remote sensing

MDPI

Article

Dual Attention Feature Fusion and Adaptive Context for Accurate Segmentation of Very High-Resolution Remote Sensing Images

Hao Shi [1,2,3], Jiahe Fan [1,2,3], Yupei Wang [1,2,3,*] and Liang Chen [1,2,3]

[1] Radar Research Lab, School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China; shihao@bit.edu.cn (H.S.); jiahe.fan@ieee.org (J.F.); chenl@bit.edu.cn (L.C.)
[2] Beijing Key Laboratory of Embedded Real-Time Information Processing Technology, Beijing 100081, China
[3] Chongqing Innovation Center, Beijing Institute of Technology, Chongqing 401120, China
* Correspondence: 7520190118@bit.edu.cn

Abstract: Land cover classification of high-resolution remote sensing images aims to obtain pixel-level land cover understanding, which is often modeled as semantic segmentation of remote sensing images. In recent years, convolutional network (CNN)-based land cover classification methods have achieved great advancement. However, previous methods fail to generate fine segmentation results, especially for the object boundary pixels. In order to obtain boundary-preserving predictions, we first propose to incorporate spatially adapting contextual cues. In this way, objects with similar appearance can be effectively distinguished with the extracted global contextual cues, which are very helpful to identify pixels near object boundaries. On this basis, low-level spatial details and high-level semantic cues are effectively fused with the help of our proposed dual attention mechanism. Concretely, when fusing multi-level features, we utilize the dual attention feature fusion module based on both spatial and channel attention mechanisms to relieve the influence of the large gap, and further improve the segmentation accuracy of pixels near object boundaries. Extensive experiments were carried out on the ISPRS 2D Semantic Labeling Vaihingen data and GaoFen-2 data to demonstrate the effectiveness of our proposed method. Our method achieves better performance compared with other state-of-the-art methods.

Keywords: deep learning; land cover classification; semantic segmentation

Citation: Shi, H.; Fan, J.; Wang, Y.; Chen, L. Dual Attention Feature Fusion and Adaptive Context for Accurate Segmentation of Very High-Resolution Remote Sensing Images. *Remote Sens.* **2021**, *13*, 3715. https://doi.org/10.3390/rs13183715

Academic Editor: Melanie Vanderhoof

Received: 28 June 2021
Accepted: 14 September 2021
Published: 17 September 2021

1. Introduction

With the development of very high resolution (VHR) remote sensing technology, large amounts of satellite remote sensing images with very high resolution are obtained every day [1]. Semantic segmentation is a computer vision task that predicts the semantic category for every pixel in an image, and such comprehensive image understanding is essential for many vision-based applications such as orbital remote sensing, autonomous driving [2,3], medical image analysis, and so on [4–6]. However, there are still lots of challenges for the task of semantic segmentation in VHR remote sensing images with complex scenes, such as poor accuracy of multi-category semantic segmentation, poor speed of multi-category semantic segmentation, and so on.

Traditional machine learning-based methods [7–9] rely on human experience and complex feature engineering. The segmentation performance mainly depends on whether researchers can obtain the accurate features of their targets. Since feature extraction is done manually by the researcher, these human-designed features may fail to handle various complex applications. With the development of deep learning [10–14], there are lots of CNN-based methods [15–21] applied in the semantic segmentation of the VHR remote sensing images. Previous methods [22–24] have used ConvNets for semantic segmentation, in which each pixel is labeled with the class of its enclosing object or region. The fully

convolutional networks (FCNs) [25] replaces the fully connected layers in traditional classification network with convolutional layers to get a segmentation result. Compared with previous methods [22–24], FCN-based methods [26–29] have made great progress in semantic segmentation, such as: accepting input images of any size, obtaining features automatically, and so on.

A general semantic segmentation architecture can be broadly thought of as an encoder network followed by a decoder network [30]: The encoder is usually a pre-trained classification network like VGG/ResNet [11,13] followed by a decoder network. The encoder usually consists of a series of convolution or pooling operations. The task of the decoder is to semantically project the discriminative features (lower resolution) learned by the encoder onto the pixel space (higher resolution) to get a dense classification. VHR remote sensing images generally consist of large and complex scenes with lots of scale-varying objects. The large variability between some targets in the same category, and small difference between some targets in different categories, brings a significant challenge to obtain precise segmentation results, especially at the boundary regions [27]. Generally, long-range contextual cues mean the relationship between pixels in long-range distance. For instance, boats usually appear in the sea or rivers but not in indoor environments. With the long-range context of a water scene, water-related feature channels should be weighted higher to increase the probability of predicting boat pixels.

In order to overcome these challenges, Chen et al. [20] and Zhao et al. [17] introduced long-range contextual cues into the top feature maps to better distinguish targets with different scales and similar features. Hu et al. [31] aim to encode the global context to generate a channel-wise feature weight that is used to re-weigh the feature channels for improving segmentation accuracy. For instance, the surface of the road may have the same color as the surface of the roof. However, the vehicles usually appear on the road but not on the roof. Chen et al. proposed Deeplabv3 [21] to achieve better segmentation performance. It uses an atrous spatial pyramid pooling (ASPP) module to aggregate the context in different distances, which is achieved by a series of dilated convolutions with different scales. These methods make progress in improving the accuracy of the segmentation in a natural sense. However, these methods are suboptimal for remote sensing images since the scenes in remote sensing applications are more complex than in natural scene applications, and the background of the target may have numerous kinds of interference.

Moreover, CNN encodes an input image by a series of convolution operations and learns suitable feature expressions for image recognition from an input image [32]. Each convolutional layer in CNN utilizes a convolution kernel to process the input image or the output of the previous layer. The feature maps at the deep layer in CNNs encode rich high-level semantic information generated by multiple stages of spatial pooling and convolution. However, it exhibits clear limitations that makes it hard to obtain the accurate segmentation result by simply up-sampling the feature maps, losing much of the low-level fine image structural details; it also further reduces the classification accuracy of pixels at the boundary [27]. Previous methods [17,20] utilized feature maps in the deep layer to obtain the final segmentation results, which fail to generate good predictions for pixels near the object boundary. Thus, some methods [16,19,33] make use of the lower feature maps to improve the semantic segmentation accuracy near the boundary region. The feature map generated by the lower layer of CNN has poor semantic information but high resolution. U-Net [16] adopts an encoder-decoder architecture to improve segmentation results. It adds skip connections between the encoder and decoder, which can combine detailed information and category information of different scale features. SegNet [15] also records the pooling indices in the encoder and reuses them in the decoder to improve the resolution of the result. However, the feature maps in neighboring layers have different channels associated with different targets and each channel has different semantic information. Simply combining high-level feature maps and high-resolution feature maps will drown useful information in massive amounts of useless information and cannot reach an informative

high-level and high-resolution feature map [26]. There is a semantic gap between the feature maps in the different layers, which may have a negative influence on the result.

To address the above problems and improve the segmentation accuracy of targets, we propose a novel framework for remote sensing image segmentation, which is illustrated in Figure 1. Inspired by [34], we adopt an adaptive context aggregating module to re-weigh pixels in different channels with the weight vectors generated by the global context information. We introduce matrix multiplication to generate the spatially-varying feature weight factors, which are utilized as the parameters of a series of dilated depth-wise convolutions with different dilation factors to capture information in multiple scales. In this way, we integrate the contextual cues into the feature maps with predictable and input-variant convolutions and the module re-weighs features at different spatial locations automatically for fine semantic segmentation results. Furthermore, we introduce a channel attention module to enhance the consistency of the feature maps.

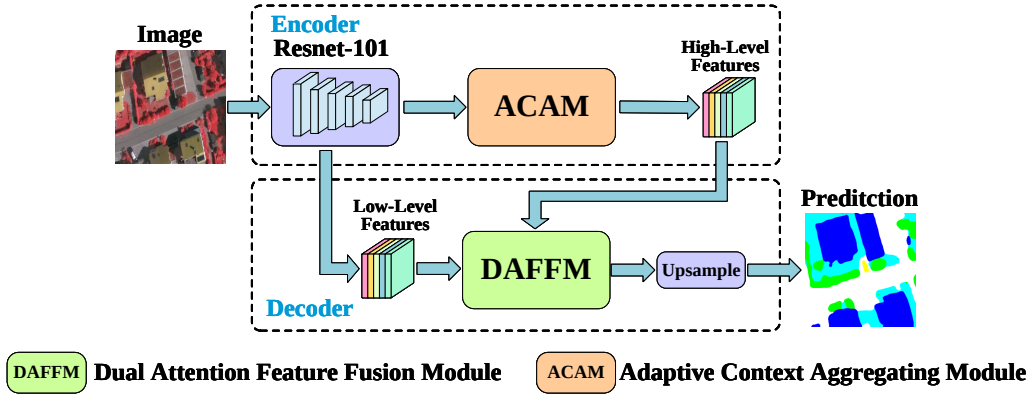

Figure 1. Overview of the proposed network for semantic segmentation of remote sensing images.

