*remote sensing*

MDPI

*Article*

# Building Extraction from Very High Resolution Aerial Imagery Using Joint Attention Deep Neural Network

**Ziran Ye [1], Yongyong Fu [1], Muye Gan [1], Jinsong Deng [1], Alexis Comber [2] and Ke Wang [1,*]**

[1] Institute of Applied Remote Sensing and Information Technology, College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China; smd_ye@zju.edu.cn (Z.Y.); yyong_fu@zju.edu.cn (Y.F.); ganmuye@zju.edu.cn (M.G.); jsong_deng@zju.edu.cn (J.D.)

[2] School of Geography and Leeds Institute for Data Analytics, University of Leeds, Leeds LS2 9JT, UK; a.comber@leeds.ac.uk

* Correspondence: kwang@zju.edu.cn

Received: 11 November 2019; Accepted: 10 December 2019; Published: 11 December 2019


**Abstract:** Automated methods to extract buildings from very high resolution (VHR) remote sensing data have many applications in a wide range of fields. Many convolutional neural network (CNN) based methods have been proposed and have achieved significant advances in the building extraction task. In order to refine predictions, a lot of recent approaches fuse features from earlier layers of CNNs to introduce abundant spatial information, which is known as skip connection. However, this strategy of reusing earlier features directly without processing could reduce the performance of the network. To address this problem, we propose a novel fully convolutional network (FCN) that adopts attention based re-weighting to extract buildings from aerial imagery. Specifically, we consider the semantic gap between features from different stages and leverage the attention mechanism to bridge the gap prior to the fusion of features. The inferred attention weights along spatial and channel-wise dimensions make the low level feature maps adaptive to high level feature maps in a target-oriented manner. Experimental results on three publicly available aerial imagery datasets show that the proposed model (RFA-UNet) achieves comparable and improved performance compared to other state-of-the-art models for building extraction.

**Keywords:** building extraction; fully convolutional neural network (FCN); attention mechanism; high resolution aerial images

## 1. Introduction

Automatic extraction of buildings from remote sensing imagery is of paramount importance in many application areas such as urban planning, population estimation, and disaster response [1]. Assigning a semantic building class label to each pixel in very high resolution (VHR) imagery of urban areas is a challenging task because of high intra-class and low inter-class variabilities [2,3]. This is because in high resolution images, the building category contains many different sized manmade-objects in urban areas, where the amount of clutters is increasing—e.g., the shadow of tall buildings—the similarity of rooftops to some roads. The result is that it is difficult to label buildings reliably and accurately.

We have witnessed a rapid, revolutionary change in computer vision research, mainly driven by convolutional neural network (CNN) [4] and the availability of large scale training data [5]. Recently, several CNNs-based semantic segmentation methods have been used in building extraction from earth observation images [6–8]. The patch-based CNNs methods [9–13] were initially adopted for prediction in dense urban areas. These patched-CNNs label the center pixel by processing an image patch through a neural network. They tend to be computationally expensive and are usually used

to detect large objects [14,15]. Since Long et al. [16] adapted the classification network into fully convolutional network (FCN) for semantic segmentation, FCN and its extensions have gradually become the preferred solution in the field of semantic labeling [17–20]. Though FCN-based methods can produce dense pixel-wise output directly, the pixel-wise classification derived from the final score map is quite coarse because of the sequential sub-sampling operations in the FCN.

To address the problem of coarse predictions, recent research [21–26] have further improved FCN-based methods for semantic labeling of remote sensing images. There is a growing body of literature that many studies [27–31] employ the encoder–decoder architecture with skip connection. UNet [32], a typical model in the style of encoder–decoder, reuses low-level information to refine the output, and results in better performance. For obtaining accurate labeling of VHR images, an effective structure to integrate the high-resolution, low-level features, and the low-resolution, high-level features is needed. The skip connection fuses features so as to compensate the loss of spatial information caused by repeating local operations (e.g., pooling and strided convolution). Features via skip connection are multi-scale in nature due to the increasingly large receptive field sizes [33]. However, one thing to note is that most existing approaches that are built on top of a contemporary classification network are good at aggregating global contexts. While the reuse of information from early encoding layers contributes to localization in the decoding phase, it may introduce redundant information which results in over-segmentation [34] and unexpected ambiguous representations [35,36]. To be specific, the low level features in the encoder are computed in the shallow layers of the network, while the high level features in the decoder are computed in the deep layers of the network. Obviously, we can assume that the latter has undergone more processing and there is a semantic gap between the features of encoder and decoder. For example, a deep layer in the decoding stage may confidently discriminate between a gray pixel belonging to 'asphalt roads' or 'rooftops', because more global contexts are passed through a long path from the low layers to the high layers. However, the signals from the symmetric layer early have different levels of discrimination that are specific to the primary class 'impervious surface' and therefore express confidence in both subclasses. As a result, integrating these features directly through skip connection may decrease the accuracy of prediction. A new research has shown that fusing semantically dissimilar features from the encoder and decoder subnetworks directly can degrade segmentation performance [37]. Thus, it is important to bridge the semantic gap between features of encoder and decoder prior to fusion.

