(\mathbf{X'}\mathbf{A}^T\right) \oplus \mathbf{X} \tag{6}$$
where f_r denotes a reshape function that transforms the size of the feature into $\mathbb{R}^{C\times H\times W}$, and $\oplus$ denotes the element-wise summation operation.

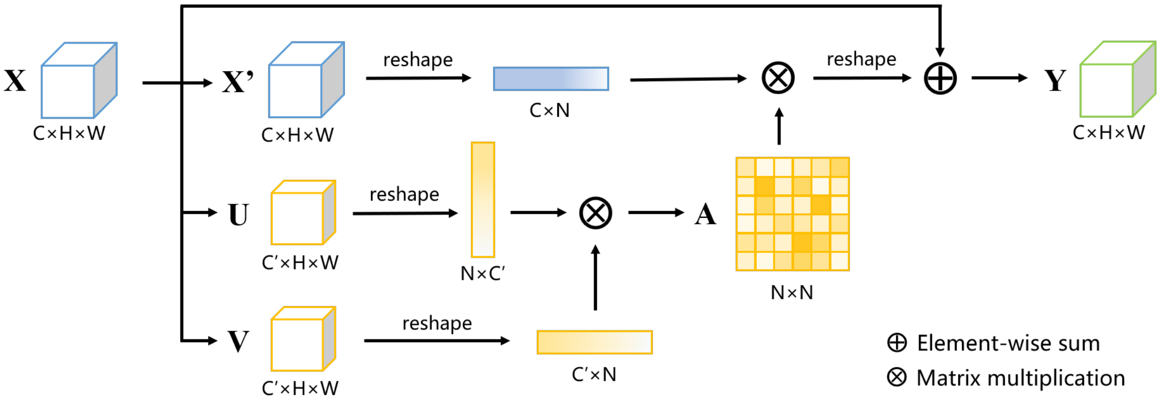


Figure 5. Detailed design of the spatial reasoning module.

4. Description Datasets and Implementation Settings

In this section, we introduce the experimented datasets, the evaluation metrics, and some implementation details.

4.1. Datasets

We test the proposed approach on two commonly used and high-quality RSI benchmark datasets: the Potsdam and the Vaihingen datasets. They are both published under the ISPRS 2D labeling contest (http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html). The Potsdam dataset covers a typical historic city with large building blocks, narrow streets, and dense settlement structures, while the Vaihingen dataset shows a relatively small village with many detached buildings.

4.1.1. The Potsdam Dataset

The dataset contains 38 tiles extracted from true orthophotos and the corresponding registered normalized digital surface models (DSMs). Twenty-four image patches are used for the training phase, and the remaining 14 for the testing phase. The orthophotos contain four spectral bands: red, green, blue, and infrared (RGBIR). Each image has the same spatial size of 6000 × 6000 pixels. The ground sampling distance (GSD) of this dataset is 5 cm. There are six categories in the reference maps: impervious surfaces, building, low vegetation, tree, car, and clutter/background. The clutter class contains uninteresting objects (e.g., swimming pools, containers, and trash cans) but should be classified and calculated in the evaluation.

4.1.2. The Vaihingen Dataset

The dataset contains 33 tiles extracted from true orthophotos and the corresponding registered normalized digital surface models (DSMs). Sixteen image patches are used for training phase, and the remaining 17 for the testing phase. The orthophotos contain three spectral bands: infrared, red, and green (IRRG). The spatial size of images varies from 1996 × 1995 to 3816 × 2550 pixels. The ground sampling distance (GSD) of this dataset is 9 cm. The defined object classes are the same as those in the Potsdam dataset.