Moreover, in order to make better use of the multi-scale feature maps in CNNs, we adopt a dual attention feature fusion module to fuse the feature maps in CNNs into the different layers based on both channel and spatial attention mechanisms. Generally speaking, our goal is to import extra low-level details in the object boundary, where it is difficult to obtain accurate category labels for the pixels. As the feature maps in two different levels of CNN layers may have the semantic gap, we utilize a matrix multiplication mechanism to measure the relevance of two feature maps at both the channel and spatial dimension, which is the basis of the weight vectors. As for the channel dimension, we capture the channel dependencies between any two channels in the different layers and update the lower channel features with the weight vector. For spatial attention, any two positions in the different layers with similar features can contribute mutual improvement regardless of their distance in spatial dimension. Finally, the outputs of these two attention modules consist of the final output. This feature fusion module aims at reducing the semantic gap for fine segmentation results. Overall, the contributions of this paper are summarized as follows:

- In order to utilize global contextual cues, we integrate the contextual cues with the spatially-varying feature weighing factors.
- To improve the classification accuracy of the pixel near the object boundary, we propose a multi-scale feature fusion module based on the attention mechanism on both the spatial and channel dimensions.
- To validate the effectiveness of our method, we conduct extensive experiments based on the ISPRS 2D Semantic Labeling Vaihingen data [35] and GaoFen-2 data [36]. The results show that our method has led to significant improvements and demonstrate the effectiveness and robustness of our method.

This article is organized as follows: Section 2 is about related work. In Section 3, we introduce our semantic segmentation framework in detail. Section 4 presents the results of

experiments. In Section 5, we analyze the results of the experiments and discuss future work. Finally, we summarize the paper in Section 6.

2. Related Work

In this section, we review works related to semantic segmentation on three different aspects: FCN-based semantic segmentation, contextual cue extraction, and multi-scale feature fusion.

2.1. Semantic Segmentation

Fully Convolutional Network (FCN) [25] based methods have made significant progress in semantic segmentation, which first replaces the fully connected layer in the traditional classification network by convolutional layer to get a segmentation result and it achieves end-to-end training by adopting the output feature map to match the resolution of the input image with up-sampling. However, the spatial resolution will be reduced due to the use of the downsampling and pooling operation. To improve the resolution of the output, researchers have adopted a variety of methods. Badrinarayanan et al. [15] utilize the convolution and deconvolution layers to construct a symmetric auto-encoder architecture, which maintains high-frequency details in the input image. Ronneberger et al. [16] adopt an encoder-decoder architecture to improve segmentation results. In order to preserve more detailed information, Chen et al. [18,20] adopt atrous spatial pyramid pooling (ASPP) to expand the receptive fields and embed contextual cues, which consists of parallel dilated convolutions with different dilated rates. However, some drawbacks, like the grid effect, bring new challenges for improving the accuracy of the segmentation results. Chen et al. [21] employ a new joint upsampling module to solve the above issue generated with dilated convolution. Moreover, Zhao et al. [17] propose a pyramid pooling module to collect the effective contextual prior, containing information of different levels. ASPP module [18] has been utilized in many methods [37,38] to capture multi-scale contextual cues from the final convolutional feature map. Lin et al. [39] and Ding et al. [40] obtain context in different scales by fusing different feature maps.

Meanwhile, in remote sensing, researchers [41,42] are also inspired by the development of segmentation in natural scenes. Wang et al. [43] present a gated convolutional neural network to automatically select adaptive features when merging different-layer feature maps. Panboonyuen et al. [27] introduce the global convolutional network to capture different resolutions by extracting multi-scale features for better results on remotely sensed images. Li et al. [44] present an auto encoder-based architecture of deep learning that makes extensive use of residual learning and multiscaling for better semantic segmentation of remote sensing images. Kang et al. [45] design the dense spatial pyramid pooling to extract dense and multi-scale features simultaneously and use the focal loss to suppress the impact of the error labels in ground truth.

2.2. Contextual Cues Modeling

Although FCN [25] based methods have made great progress in semantic segmentation, some new problems have emerged with the development of the research. A series of convolution and down-sampling operations capture information with larger receptive fields. However, they still cannot take advantage of the global or long-range contextual cues effectively. Liu et al. [46] encode the global pooling feature, which is concatenated with the original feature maps to integrate the global context. PSPNet [17] adopt a spatial pyramid pooling module consist of a series of pooling operations to collect contextual cues in different scales. Deeplab series [18,20] develop ASPP to obtain multi-scale contextual cues by dilated convolutional layers with different dilation rates. Yang et al. [47] and Bilinski et al. [48] encode contextual cues in a dense way. Huang et al. [28] propose a network structure whose multiple branches with different atrous rates can share a single kernel effectively. In order to increase the receptive field size, Peng et al. [33] directly utilize a large filter to capture the contextual cues. Although, the above papers utilize

various methods to capture the contextual cues, they treat all pixels in each sub-region with uniform weight for feature aggregation, which cannot capture the information in each channel with different weight vectors.

To solve this issue, some researchers try to aggregate the feature in an adaptive and flexible way. He et al. [29] adopt an adaptive context module to estimate inter-pixel affinity weights for feature aggregation. Zhang et al. [49] propose an aggregated co-occurring feature (ACF) module to aggregate the co-occurrent context. Based on the ACF module, Zhang et al. [50] propose the attentional class feature module to make different pixels adaptively focus on different class centers to improve the semantic segmentation. Zhao et al. [51] predict that the attention map will aggregate contextual cues for each pixel. Fu et al. [52] propose the dual attention module consists of a position attention module and channel attention module with the self-attention mechanism to aggregate features. These methods show robustness to aggregate the contextual cues and encode the long-range context, which is able to boost the segmentation performance.

2.3. Multi-Scale Feature Fusion

FCN-based methods utilize a series of convolution and pooling operations to obtain semantic information of the target. However, it generates a new problem that successive convolution and pooling operations lead to the reduction of the feature resolution and the detail information, which influence the accuracy of the result. In order to solve this issue, it is essential to make use of both high-level categorical semantic, and low-level spatial details.

Unet [16] adopt an encoder-decoder architecture with skip connections to combine categorial semantic and spatial details in different scale feature maps. Lin et al. [53] adopt the same architecture as Unet [16] with predictions from each level of the feature pyramid. Lin et al. [19] propose a multi-path refinement network to exploit features at multiple levels of abstraction for high-resolution semantic segmentation. Panboonyuen et al. [27] extract multi-scale features from different stages of the network and fuse these features for better results. Wang et al. [43] adopt a gate mechanism to integrate the feature maps more effectively. In order to take advantage of the redundancy in the label space of semantic segmentation, while Tian et al. [54] propose a data-dependent upsampling to replace the bilinear one. He et al. [55] propose a dynamic multi-scale network to adaptively capture and fuse multi-scale contents for predicting pixel-level semantic labels. Li et al. [26] propose a new architecture to selectively fuse features from multiple levels using gates in a fully connected way. Yu et al. [56] adopt an encoder-decoder architecture containing new modules to select the more discriminative features and make the bilateral features of the boundary distinguishable with deep semantic boundary supervision.

3. Methods

In this section, we first present a general framework of our network and then introduce the adaptive context aggregating module (ACAM) and the dual attention feature fusion module (DAFFM). Finally, we describe how to aggregate them together for further refinement.

3.1. Overview

Given the context of remote sensing, objects are diverse on scales, lighting, and views. High-resolution remote sensing of images involves complex scenes where objects in the same category can be diverse in appearance and features. Meanwhile, different semantic categories may have similar features. Since a series of convolution operations can lead to a local receptive field, the features corresponding to the pixels with the same label may differ significantly, which brings additional difficulties for accurate classification at the pixel-level.