In recent years, several researchers have begun to apply attention mechanisms to CNNs. Initially, attention in CNNs was used to interpret the gradient of a class output score with respect to the input image [38]. Later trainable self-attention was deployed for image captioning, image classification, object detection, and image segmentation [39–42]. A large body of literature exploring different gating architectures has emerged. For instance, Oktay et al. [43] proposed a self-attention gating module that can be utilized in FCN models for medical image segmentation. Zagoruyko et al. [44] improved the performance of a student CNN by transferring the attention maps from a teacher network. Different from the above, where they used the grid-attention technique to capture spatial salient regions, Hu et al. [45] proposed channel-wise attention to highlight important feature dimensions. Subsequent studies [46–49] have demonstrated the performance of channel-wise attention mechanism in the semantic segmentation task. In remote sensing, some attempts [50–52] have been made to adopt attention mechanisms on the building extraction task. Yang et al. [52] used a spatial attention module that weights map generated by applying sigmoid function at the deep features. Pan et al. [50] used a generative adversarial network with spatial [34] and channel [45] attention to extract buildings. Though there are a few differences in the above attention modules, most of these implementations can be attributed to the use of self-attention to enhance the representation of single-layer features.

Since the attention can model interdependency and adjust the response of a position or a channel in the input feature maps, we expect to exploit it to alleviate the semantic difference between features from different depths in the skip connection. Similar to [39,46], we employ a joint attention module (RFA) in the deep neural network, while our focus is to bridge gap between hierarchical representations. To this

end, we proposed an attention re-weighting process that could be integrated into UNet model for the building extraction task in VHR images. The proposed attention module emphasizes meaningful features and suppress insignificant features along both channel and spatial dimensions adaptively, under the guidance of deep features. Benefitting from global context information captured by joint attention, the semantic information of high spatial resolution but low level features in the encoder are gradually enriched in a task-oriented direction before fusion. In summary, the contributions of our work are summarized as follows:

(1) We implement joint spatial and channel-wise attention mechanism to enhance consistency of features across layers in the U-shaped FCN. Experimental results show that using attention jointly is effective to reduce semantic differences between features.

(2) We integrate the proposed attention module into existing UNet model and propose an end-to-end method (RFA-UNet) for the building extraction task, which attains comparable and stable performance with other state-of-the-art model on three public datasets.

The remainder of this paper is organized as follows. Section 2 introduces the proposed method. The experimental results are presented in Section 3. The discussion about the method and experiments is given in Section 4. Section 5 concludes this paper.

## 2. Methods

In this section, we wish to put forward an end-to-end method (RFA-UNet) based on the common semantic segmentation architecture UNet with a new attention module. Our approach leverages the benefits of typical segmentation architecture with the skip connections. An overview of the proposed architecture is shown in Figure 1. First, input images are progressively filtered (convolution) and downsampled by factor of 2 (pooling) at each level in the contracting path. Second, features from each encoding stage are filtered by attention module before skip connection. Different from most existing methods, we introduce discriminative information from coarser scales to help generate joint attention maps. After that, the refined features are concatenated to the corresponding decoding features again through skip connection. The rest of this section describes the details of RFA-UNet.