4.2. Evaluation Metrics

The performance of the tested methods is evaluated by two measurements: the overall accuracy (OA) and the F1 score. These are the evaluation metrics advised by the data publisher [36] and are commonly used in the existing literature [26,27]. Overall accuracy is the percentage of correctly classified pixels among all the pixels. The F1 score of a certain object category is the harmonic mean of the precision and recall, calculated as

$$\mathrm{F1} = 2 \times \frac{\mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{7}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{8}$$

where TP, FP, and FN stand for true positive, false positive, and false negative, respectively. True positive denotes the number of correctly classified pixels, false positive denotes the number of pixels of other classes that are wrongly classified into a specific class, and false negative denotes the number of pixels of a given class that are wrongly classified into other categories.

4.3. Implementation Settings

The proposed architecture is implemented using the PyTorch library [37]. The hardware device is a server equipped with 2 Nvidia P100 GPUs (each has 16 GB of memory) and 128 GB of RAM.

The same preprocessing, i.e., data augmentation and weight initialization settings, is used in all the experiments. The DSMs are concatenated with the RSIs as input data, so there are five channels for the Potsdam dataset and four channels for the Vaihingen dataset. Due to the limitation of computational resources, the input data are cropped using a 512 × 512 window during the training phase. However, the prediction for the test set is performed whole-image-wise to obtain an accurate evaluation of the compared methods. Random-flipping and random-cropping operations are conducted during each iteration of the training phase as a way of data augmentation.

The features pretrained on the PASCAL-Context dataset (https://cs.stanford.edu/~roozbeh/pascal-context/) (provided by the author of HRNet (https://github.com/HRNet/HRNet-Semantic-Segmentation)) are used to initialize the weights of HRNet. During the training stage, the learning rate is initialized to 0.01 and the Adam algorithm [38] is adopted to optimize the learning rate after each iteration. The cross-entropy loss is used to calculate the differences between the predictions and the reference maps.

5. Experimental Results

5.1. Ablation Study

In order to verify the effectiveness of the proposed modules and the overall architecture, ablation studies are performed on the two datasets. Since the tested modules operate on different branches of the HRNet, the feature maps from each branch are also evaluated (before and after the use of the corresponding processing module) to quantify their contributions to the final results. The results are obtained by using an additional classifier to classify the tested features and then evaluating the outputs.

Table 1 shows the results of the ablation study on the Potsdam dataset. Branch 1 refers to the branch with the highest resolution (scaling ratio: 1/4), while Branch 4 refers to the branch with the lowest resolution (scaling ratio: 1/32). The OAs of the four branches are 87.41, 90.59, 90.14, and 89.61%, respectively. The OA and the average F1 score of concatenating the four branches are 91.01 and 92.18%, respectively. Therefore, the second branch (scaling ratio: 1/8) contributes the most to the final results. After the use of the ASP module, the OAs of the first and the second branches increased by 3.15 and 0.62%, respectively. The average F1 scores increased by 3.00 and 0.67%. This proves the effectiveness of the proposed ASP module on high-resolution feature maps. However, using ASP on the third and the fourth branches does not bring any significant improvements to the results. This is because the low-resolution feature maps are related to relatively large RFs, which are not greatly enlarged after adding the ASP modules (as discussed in Section 3). Therefore, we only apply the ASP modules to the first two branches in the proposed architecture. The SR module brings an increase of 0.77% in the OA and 1.55% in the average F1 score compared with the high-resolution branches (Branch 3 and Branch 4). Compared with the baseline HRNet, the proposed architecture with both ASP and SR modules has an advantage of 0.47% in OA and 0.59% in average F1 score.

Table 1. Results of the ablation study on the Potsdam dataset.