Feature re-weighing has proven to be an efficient approach to capture semantic contexts at different distances according to the channel-wise weight factors from the global contextual cues. However, there is the limitation that the weight vector is shared by all spa-

tial locations of the 2D feature map. Actually, feature maps of each channel have different contextual cues. We cannot make full use of these differences by a single weight vector [34]. Thus, it is not a suitable choice to use a globally-shared weight to re-weigh different spatial locations belonging to objects of different categories. Therefore, it is necessary to aggregate the contextual cues for better segmentation according to two principles: (1) learning the feature weight factors from the global context and (2) capturing different locations' unique characteristics in a spatially-varying way. To achieve this goal, we propose the ACAM to obtain the channel-wise vectors to re-weigh the 2D feature maps within the global context.

The top layers in CNNs encode rich global category semantics. However, local spatial details are missing [57–60]. On the contrary, the lower-level feature maps capture rich spatial details. However, the lower-level feature layers fail to encode global semantic cues due to limited discriminative ability [19]. Therefore, it is essential to fuse the global semantic cues and local spatial details. The authors of [16] address this issue by concatenating different levels' feature maps, whose improvement is limited by the large semantic gap. Inspired by some successfully applied attention mechanisms [61–63], we introduce a feature fusion module based on both position and channel attention to effectively combine the feature maps at different scales.

As illustrated in Figure 1, we introduce the ACAM to capture multi-scale contextual cues. On this basis, the DAFFM is utilized to fuse the multi-scale feature maps with both spatial and channel attention mechanisms. We adopt a pre-trained residual network [13] with the dilated strategy as the backbone. We replace the down-sampling operations with dilated convolutions in the last two layers. Thus, the size of the final feature map is $1/8$ of the input images. It can retain more details without adding extra parameters. The backbone encodes each remote scene image into a feature map of $X \in \mathbb{R}^{c \times h \times w}$, where h, w, c are the height, width, and feature channels of the feature map. First, we adopt the ACAM to capture contextual cues in the top feature map X, then re-weigh the top feature map by the spatially-varying weights generated by the global context. Then, we feed the feature maps generated by the ACAM and the feature maps from the fourth layer of the backbone CNN into the DAFFM to achieve a balanced fusion of features at different levels. Finally, we obtain the segmentation result by concatenating the feature maps generated by the DAFFM.

3.2. Adaptive Context Aggregating Module

The adaptive context aggregating module (ACAM) consists of three submodules: channel re-weighing module, convolution kernels predicting module, and context-adaptive capturing module.

First, we utilize the channel re-weighing module based on channel attention to enhance the consistency of the feature maps in the top layer. We will illustrate the details of the re-weighing module in Section 3.4.

In order to control the computation cost and maintain the spatial information, we utilize matrix multiplication to predict the convolution kernels. The input feature map is first transformed into the query feature map $Q \in \mathbb{R}^{c \times h \times w}$ and the key feature map $K \in \mathbb{R}^{c \times h \times w}$, respectively, which is implemented by 1×1 convolutions to reduce the computation.

To aggregate the global spatial information, we first reshape the feature map K and Q into $K_1 \in \mathbb{R}^{c \times n}$ and $Q_1 \in \mathbb{R}^{c \times n}$ where $n = h \times w$. Then, we transpose feature map Q_1 for performing a matrix multiplication with feature map K_1 to generate the feature map W as illustrated in Equation (1):

$$W = Q^{\top} K, \tag{1}$$

where $W \in \mathbb{R}^{C \times C}$. The element in the feature maps W, represents the overall spatial distribution of each channel in the feature map K. The result measures the similarity of the spatial distribution between the feature maps Q and K. In this way, the global spatial information of each channel is concentrated into the feature map W. Then, we expand the dimension of W and obtain the feature map $W_1 \in \mathbb{R}^{1 \times C \times C}$. After that, we reduce the channels of W_1 to s by 1×1 convolutions and compress its dimension to obtain the feature

map $W_2 \in R^{C\times 9}$. Then, we obtain the predicted feature map from W_2 as the convolution kernel by reshaping the batch normalization operation.

Then, in order to re-weigh each pixel of the input feature maps, we adopt depth-wise convolution with the predicted convolution kernels to generate the spatially-varying weight map. Thus, each channel of the predicted kernel is utilized to re-weigh one channel of the input feature map. In this way, the weight vector is independent with each other and the contextual cues can be aggregated according to its spatial information. Moreover, we denote the original kernels $S \in \mathbb{R}^{c\times 3\times 3}$ with dilation rate 1 as S_1 and obtain S_2, S_3 with different dilation rates 2 and 3 to expand the receptive field without introducing extra parameters and computations.

As shown in Figure 2, we perform depth-wise convolution on the input feature map with convolution kernels S_1, S_2 and S_3, independently and use the sigmoid function to generate the weight feature maps $R_1 \in \mathbb{R}^{c\times h\times w}$, $R_2 \in \mathbb{R}^{c\times h\times w}$ and $R_3 \in \mathbb{R}^{c\times h\times w}$, which is added to generate the final weight feature map $R \in \mathbb{R}^{c\times h\times w}$,

$$R = R_1 \oplus R_2 \oplus R_3, \tag{2}$$

where $\oplus$ represents the element-wise addition.

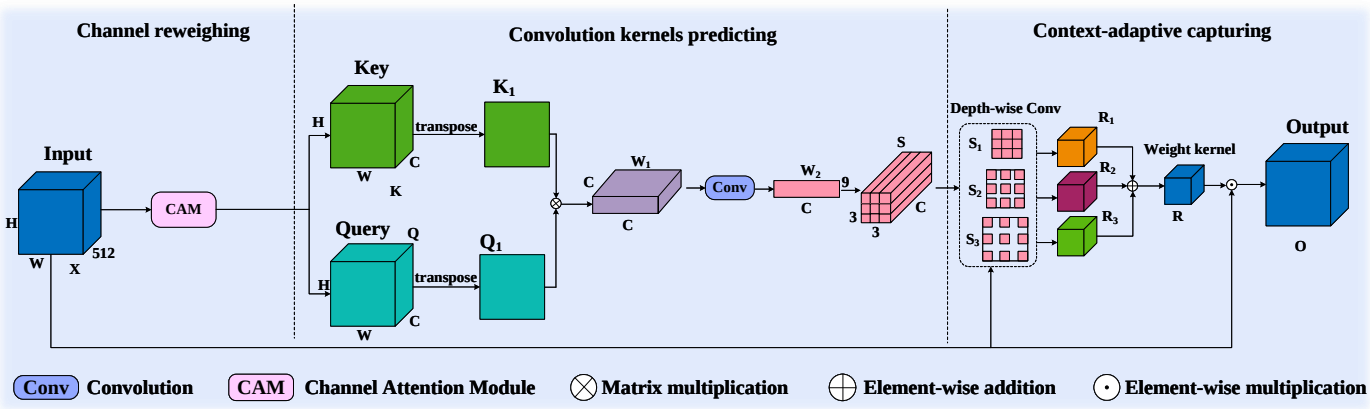

Figure 2. Overview of the proposed ACAM. To make better use of the contextual cues, we introduce ACAM to aggregate the contextual cues adaptively. Different from current context aggregating methods, the ACAM generates the weight vectors according to the global context cues of all feature channels. Moreover, the dilated depth-wise convolutions with different dilation factors are introduced to capture contextual cues. In this way, we integrate the contextual cues adaptively for better performance.

Finally, we obtain the output feature map $O \in \mathbb{R}^{c\times h\times w}$ by performing the element-wise multiplication between R and the input feature map X as shown in Equation (3),

$$O = X \odot R, \tag{3}$$

where $\odot$ donates the element-wise multiplication. Since the scenes in remote sensing applications are more complicated, and the categories of targets are richer compared with the natural scene, the contextual cues in the background of the target varies. Thus, we utilize more channels in the feature map Q to explore more complex relationships between the different channels.

3.3. Dual Attention Feature Fusion Module

The feature maps in deeper layers of CNNs encode richer semantic cues but with smaller spatial resolution. On the contrary, the spatial resolution of the lower feature maps is larger, but local spatial details are lack. Although the existing multi-scale feature fusion mechanism is a reasonable solution, the improvement is limited by the large semantic gap among the multi-level features.