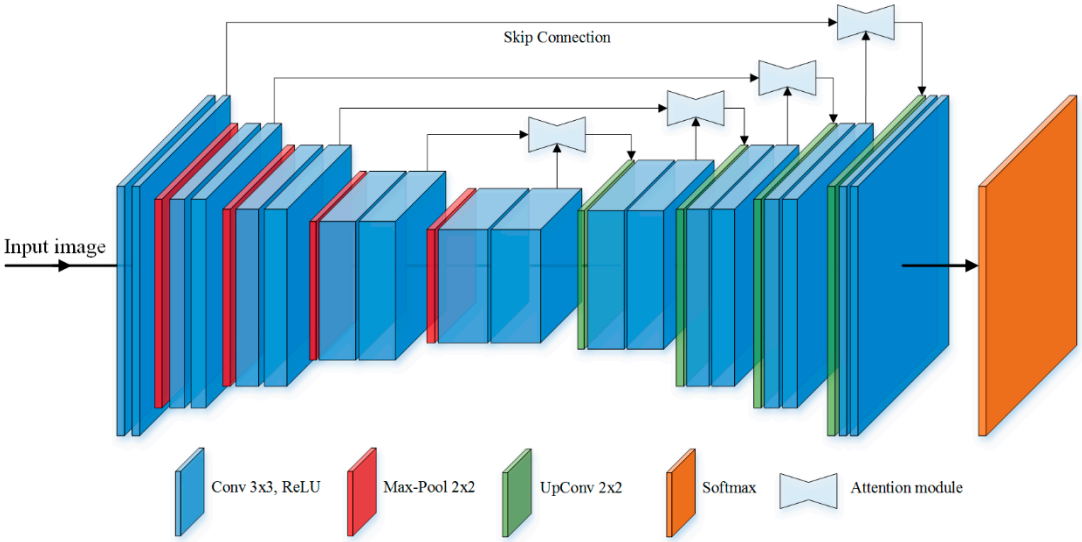



**Figure 1.** Proposed architecture (RFA-UNet) for building extraction of VHR images.

### 2.1. Semantic Labeling Using UNet

In general, the UNet model including an encoder and a decoder can make dense pixel-wise prediction naturally. UNet constructs the encoder part by applying a classification network. A cascaded convolution and pooling operations downsamples the output feature maps at each stage and increases the feature map channel number simultaneously through the contracting path. To generate results that

are the same size as the input image, an expansive path symmetric to the contracting path is adopted as the decoder. The size of feature maps is retrieved gradually, and each upsampling operation is followed by two convolution layers. In order to yield more precise localization, the network propagates spatial precision from earlier layer in encoding to deep layers at the decoding stage, i.e., skip connection.

It is well known that a deep network is built upon supervised training in a big dataset such as the ImageNet dataset [53]. In most cases, however, manual labelling for training data is a costly task, and it is also the same when using CNN for remote sensing classification. UNet has proven itself useful for segmentation problems with a relatively small datasets, e.g., satellite image analysis and medical image analysis [54,55]. For this reason, we choose UNet as the baseline architecture for our study.

### 2.2. Residual Feature Map Attention

The complex structure of different buildings increases the difficulty of determining the building outlines in VHR images. Though the low-level features captured by the earlier layers can help refine the feature maps in the deepen layers via skip connections, these should be employed with caution to avoid introducing inconsistencies across different stages. In this work, we wish to make the low-level feature maps $X_l \in \mathbb{R}^{H\times W\times C_l}$ adaptive to high-level feature maps $X_h \in \mathbb{R}^{H\times W\times C_h}$ in the skip connection. First, the feature maps are re-weighted by a channel-wise weighted vector, thus the network tends to learn the most salient features that contribute to the classification. Then, a spatial attention map $\widetilde{X}_S$ shows where the network focuses in order to highlight informative regions, as a complement to the channel attention map $\widetilde{X}_C$. Figure 2 illustrates the structure of the residual feature map attention (RFA). The proposed joint attention module can be summarized as

$$\widetilde{X}_C = f_C(X_l, \alpha), \tag{1}$$

$$\widetilde{X}_S = f_S(\widetilde{X}_C, \beta), \tag{2}$$

$$\widetilde{X} = X_l \oplus \widetilde{X}_S, \tag{3}$$

where $\alpha$ represents channel attention weights, $\beta$ represents spatial attention weights, $f_C(\cdot)$ denotes multiplication of feature maps and corresponding weights on the channel dimension, $f_S(\cdot)$ is the pixel-wise multiplication between spatial regions of feature maps and corresponding weights and $\oplus$ denotes element-wise addition.