Method	Number of Branch	ASP	SR	Average F1	OA
HRNet (Branch 1)	1			88.87	87.41
HRNet (Branch 2)	2			91.71	90.59
HRNet (Branch 3)	3			90.10	90.14
HRNet (Branch 4)	4			88.33	89.61
HRNet (Branch 3, 4)	3, 4			90.30	90.33
HRNet	1, 2, 3, 4			92.18	91.01
HRNet+ASP (Branch 1)	1	√		91.87	90.56
HRNet+ASP (Branch 2)	2	√		92.38	91.21
HRNet+ASP (Branch 3)	3	√		90.02	90.11
HRNet+ASP (Branch 4)	4	√		88.75	89.65
HRNet+SR (Branch 3, 4)	3, 4		√	91.85	91.04
HRNet+ASP+SR	1, 2, 3, 4	√	√	**92.77**	**91.48**

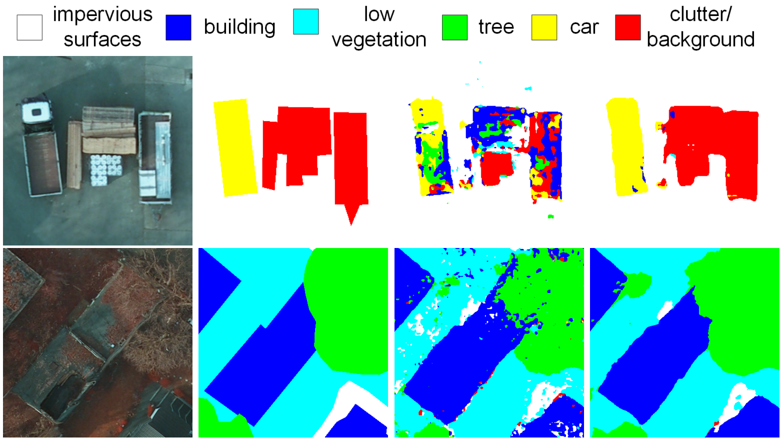
Table 2 shows the results of the ablation study on the Vaihingen dataset. The ASP modules increases the OA by 1.03 and 0.46% and the average F1 score by 1.02 and 0.40%, respectively, for the first and the second branches. The use of the SR module brings an increase of 0.64% in OA and 1.14% in average F1 score. The proposed architecture has an advantage of 0.67% in OA and 0.96% in the average F1 score compared to the baseline HRNet.

Table 2. Results of the ablation study on the Vaihingen dataset.

Method	Number of Branch	ASP	SR	Average F1	OA
HRNet (Branch 1)	1			86.46	87.86
HRNet (Branch 2)	2			87.22	88.89
HRNet (Branch 3)	3			84.86	88.30
HRNet (Branch 4)	4			81.38	87.78
HRNet (Branch 3, 4)	3, 4			84.96	88.49
HRNet	1, 2, 3, 4			87.84	89.40
HRNet+ASP (Branch 1)	1	√		87.48	88.89
HRNet+ASP (Branch 2)	2	√		87.62	89.35
HRNet+ASP (Branch 3)	3	√		85.03	88.44
HRNet+ASP (Branch 4)	4	√		81.53	87.90
HRNet+SR (Branch 3, 4)	3, 4		√	86.10	89.13
HRNet+ASP+SR	1, 2, 3, 4	√	√	**88.80**	**90.07**