To effectively combine the multi-scale feature maps, we propose the dual attention feature fusion module (DAFFM), as illustrated in Figure 3, which is based on both spatial and channel attention mechanisms. Given the deeper-level feature map A and the lower-level feature map B, where A is generated by the ACAM, we first re-weigh the lower feature map by the module illustrated in Figure 4 and utilize the 1×1 convolution to compress the channels to generate the feature map $B_1 \in \mathbb{R}^{c \times h \times w}$, where c, h and w represent the height, width, and channels of the feature map.

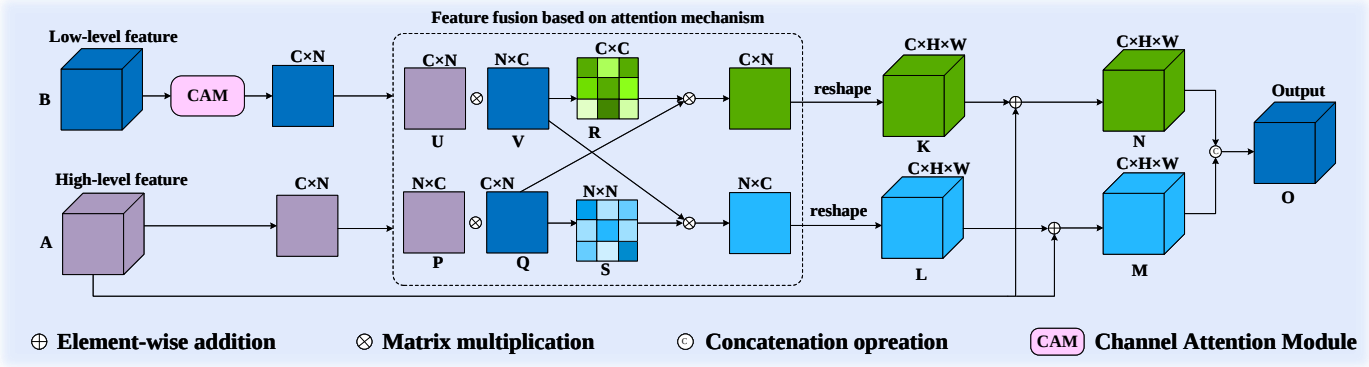

Figure 3. Overview of the proposed DAFFM. In order to make better use of multi-scale feature maps, we adopt the feature fusion module DAFFM based on spatial attention and channel attention mechanism to fuse the detail and semantic information for better segmentation results at the boundary. Benefiting from the detail and semantic information aggregating based on attention mechanisms, the semantic gap between different layers is reduced. In this way, we utilize the multi-scale feature maps more effectively.

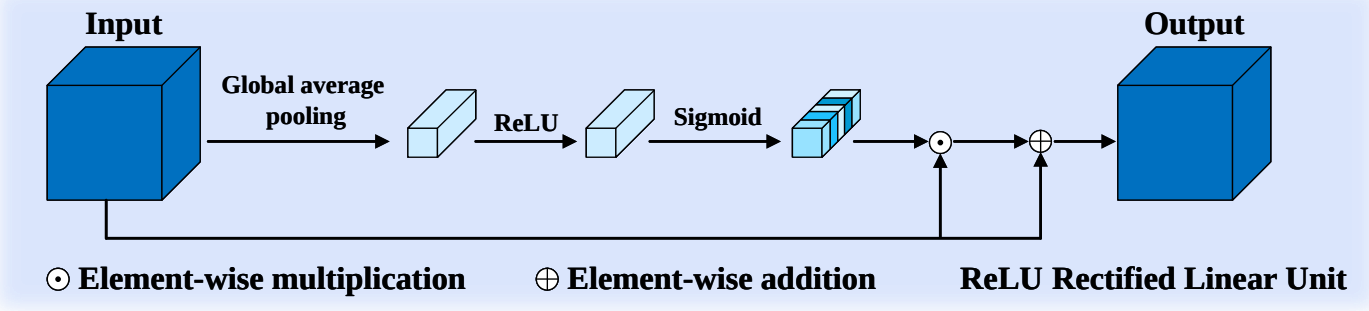

Figure 4. Overview of the implied channel attention module. The CAM utilizes a global average pooling layer and sigmoid function to generate the weight vectors. In this way, we re-weigh feature maps to improve the consistency of different channels for better segmentation results.

Firstly, we fuse the feature map based on the attention mechanisms on spatial dimensions. We reshape A and B_1 to $P \in \mathbb{R}^{c \times n}$ and $Q \in \mathbb{R}^{c \times n}$, where $n = h \times w$ represents the number of pixels in each feature map. After that, we apply the matrix multiplication between P and Q and utilize a softmax layer to calculate the spatial attention map $S \in \mathbb{R}^{n \times n}$,

$$s_{ji} = \frac{\exp(P_i \cdot Q_j)}{\sum_{i=1}^{N} \exp(P_i \cdot Q_j)}, \tag{4}$$

where s_{ji} measures the relevance of pixels between i-th position in the lower feature map and j-th position in the higher feature map.

It can be inferred from Equation (4) that each position of the final feature $O \in \mathbb{R}^{c \times h \times w}$ is a weighted sum of the features across all positions of the deeper-level features. As the final feature map is generated by the deeper-level feature map, the high-level semantic is well preserved in the outputs.

Then we transpose the feature map Q to $V \in \mathbb{R}^{c\times n}$ for performing a matrix multiplication with the spatial attention map S and reshape the output to generate the feature map $L \in \mathbb{R}^{c\times h\times w}$. Finally, we utilize an element-wise sum operation between A and L to obtain the final output $M \in \mathbb{R}^{c\times h\times w}$ as follows:

$$M_j = \alpha \sum_{i=1}^{N} (s_{ji} V_i) + A_j, \tag{5}$$

where α is initialized as 0 and gradually learns to assign a reasonable weight factor and V_i represents the pixels of the *i-th* position in the lower feature map and A_j represents the *j-th* channel of the deeper-level feature map.

On the other hand, we reshape the B into $U \in \mathbb{R}^{c\times n}$ and then perform a matrix multiplication between U and V. Then, we utilize a softmax layer to obtain the channel-wise attention map $R \in \mathbb{R}^{c\times c}$, where c is the number of the channels:

$$r_{ji} = \frac{\exp(U_i \cdot V_j)}{\sum_{i=1}^{C} \exp U_i \cdot V_j}, \tag{6}$$

where r_{ji} measures the *i-th* channel's impact on the *j-th* channel. Then we perform a matrix multiplication between the transpose of R and Q to generate the feature map $K \in \mathbb{R}^{c\times h\times w}$. Then we multiply the results by a scale parameter β and perform an element-wise sum operation with A to obtain the final output $N \in \mathbb{R}^{c\times h\times w}$,

$$\mathrm{N}_j = \beta \sum_{i=1}^{C} (r_{ji} Q_i) + A_j, \tag{7}$$

where β gradually learns a weight from 0. A_j represents the *j-th* channel of the A and Q_i represents the *i-th* channel of the Q. We can learn from Equation (7) that each channel of the final output is a weighted sum of the features Q. It models the relevance of different channels in feature maps and helps to boost feature fusion. Finally, we obtain the final fusion result $O \in \mathbb{R}^{c\times h\times w}$ by concatenating M and N.

In summary, we utilize matrix multiplication to measure the relevance between feature maps with different scales in both spatial and channel dimensions, which is utilized as the guidance of the fusion operation. In this way, we improve the semantic segmentation accuracy in the boundary and reduce the negative influence of the semantic gap for feature fusion.

3.4. Channel-Wise Feature Re-Weighing

Generally, each channel of the feature map can be regarded as a class-specific response. In order to enhance the consistency of the feature maps in each layer, we utilize the channel attention module (CAM) to change the weights of the features in each channel, as illustrated in Figure 4. We first employ a global average pooling layer to squeeze the spatial information and then utilize the sigmoid function to generate the weight vectors, which are finally combined with the input feature maps by an element-wise multiplication operation to generate the output feature map. The overall information is integrated into the weight vectors and strengths the feature maps, which are more relevant to the ground-truth.