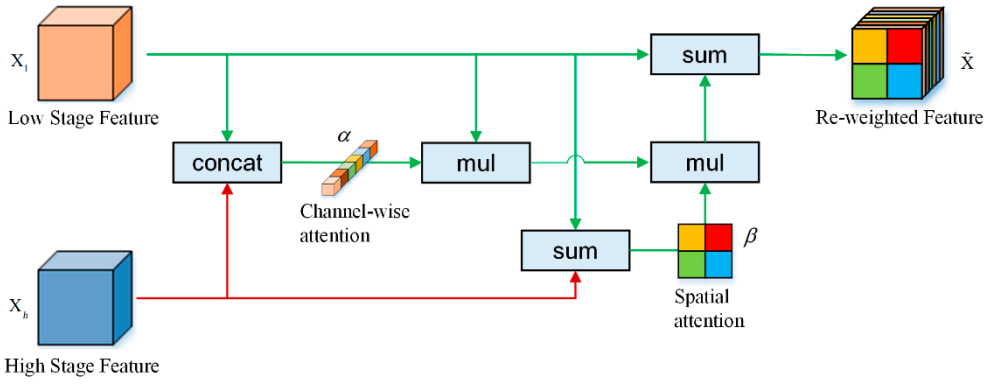



**Figure 2.** Diagram of the proposed attention module. Under the guidance of high stage features, low stage features are refined adaptively before passing through skip connection.

At last, a residual mapping is used to obtain the output of attention schemes. This short connection draws on the idea of the residual network [56], making the network easier to optimize.

(1) Channel Attention: Each filter performs pattern detection, and each channel of the feature map is a response activation of the corresponding convolutional filter [39]. In a standard CNN, the importance of each channel is considered to be the same. Thus, applying attention mechanisms

to feature channels intuitively distinguishes the features in different stages, and a channel attention map is produced to enhance feature consistency. For channel-wise attention, we first upsample the high stage features $X_h$ using bilinear interpolation to make them have the same shape (except in channel dimension) as the low stage features $X_l$. Then the two sequences of features are concatenated to generate channel-wise statistics. The global information are computed using global average pooling for the concatenated features. The generated vector $z \in \mathbb{R}^{H \times W \times C_l}$, which can be interpreted as a set of the channel descriptors for the image, is calculated by

$$\begin{aligned} z &= \frac{1}{H \times W} \sum_{i=1}^{W} \sum_{j=1}^{H} ([X_l; X_h]) \\ &= [z_1, z_2, \ldots, z_C], \end{aligned} \tag{4}$$

where [;] denotes the concatenate operation, scalar $z_i$ represents $i$-th channel descriptor.

To ensure that the module can learn nonlinear interaction between channels, the channel vector $z$ is passed through two fully connected layers. Then the channel attention vector $\alpha$ is obtained with a sigmoid activation

$$\alpha = \sigma_2(W_2 \sigma_1(W_1 z)) \tag{5}$$

where $\sigma_1$ denotes the ReLU function and $\sigma_2$ the sigmoid activation. Two fully connected layers with parameters where $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ with ratio $r$ to reduce dimensions for simplicity and $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$ to restore dimensions.

(2) Spatial Attention: In a standard CNN, a global image descriptor derived from fully connected layers maps the input into a high-dimensional space in order to make the classes linearly separable [38]. However, using only the global feature representation to classify pixels ignores local spatial structural characteristics that need to be considered for semantic labeling tasks. Although FCN has made the architecture more suitable for local positioning, the importance of each pixel location is considered equal. In general, the identification of a pixel needs to consider its spatial context, and near pixels are more related to each other [57]. Therefore, our spatial attention module is designed to pay more attention to the semantic regions. A spatial attention map is produced to emphasize or suppress feature responses in different spatial locations. We reshape current low stage features $X_l$ by flattening the height and width to $V_l = [v_1, v_2, \ldots, v_m]$, where $v_i$ represents the $i$-th location pixel-wise vector of length $C$, and $m = W \times H$. Similarly, a gating vector $V_h$ obtained from the upsampled high-level features is introduced as a guidance. We employ linear transformations to make two vectors have the same length, i.e., the same dimensional space. Finally, we add them to generate spatial attention map $\beta$, formulated as

$$\beta = \sigma_2\Big(W_3\Big(\sigma_1\Big(W_1^T V_l + {W_2}^T V_h\Big)\Big)\Big), \tag{6}$$

where $\sigma_1$ denotes the ReLU function and $\sigma_2$ the sigmoid activation, linear transformations $W_1 \in \mathbb{R}^{C_l \times \text{int}}$, $W_2 \in \mathbb{R}^{C_g \times \text{int}}$, $W_3 \in \mathbb{R}^{\text{int} \times 1}$ are computed using channel-wise $1 \times 1$ convolutions for the two inputs.