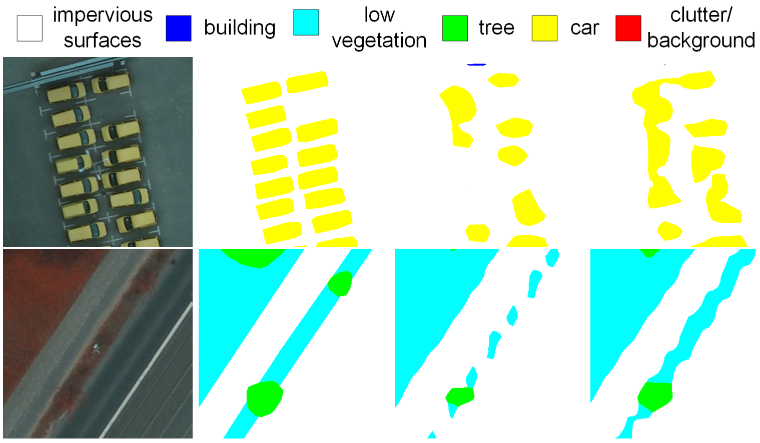
5.2. Visualization of Features

To qualitatively evaluate the effectiveness of both processing modules, we visualize the selected features by directly classifying them into the given categories. Figure 6 shows some examples of the effects of using the ASP module. They are selected from the outputs of the high-resolution branches on the Potsdam dataset. Since the original RFs of these branches are relatively small, the classified results are fragmented. After the use of the ASP module, context information is aggregated and the classified segments are more complete. Figure 7 shows the effects of using the SR module. The classified features are all produced by the low-resolution branches of HRNet. After using the SR module, the semantic representation of objects is enhanced, and some weak areas appear.



(**a**) Input image (IRRG); (**b**) Ground truth; (**c**) Before ASP; (**d**) After ASP.

Figure 6. Comparison of classified high-resolution features before and after the use of the adaptive spatial pooling (ASP) module (Potsdam dataset).



(**a**) Input image (IRRG); (**b**) Ground truth; (**c**) Before SR; (**d**) After SR.

Figure 7. Comparison of classified low-resolution features before and after the use of the spatial reasoning (SR) module (Potsdam dataset).

5.3. Comparisons with State-of-the-Art Methods

We further compare the proposed method with several semantic segmentation networks to evaluate its performance. The compared networks include the FCN [1], PSPNet [39], DeepLabv3+ [12], SENet [21], CBAM[40], and the most recent networks GloRe [41] and DANet [35]. The quantitative results on the Potsdam dataset are shown in Table 3. Among the compared methods, DANet and DeepLabv3+ achieve the best results in OA and average F1 score. Both of them have RF-enlarging designs that are able to aggregate context information. DANet contains a position attention module to aggregate long-range information, while DeepLabv3+ has a local information aggregation design at the top of its encoder. The PSPNet shows limited improvement compared with the basic FCN, the reason for which is discussed in Section 3. With the use of HRNet and the integrated modeling of both long-range and local context information, the proposed method shows a dominant advantage over all of the compared approaches and achieves the best results in both OA and per-class F1 score. It improves the OA by 1.74% and the average F1 score by 1.83% compared with DeepLabV3+.

Table 4 shows the quantitative results on the Vaihingen dataset. The differences in OA and F1 scores are smaller compared with those on the Potsdam dataset. This is partly due to the relatively smaller amount of training data. DANet, CBAM, and DeepLabV3+ show advantages compared with other methods described in the literature. The proposed approach outperforms the compared SOTA methods both in OA and in F1 score.

Table 3. Results on the Potsdam dataset in terms of per-class F1 score, average F1 score, and overall accuracy (OA).

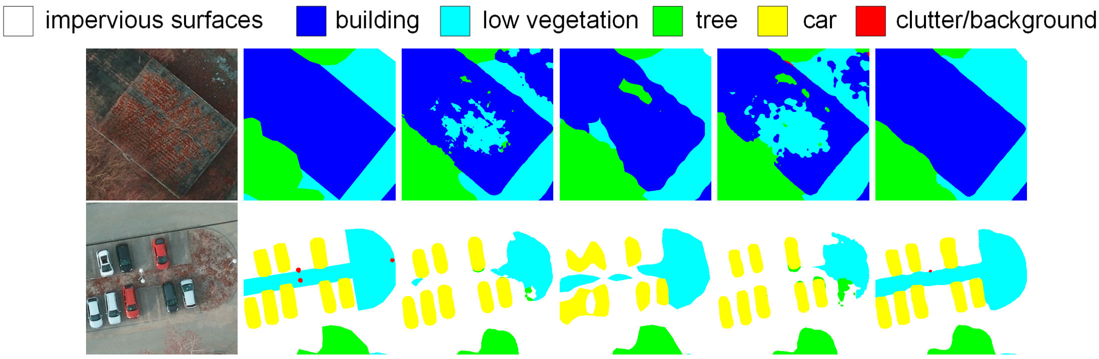
Method	**Per-class F1 Score (%)**					**Average F1 (%)**	**OA (%)**
	Impervious Surface	**Building**	**Low Vegetation**	**Tree**	**Car**		
FCN	91.46	96.63	85.99	86.94	82.28	88.66	89.42
PSPNet [39]	91.61	96.30	86.41	86.84	91.38	90.51	89.45
FCN+SE [21]	91.47	96.57	86.21	87.51	81.07	88.56	89.55
FCN+CBAM [40]	91.37	96.49	86.00	87.40	83.22	88.89	89.46
FCN+GloRe [41]	91.55	96.54	86.17	87.42	82.69	88.87	89.57
DANet [35]	91.61	96.44	86.11	88.04	83.54	89.14	89.72
DeepLabv3+ [12]	92.35	96.77	85.22	86.79	93.58	90.94	89.74
Proposed	**93.75**	**97.54**	**87.75**	**88.78**	**96.04**	**92.77**	**91.48**