3.5. Implementation Details

Our implementation is based on Pytorch [64]. For better training results, we adopt the pre-trained ResNet model [13] to initialize the backbone CNN and initialize the other layer with normal distribution. Moreover, we replace the down-sampling operations with dilated convolutions in the last two layers, and the hyperparameters of training epochs, batch-size, initial learning rate are set to 200, 6 and 0.005.

Many public datasets have been published to advance semantic segmentation in remote sensing. We select two widely used datasets to evaluate our proposed method: ISPRS 2D Semantic Labeling Vaihingen data [35] and GaoFen-2 dataset [36].

The ISPRS 2D Semantic Labeling Vaihingen data [35] is provided by the International Society for Photogrammetry and Remote Sensing, which consists of 33 high-resolution true orthophoto tiles and corresponding digital surface models, as well as ground-truth labels, and we adopt the DSM band channel to perform our experiments. The labels are classified into 6 categories: impervious surfaces, building, low vegetation, tree, car and clutter. We select 11 tiles for the training dataset and 5 tiles are used as the validation set. The rest of the tiles are used for testing the performance of the method. Based on the 33 tiles, we perform experiments on the whole 6 categories.

Comparing with ISPRS 2D Semantic Labeling Vaihingen data [35], GaoFen-2 dataset [36] is more challenging, which contains 500 satellite images collected from GaoFen-2 satellite over different geographic locations in China. It contains 500 labeled images of size 512×512, which have a large intra-class difference and small inter-class diversities. Thus, the GaoFen-2 [36] is more convincing to test the effectiveness and robustness of our method. The GaoFen-2 [36] is split into 400 training and 100 validation images with annotation containing 9 categories: road, building, tree, grass, bare land, water, transportation, impervious surfaces and others.

In order to evaluate the performance of our method, we utilize overall pixel accuracy (OA) and mean intersection over union (mIOU) as the metrics. The OA represents the accuracy of all pixels for all categories. The mIOU means the intersection of prediction and target divided by the union, which is the main criterion for evaluating the performance of each method [17,52].

4. Experimental Results

In this section, to evaluate the proposed method, we carry out comprehensive experiments and evaluate the performance of our method qualitatively and quantitatively.

4.1. Ablation Study for Each Module

We employ the ACAM and DAFFM for better segmentation results. To verify the performance of the two modules and help us to understand them better, we benchmark the whole modules based on the above two datasets. The experiment results are shown in Tables 1 and 2.

Table 1. Evaluation of land cover classification accuracy on ISPRS 2D Semantic Labeling Vaihingen data [35].

Methods	OA (%)	mIOU (%)
Baseline	86.66	68.02
Baseline + ACAM	86.44	68.75
Baseline + DAFFM	**87.17**	69.86
Baseline + DAFFM + ACAM	87.01	**70.51**

Table 2. Evaluation of land cover classification accuracy on GaoFen-2 dataset [36].

Methods	OA (%)	mIOU (%)
Baseline	79.36	53.94
Baseline + ACAM	78.50	52.18
Baseline + DAFFM	79.85	55.08
Baseline + DAFFM + ACAM	**80.91**	**56.98**

4.1.1. Baseline

The baseline has similar architectures like U-net, which are used to evaluate the effectiveness of other components, and it only fuses the feature maps in the last two layers by concatenating the feature maps directly. Baselines achieve mIOU of 68.02% on ISPRS 2D Semantic Labeling Vaihingen data [35] and 53.94% on GaoFen-2 dataset [36].

4.1.2. Adaptive Context Aggregating Module

Compared with the baseline, we introduce the ACAM to capture the contextual cues in the last layer. As shown in Tables 1 and 2, ACAM obtains mIOU of 68.75% on ISPRS 2D Semantic Labeling Vaihingen data [35] and 52.18% on GaoFen-2 dataset [36]. It can be inferred that ACAM improves the performance of the segmentation by capturing context with adaptive weight vectors.

4.1.3. Dual Attention Feature Fusion Module

A dual attention feature fusion module is introduced between the last two layers based on the baseline for better fusion of muti-scale features. DAFFM obtains an mIOU of 69.86% on ISPRS 2D Semantic Labeling Vaihingen data [35] and 55.08% on GaoFen-2 [36], outperforming the baseline by 1.84% and 1.14%, respectively. It is obvious that DAFFM achieves better segmentation results with feature fusion based on the attention mechanism. According to Tables 1 and 2, our feature fusion module is effective in aggregating multi-scale features for better segmentation results.

4.1.4. Network with Full Architecture

We integrate the ACAM and DAFFM into the baseline to generate a network which has full architecture of our method. Compared with the models mentioned above, our method utilize the ACAM to capture multi-distance context adaptively and combine the lower features by DAFFM for more accurate segmentation result. The two modules adopted in our method improve the performance in dimensions of context and multi-scale features, which improve the classification accuracy of pixels both in the internal and boundary regions. The experimental results demonstrate the effectiveness of our method which obtains the best performance with mIOU of 70.51% on ISPRS 2D Semantic Labeling Vaihingen data [35] and 56.98% on GaoFen-2 [36] and boosts over 2.49% and 3.04%, respectively, compared with the baseline.

In order to show discrimination of each module directly, we visualize the comparison results based on ISPRS 2D Semantic Labeling Vaihingen data [35] as illustrated in Figure 5. It is shown that DAFFM obtains the better segmentation performance compared to the baseline, which is consistent with our analysis mentioned above. On the other hand, ACAM also achieves better performance on some small targets visually, compared to the baseline, which contributes to aggregating the multi-scale contextual cues. By combining the advantages of DAFFM and ACAM, our method obtains the best performance according to the visual segmentation result.

According to the analysis in different aspects, the ACAM and DAFFM adopted in our method improve the segmentation accuracy, respectively. Moreover, it is consistent and feasible to combine both of them into the same architecture, which obtains more improvement, and the DAFFM brings the most improvements in our method. We adopt two modules to obtain better results in different dimensions, which is demonstrated in effectiveness based on two datasets [65,66]. On the other hand, the quantitative and qualitative experiment results further demonstrate the importance of the multi-scale contextual cues and features in different scales for remote scene image segmentation.

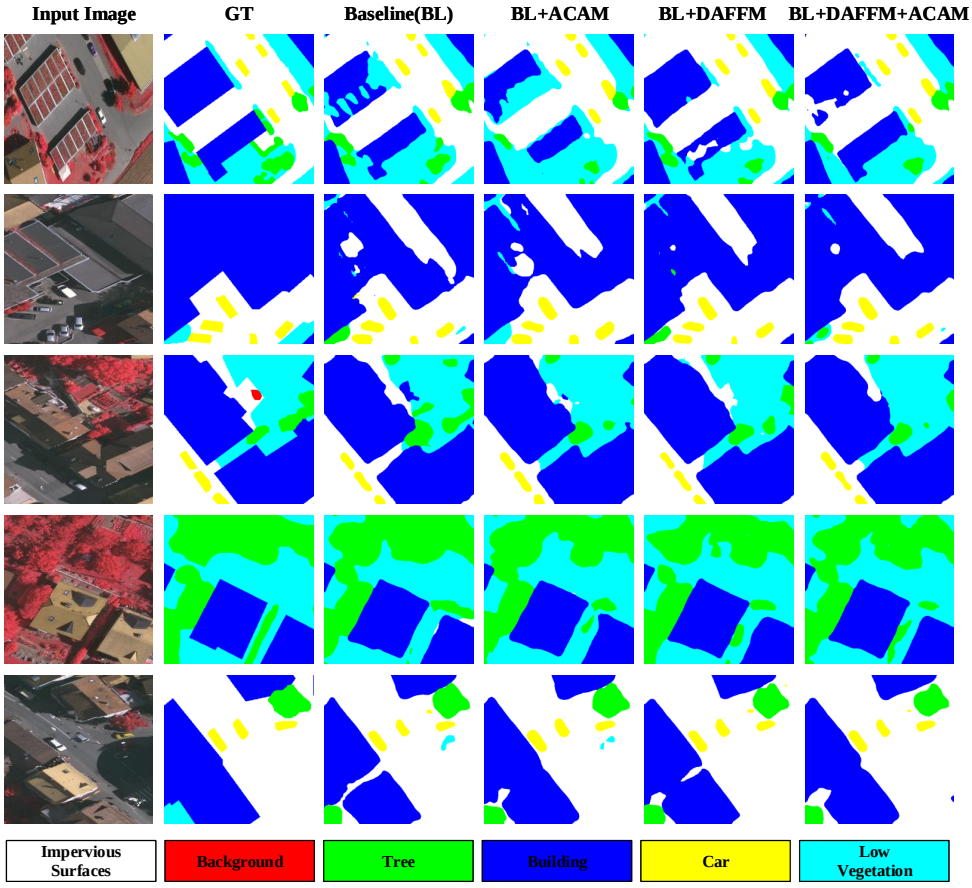

Figure 5. Visualization of the segmentation results for ablation study. From left to right, the input image, the ground-truth segmentations, and the results from variants of our methods.

4.2. Comparing with State-of-the-Art Methods

In order to show the effectiveness of the proposed method, we further perform comparisons with state-of-the-art semantic segmentation methods on both ISPRS 2D Semantic Labeling Vaihingen data [35] and GaoFen-2 [36] datasets. We select four popular methods, PSPNet [17], DeeplabV3 [21], Unet [16] and DANet [52] for comparison. We choose the mIOU of each category, overall accuracy and mIOU of all categories as the metrics to evaluate the performance of each method and the mIOU is the main evaluation metric between different categories, similar to [66]. Results based on ISPRS 2D Semantic Labeling Vaihingen data [35] and GaoFen-2 [36] datasets are shown in Tables 3 and 4, respectively.

Table 3. Comparison with state-of-the-art methods on ISPRS 2D Semantic Labeling Vaihingen data [35].

Methods	Imp.Surf.	Background	Tree	Building	Car	Low Veg.	OA (%)	mIOU (%)
Unet [16]	74.83	22.68	72.67	79.33	51.91	59.36	83.27	60.13
PSPNet [17]	78.81	44.13	75.74	84.77	63.02	64.65	86.27	68.52
Deeplabv3 [21]	79.71	42.64	76.76	86.03	67.04	65.40	86.90	69.60
DANet [52]	76.89	43.95	74.98	81.59	56.83	64.33	85.22	66.43
Baseline	79.69	36.15	76.27	85.69	65.59	64.73	86.66	68.02
Baseline + ACAF	79.30	42.31	75.81	85.23	64.64	65.19	86.44	68.75
Baseline + DAFFM	80.99	43.78	76.27	86.90	66.05	65.15	**87.17**	69.86
Baseline + DAFFM + ACAM	80.11	47.95	76.24	86.57	66.64	65.56	87.01	**70.51**

Table 4. Comparison with state-of-the-art methods on the GaoFen-2 dataset [36].

Methods	Road	Building	Tree	Grass	Bare Land	Water	Tran.	Imp. Surf.	Clutter	OA (%)	mIOU (%)
Unet [16]	38.48	81.54	67.13	60.41	61.85	39.28	46.00	0.0	0.0	73.54	43.85
PSPNet [17]	42.36	79.42	69.61	69.59	68.65	47.29	45.11	24.41	4.98	77.49	50.16
Deeplabv3 [21]	46.60	81.16	67.78	72.86	68.28	50.04	39.67	44.93	8.55	78.62	53.32
DANet [52]	43.07	80.66	66.46	67.54	66.26	47.52	41.55	49.64	4.22	76.42	51.88
Baseline	50.36	81.01	69.38	72.00	68.38	54.70	39.25	45.34	5.07	79.36	53.94
Baseline + ACAF	48.40	80.30	66.60	70.90	70.77	50.78	37.21	38.88	5.76	78.50	52.18
Baseline + DAFFM	47.39	81.16	68.76	72.14	76.09	52.50	39.78	46.16	11.72	79.85	55.08
Baseline + DAFFM + ACAM	51.80	82.21	69.18	73.11	76.50	53.87	41.90	48.75	15.53	**80.91**	**56.98**

As shown in Tables 3 and 4, DAFFM and ACAM have similar or better performance compared with most of the state-of-the-art semantic segmentation methods and our method achieves the best performance between all the state-of-the-art methods mentioned above on both datasets [65,66]. The comparison between different methods mentioned above further demonstrates the superiority of our method. We also visualize the segmentation results of each method on both datasets [65,66] respectively as illustrated in Figures 6 and 7. The comprehensive comparison mentioned above further demonstrates the effectiveness and superiority of our method.

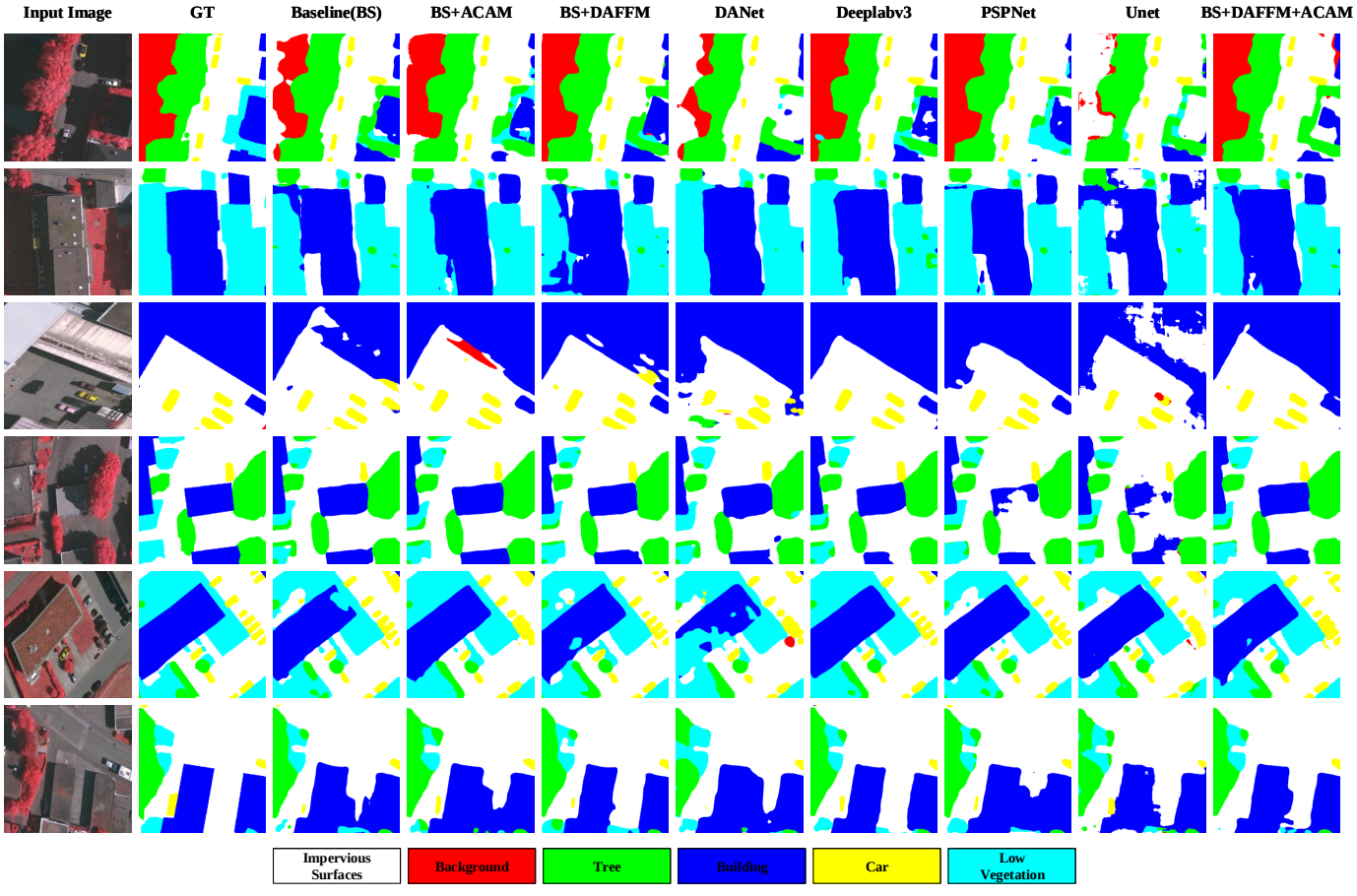

Figure 6. Visualization of the segmentation results for state-of-the-art methods based on ISPRS 2D Semantic Labeling Vaihingen data [35]. From left to right, the input image, the ground-truth segmentation results, the results from our methods and the state-of-the-art method.