Table 4. Results on the Vaihingen dataset in terms of per-class F1 score, average F1 score, and OA.

Method	Per-class F1 Score (%)					Average F1 (%)	OA (%)
	Impervious Surface	Building	Low Vegetation	Tree	Car		
FCN	94.10	90.98	81.25	87.58	76.80	86.14	88.66
PSPNet [39]	94.38	91.44	81.52	87.91	78.02	86.65	88.99
FCN+SE [21]	93.95	90.43	81.33	87.50	63.33	83.31	88.27
FCN+CBAM [40]	94.03	90.86	81.16	87.63	76.26	85.99	88.61
FCN+GloRe [41]	93.99	90.57	81.28	87.49	70.09	84.68	88.41
DANet [35]	94.11	90.78	81.40	87.42	75.85	85.91	88.59
DeepLabv3+ [12]	94.34	91.35	81.32	87.84	78.14	86.60	88.91
Proposed	**94.68**	**92.91**	**83.19**	**88.94**	**84.28**	**88.80**	**90.07**

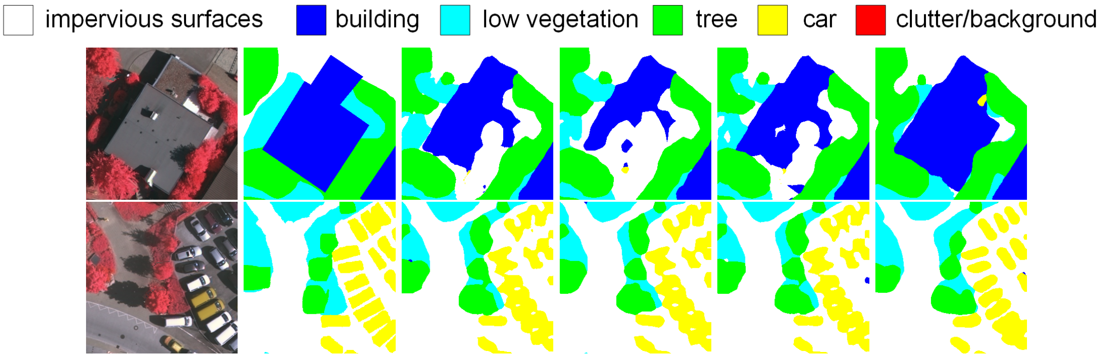
Figures 8 and 9 show some examples of the classification results on the Potsdam and the Vaihingen datasets, respectively. One can observe that the proposed method can better capture the contours of hard objects such as cars and buildings, and can better recognize confusion areas. Figures 10 and 11 show the large-scale prediction on a test image from the Potsdam dataset. The classification map produced by the proposed approach contains fewer errors while still preserving small objects.

These comparative experiments highlight two major advantages of the proposed approach: (1) It has a better capability of preserving the spatial details, which is due to the use of the parallel backbone network (HRNet); (2) its judgment of object categories is more reliable due to the multi-scale aggregation of the contextual information.



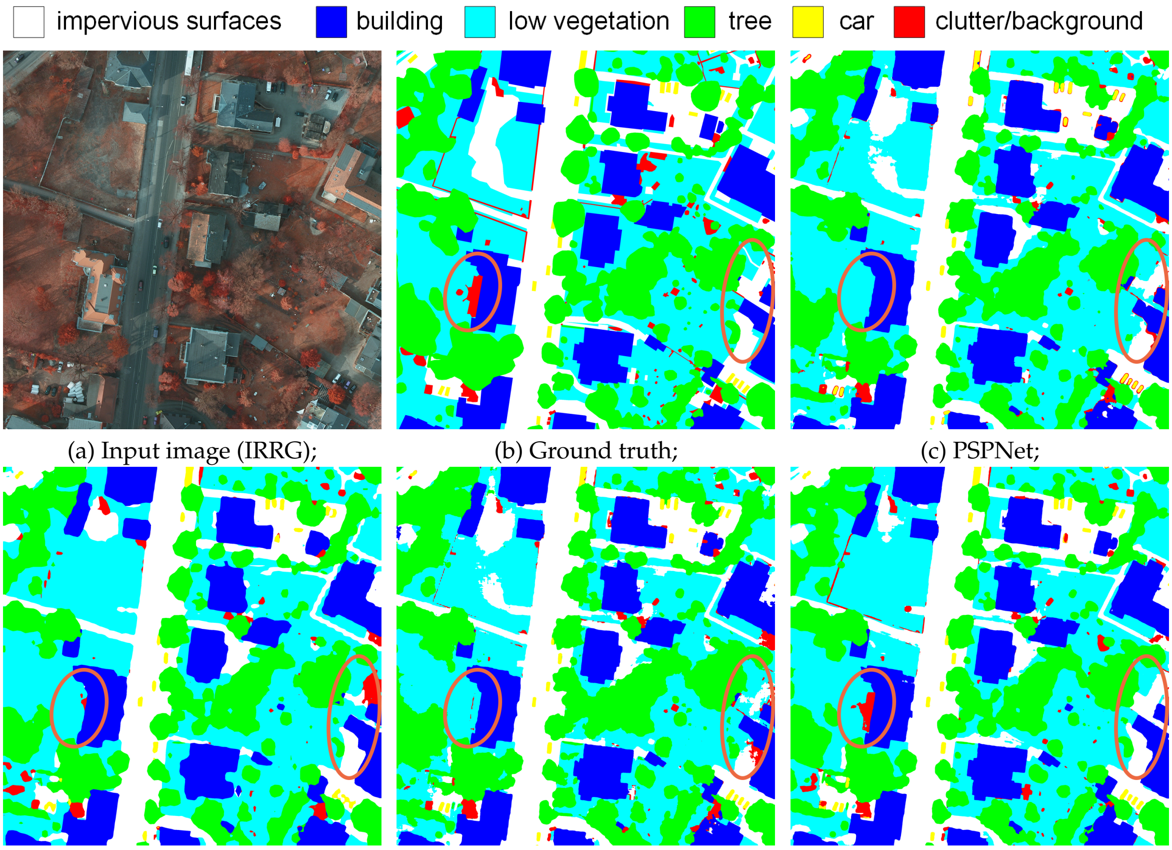
(**a**) Input image (IRRG); (**b**) Ground truth; (**c**) PSPNet; (**d**) DANet; (**e**) DeepLabv3+; (**f**) Proposed;

Figure 8. Examples of semantic segmentation results on the Potsdam dataset.