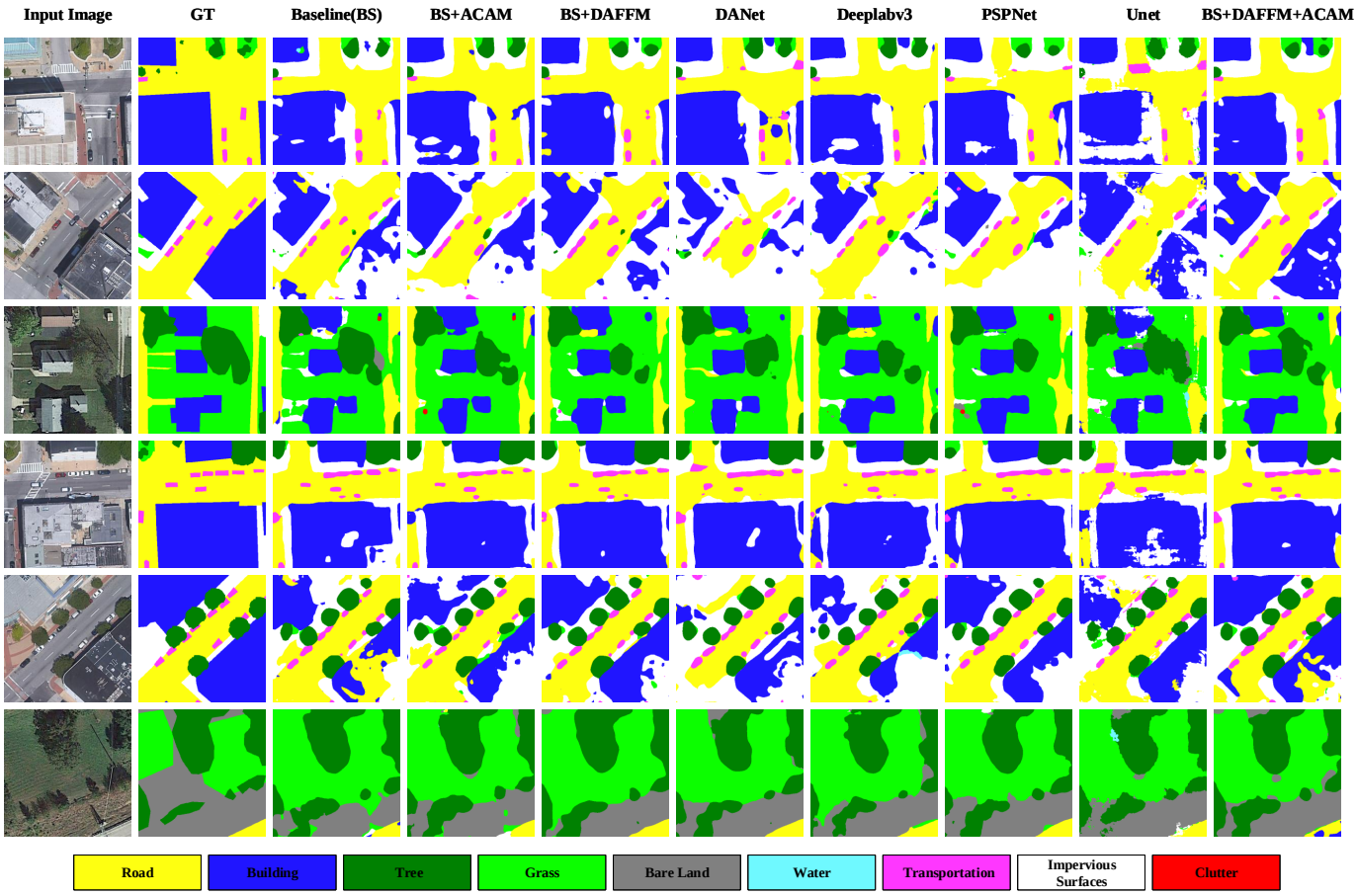

Figure 7. Visualization of the segmentation results for state-of-the-art method based on GaoFen-2 [36]. From left to right, the input image, the ground-truth segmentation results, the results from our methods and the state-of-the-art method.

5. Discussion

Previous methods [18,20,33] utilize various strategies to capture contextual cues. However, they treat all pixels with uniform weight for feature aggregation, which cannot capture the information in each channel with different weight vectors. To solve this issue, we integrate the contextual cues with the spatially-varying feature-weighting factor. Contributing to the adaptive contextual cue aggregating, the context information in the background of the objects can be aggregated by different weight vectors. Taking the second and fourth rows of Figure 7 as examples, the impervious surfaces, roads and buildings have a similar appearance. PSPNet [17] utilizes the spatial pyramid pooling module to aggregate multi-scale context. However, it cannot aggregate context adaptively and selectively, which leads to the wrong classification. Compared with PSPNet [17], Deeplabv3 [21] introduces the ASPP module which obtains better segmentation result. However, it also cannot achieve adaptive context aggregating, which leads to weakness in obtaining fine details and completing the object shape. Compared with the above methods, our method further alleviates this challenging issue by adaptive contextual cue aggregating. It is illustrated in Figures 6 and 7 that our method obviously obtains fine segmentation result for some objects.

Feature fusing is an effective strategy to obtain accurate segmentation results. However, methods [16,21,53] extensively used to fuse multi-scale feature maps ignore the differences between feature maps in different channels, which will limit the improvement of segmentation performance. To solve this issue, we integrate the DAFFM to make better use of the multi-scale feature maps, which fuses the feature maps in the different layers of CNNs based on both channel and spatial attention mechanisms. As illustrated in Figures 6 and 7, Unet [16] fuses the features by concatenating features directly without

distinguishing the difference and similarity of features and leads to the wrong prediction in the boundary region of the object. On the contrary, our method has better performance in the boundary region, which contributes to the effective feature fusion generated by the DAFFM.

Compared with the other methods mentioned above, our method improved the segmentation performance by context aggregating and feature fusion. It can be inferred that adaptive context aggregating and feature fusion are feasible to achieve better segmentation performance. Although our method achieved better performance in the experiment, it has a more complex structure. With the development of remote sensing technology, there will be more remote sensing images to be processed. Therefore, we believe that further improving the trade-off between the complexity of structure and accuracy is possible for future works that need more attention.

6. Conclusions

In this work, we presented a network for the challenging and meaningful task of precise semantic segmentation of VHR remote sensing images, which adaptively aggregates the contextual cues and flexibly fuses multi-scale feature maps based on spatial and channel attention mechanisms. Specifically, we first introduced ACAM to capture the multi-scale contextual cues of objects with adaptive weight vectors concentrating the global semantic information. Moreover, we adopted DAFFM based on spatial attention and channel attention mechanisms to explore the consistency of multi-scale features for better fusion results. In this way, our method makes better predictions for pixels of objects in complex remote scenes and improves the classification accuracy of pixels in the boundary of objects. Finally, the extensive ablation experiments based on ISPRS 2D Semantic Labeling Vaihingen data [35] and GaoFen-2 [36] data show that our method gives more precise segmentation results and achieves a state-of-the-art performance.

Author Contributions: Conceptualization, J.F. and H.S.; methodology, J.F. and L.C.; software, J.F. and Y.W.; validation, J.F.; writing-original draft preparation, J.F., H.S., L.C. and Y.W.; writing-review and editing, H.S. and L.C.; visualization, Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Chang Jiang Scholars Programme (Grant No.T2012122).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data set is public.