(**a**) Input image (IRRG); (**b**) Ground truth; (**c**) PSPNet; (**d**) DANet; (**e**) DeepLabv3+; (**f**) Proposed;

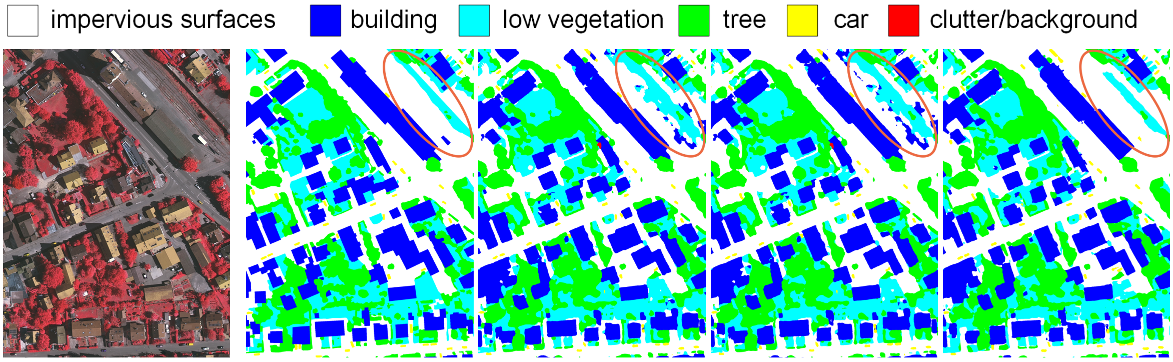
Figure 9. Examples of semantic segmentation results on the Vaihingen dataset. Major differences are highlighted (zoom in to see more details).



(a) Input image (IRRG); (b) Ground truth; (c) PSPNet;

(d) DANet; (e) DeepLabv3+; (f) Proposed.

Figure 10. Classification maps produced by the compared approaches on the Potsdam dataset.



(a) Input image (IRRG); (b) Ground truth; (c) PSPNet; (d) DeepLabv3+; (e) Proposed.

Figure 11. Classification maps produced by the compared approaches on the Vaihingen dataset. Major differences are highlighted (zoom in to see more details).

6. Conclusions

In this paper, the CNN-based semantic segmentation of RSIs is studied. Considering the limitation of conventional "encoder-decoder" architectures, we introduce the HRNet with parallel multi-scale branches to reduce the loss of spatial information. Moreover, we have designed separate context-aggregation modules for these parallel branches. On the one hand, the ASP module that adaptively calculates local descriptors is designed to enlarge the RFs of low-resolution branches. On the other hand, the SR module is introduced to learn long-range location correlations. Finally, an up-to-bottom feature fusion across the low-to-high branches is made to incorporate the aggregated multi-scale context information. The proposed architecture shows great improvements over existing approaches and achieves the best results on two RSI datasets (the Potsdam and Vaihingen datasets). One of the limitations of the proposed approach is that the HRNet reduces the spatial size of the input data in its early layers to avoid intensive calculations. In the future, we plan to further improve this approach by reducing the calculations and incorporating more multi-scale branches.

Author Contributions: J.Z. wrote the manuscript and performed the ablation study; S.L. supervised the study and revised the manuscript; L.D. designed the architecture and performed the comparative experiments; L.B. gave comments and suggestions to the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by a scholarship from China Scholarship Council, under grant NO.201703170123.

Conflicts of Interest: The authors declare that there is no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

VHR	Very High Resolution
RSI	Remote Sensing Image
HRNet	High-Resolution Network
DCNN	Deep Convolutional Neural Network
FCN	Fully Convolutional Network
RF	Receptive Field
DSM	Digital Surface Model
GSD	Ground Sampling Distance
ASP	Adaptive Spatial Pooling
SR	Spatial Reasoning
OA	Overall Accuracy
SOTA	State-of-the-Art