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

References

1. Marcos, D.; Volpi, M.; Kellenberger, B.; Tuia, D. Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 96–107. [CrossRef]
2. Fan, R.; Wang, H.; Cai, P.; Liu, M. Sne-roadseg: Incorporating surface normal information into semantic segmentation for accurate freespace detection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 340–356.
3. Fan, R.; Wang, H.; Wang, Y.; Liu, M.; Pitas, I. Graph Attention Layer Evolves Semantic Segmentation for Road Pothole Detection: A Benchmark and Algorithms. *arXiv* **2021**, arXiv:2109.02711.
4. Matikainen, L.; Karila, K. Segment-based land cover mapping of a suburban area—Comparison of high-resolution remotely sensed datasets using classification trees and test field points. *Remote Sens.* **2011**, *3*, 1777–1804. [CrossRef]
5. Tang, Y.; Zhang, L. Urban change analysis with multi-sensor multispectral imagery. *Remote Sens.* **2017**, *9*, 252. [CrossRef]
6. Yuan, Y.; Lin, J.; Wang, Q. Hyperspectral image classification via multitask joint sparse representation and stepwise MRF optimization. *IEEE Trans. Cybern.* **2015**, *46*, 2966–2977. [CrossRef] [PubMed]
7. Qian, Y.; Zhou, W.; Yan, J.; Li, W.; Han, L. Comparing machine learning classifiers for object-based land cover classification using very high resolution imagery. *Remote Sens.* **2015**, *7*, 153–168. [CrossRef]
8. Maxwell, A.E.; Warner, T.A.; Fang, F. Implementation of machine-learning classification in remote sensing: An applied review. *Int. J. Remote Sens.* **2018**, *39*, 2784–2817. [CrossRef]

9. Thanh Noi, P.; Kappas, M. Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery. *Sensors* **2018**, *18*, 18. [CrossRef] [PubMed]
10. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
11. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
12. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 12 June 2015; pp. 1–9.
13. 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.
14. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
15. 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]
16. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
17. 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. 2881–2890.
18. 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.
19. Lin, G.; Milan, A.; Shen, C.; Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1925–1934.
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.* **2017**, *40*, 834–848. [CrossRef]
21. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
22. Ciresan, D.; Giusti, A.; Gambardella, L.; Schmidhuber, J. Deep neural networks segment neuronal membranes in electron microscopy images. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 2843–2851.
23. Farabet, C.; Couprie, C.; Najman, L.; LeCun, Y. Learning hierarchical features for scene labeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 1915–1929. [CrossRef] [PubMed]
24. Gupta, S.; Girshick, R.; Arbeláez, P.; Malik, J. Learning rich features from RGB-D images for object detection and segmentation. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014, pp. 345–360.
25. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
26. 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.
27. Panboonyuen, T.; Jitkajornwanich, K.; Lawawirojwong, S.; Srestasathiern, P.; Vateekul, P. Semantic segmentation on remotely sensed images using an enhanced global convolutional network with channel attention and domain specific transfer learning. *Remote Sens.* **2019**, *11*, 83. [CrossRef]
28. Huang, Y.; Wang, Q.; Jia, W.; He, X. See More Than Once–Kernel-Sharing Atrous Convolution for Semantic Segmentation. *arXiv* **2019**, arXiv:1908.09443.
29. He, J.; Deng, Z.; Zhou, L.; Wang, Y.; Qiao, Y. Adaptive pyramid context network for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7519–7528.
30. Wang, H.; Fan, R.; Sun, Y.; Liu, M. Applying surface normal information in drivable area and road anomaly detection for ground mobile robots. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 24 October–24 January 2020; pp. 2706–2711.
31. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
32. Fan, R.; Bocus, M. J.; Zhu, Y.; Jiao, J.; Wang, L.; Ma, F.; Cheng, S.; Liu, M. Road crack detection using deep convolutional neural network and adaptive thresholding. In Proceedings of the IEEE Intelligent Vehicles Symposium (IV), Paris, France, 9–12 June 2019; pp. 474–479.
33. Peng, C.; Zhang, X.; Yu, G.; Luo, G.; Sun, J. Large kernel matters–improve semantic segmentation by global convolutional network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4353–4361.
34. Liu, J.; He, J.; Ren, J.S.; Qiao, Y.; Li, H. Learning to Predict Context-adaptive Convolution for Semantic Segmentation. *arXiv* **2020**, arXiv:2004.08222.

35. Available online: https://www2.isprs.org/commissions/comm2/wg4/benchmark/2d-sem-label-vaihingen/ (accessed on 8 June 2021).
36. Available online: http://sw.chreos.org/dataset/3 (accessed on 8 June 2021).
37. Li, Z.; Wang, R.; Zhang, W.; Hu, F.; Meng, L. Multiscale Features Supported DeepLabV3+ Optimization Scheme for Accurate Water Semantic Segmentation. *IEEE Access* **2019**, *7*, 155787–155804. [CrossRef]
38. Kuo, T.S.; Tseng, K.S.; Yan, J.W.; Liu, Y.C.; Wang, Y.C.F. Deep Aggregation Net for Land Cover Classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–23 June 2018; pp. 252–256.
39. Lin, D.; Ji, Y.; Lischinski, D.; Cohen-Or, D.; Huang, H. Multi-scale context intertwining for semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 603–619.
40. Ding, H.; Jiang, X.; Shuai, B.; Qun Liu, A.; Wang, G. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2393–2402.
41. Li, X.; Xu, F.; Lyu, X.; Gao, H.; Tong, Y.; Cai, S.; Li, S.; Liu, D. Dual attention deep fusion semantic segmentation networks of large-scale satellite remote-sensing images. *Int. J. Remote Sens.* **2021**, *42*, 3583–3610. [CrossRef]
42. Liu, Y.; Zhu, Q.; Cao, F.; Chen, J.; Lu, G. High-Resolution Remote Sensing Image Segmentation Framework Based on Attention Mechanism and Adaptive Weighting. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 241. [CrossRef]
43. Wang, H.; Wang, Y.; Zhang, Q.; Xiang, S.; Pan, C. Gated convolutional neural network for semantic segmentation in high-resolution images. *Remote Sens.* **2017**, *9*, 446. [CrossRef]
44. Li, L. Deep Residual Autoencoder with Multiscaling for Semantic Segmentation of Land-Use Images. *Remote Sens.* **2019**, *11*, 2142. [CrossRef]
45. Kang, W.; Xiang, Y.; Wang, F.; You, H. EU-Net: An Efficient Fully Convolutional Network for Building Extraction from Optical Remote Sensing Images. *Remote Sens.* **2019**, *11*, 2813. [CrossRef]
46. Liu, W.; Rabinovich, A.; Berg, A.C. Parsenet: Looking wider to see better. *arXiv* **2015**, arXiv:1506.04579.
47. 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, 18–23 June 2018; pp. 3684–3692.
48. Bilinski, P.; Prisacariu, V. Dense decoder shortcut connections for single-pass semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6596–6605.
49. Zhang, H.; Zhang, H.; Wang, C.; Xie, J. Co-occurrent features in semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 548–557.
50. Zhang, F.; Chen, Y.; Li, Z.; Hong, Z.; Liu, J.; Ma, F.; Han, J.; Ding, E. Acfnet: Attentional class feature network for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 6798–6807.
51. 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 (ECCV), Munich, Germany, 8–14 September 2018; pp. 267–283.
52. 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, 15–20 June 2019; pp. 3146–3154.
53. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
54. Tian, Z.; He, T.; Shen, C.; Yan, Y. Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3126–3135.
55. He, J.; Deng, Z.; Qiao, Y. Dynamic multi-scale filters for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 3562–3572.
56. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Learning a discriminative feature network for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1857–1866.
57. 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.
58. Eigen, D.; Fergus, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–12 June 2015; pp. 2650–2658.
59. Liu, F.; Shen, C.; Lin, G. Deep convolutional neural fields for depth estimation from a single image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, MA, USA, 7–12 June 2015; pp. 5162–5170.
60. Liu, F.; Shen, C.; Lin, G.; Reid, I. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 2024–2039. [CrossRef] [PubMed]
61. Lin, Z.; Feng, M.; Santos, C.N.d.; Yu, M.; Xiang, B.; Zhou, B.; Bengio, Y. A structured self-attentive sentence embedding. *arXiv* **2017**, arXiv:1703.03130.

62. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
63. Shen, T.; Zhou, T.; Long, G.; Jiang, J.; Pan, S.; Zhang, C. Disan: Directional self-attention network for rnn/cnn-free language understanding. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
64. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic Differentiation in Pytorch. 2017. Available online: https://openreview.net/forum?id=BJJsrmfCZ (accessed on 8 June 2021).
65. Mou, L.; Zhu, X.X. RiFCN: Recurrent network in fully convolutional network for semantic segmentation of high resolution remote sensing images. *arXiv* **2018**, arXiv:1805.02091.
66. Sun, X.; Shi, A.; Huang, H.; Mayer, H. BAS^4 Net: 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]