References

1. Jonathan, L.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
2. 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]
3. 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.
4. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
5. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing And Computer-Assisted Intervention*; Springer: New York, NY, USA, 2015; pp. 234–241.
6. Zhang, Z.; Zhang, X.; Peng, C.; Xue, X.; Sun, J. Exfuse: Enhancing feature fusion for semantic segmentation. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 269–284.
7. Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Xiao, B.; Liu, D.; Mu, Y.; Wang, X.; Liu, W.; Wang, J. High-Resolution Representations for Labeling Pixels and Regions. *arXiv* **2019**, arXiv:1904.04514.
8. Mou, L.; Hua, Y.; Zhu, X.X. A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 12416–12425.
9. Ding, L.; Zhang, J.; Bruzzone, L. Semantic Segmentation of Large-Size VHR Remote Sensing Images Using a Two-Stage Multiscale Training Architecture. *IEEE Trans. Geosci. Remote Sens.* **2020**. [CrossRef]

10. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations; San Diego, CA, USA, 7–9 May 2015.
11. Lin, G.; Milan, A.; Shen, C.; Reid, I.D. RefineNet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation. *Cvpr* **2017**, *1*, 5.
12. 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 (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
13. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. Denseaspp for semantic segmentation in street scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3684–3692.
14. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: New York, NY, USA, 2018; pp. 3–11.
15. Liu, W.; Rabinovich, A.; Berg, A.C. Parsenet: Looking wider to see better. *arXiv* **2015**, arXiv:1506.04579.
16. 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 (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
17. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. In Proceedings of the 4th International Conference on Learning Representations, ICLR, San Juan, Puerto Rico, 2–4 May 2016.
18. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062.
19. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
20. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef] [PubMed]
21. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition, CVPR, Salt Lake City, UT, USA, 18–23 June 2018.
22. 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.
23. Zhao, H.; Zhang, Y.; Liu, S.; Shi, J.; Change Loy, C.; Lin, D.; Jia, J. Psanet: Point-Wise Spatial Attention Network for Scene Parsing. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 267–283.
24. Paisitkriangkrai, S.; Sherrah, J.; Janney, P.; Hengel, V.D. Effective semantic pixel labelling with convolutional networks and conditional random fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 7–12 June 2015; pp. 36–43.
25. Zhang, C.; Sargent, I.; Pan, X.; Gardiner, A.; Hare, J.; Atkinson, P.M. VPRS-based regional decision fusion of CNN and MRF classifications for very fine resolution remotely sensed images. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 4507–4521. [CrossRef]
26. 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*, 1–10. [CrossRef]
27. 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]
28. Liu, S.; Ding, W.; Liu, C.; Liu, Y.; Wang, Y.; Li, H. ERN: Edge Loss Reinforced Semantic Segmentation Network for Remote Sensing Images. *Remote Sens.* **2018**, *10*, 1339. [CrossRef]
29. Kaiser, P.; Wegner, J.D.; Lucchi, A.; Jaggi, M.; Hofmann, T.; Schindler, K. Learning aerial image segmentation from online maps. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 6054–6068. [CrossRef]
30. Sun, W.; Wang, R. Fully Convolutional Networks for Semantic Segmentation of Very High Resolution Remotely Sensed Images Combined With DSM. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 474–478. [CrossRef]

31. 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; pp. 1–4.
32. Audebert, N.; Le Saux, B.; Lefèvre, S. Joint learning from earth observation and openstreetmap data to get faster better semantic maps. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017.
33. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5693–5703.
34. 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, 18–22 June 2018; pp. 7794–7803.
35. 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 Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3146–3154.
36. ISPRS. 2D Semantic Labeling Contest-Potsdam. Available online: http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html (accessed on 4 September 2018).
37. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. *Automatic Differentiation in PyTorch*; NIPS-W: New York, NY, USA, 2017.
38. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
39. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *TPAMI* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
40. Woo, S.; Park, J.; Lee, J.Y.; So Kweon, I. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3-19.
41. Chen, Y.; Rohrbach, M.; Yan, Z.; Shuicheng, Y.; Feng, J.; Kalantidis, Y. Graph-Based Global Reasoning Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 433-442.