### *2.3. Network Architecture*

We use VGG16 that consists of 16 sequential layers as the feature encoder, but we remove the full connected layer from the network. The architecture of the encoder is presented in Table 1. All convolutional layers followed by a ReLU activation function have $3 \times 3$ kernels and the number of output channels doubles after the max pooling operation. The output of 512 channels feature maps is served as a bottleneck of network, separating the encoder from the decoder.

**Table 1.** Architecture of encoder.

| Input (320 × 320 RGB Image) | | |
|---|---|---|
| **Stage** | **Output Size** | **Template** |
| 1 | 320 × 320 | Conv3-64<br>Conv3-64 |
| | 160 × 160 | Maxpool |
| 2 | 160 × 160 | Conv3-128<br>Conv3-128 |
| | 80 × 80 | Maxpool |
| 3 | 80 × 80 | Conv3-256<br>Conv3-256<br>Conv3-256 |
| | 40 × 40 | Maxpool |
| 4 | 40 × 40 | Conv3-512<br>Conv3-512<br>Conv3-512 |
| | 20 × 20 | Maxpool |
| 5 | 20 × 20 | Conv3-512<br>Conv3-512<br>Conv3-512 |
| **Bottleneck (20 × 20 × 512)** | | |

In the symmetric decoder part, low spatial resolution deep features are upsampled with a deconvolution layer. The upsampled feature map is regarded as providing consistency guidance for the corresponding earlier low-level feature map in encoding, and both features are transmitted into the proposed attention module (RFA) to obtain weighted feature maps. The reinforced meaningful features are then concatenated with the upsampled high level features via skip connection (see Figure 3).

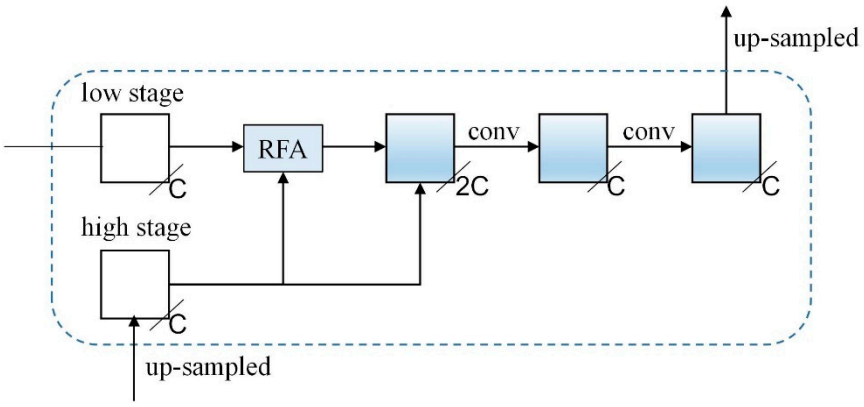



**Figure 3.** A decoding stage embedded with proposed attention module, where 'RFA' refers to the attention module, and 'C' refers to channel number of features.

The output of the network model passes through the softmax nonlinearity layer and results in an image where each pixel corresponds to a probability of belonging to buildings.

### *2.4. Training and Inference*

In the training stage, aerial images, and corresponding annotation images are both inputs to the model. As the building extraction task is a pixel-wise classification problem, we add cross entropy loss and dice coefficient loss together as the loss function of segmentation, defined by

$$L_{be} = -\frac{1}{N}\left(\sum_{i=1}^{N} y_i \log(\hat{y}_i)\right), \tag{7}$$

$$L_{dice} = \frac{1}{N}\sum_{i=0}^{N} \frac{2 \cdot y \cdot \hat{y}}{y + \hat{y}}, \tag{8}$$

$$L_{seg} = L_{be} - \log L_{dice} \tag{9}$$

where $\hat{y}_i$ is the label (1 for building, 0 for background) of pixel and $y_i$ is predicted probability for the pixel.

As the satellite images are large, we crop patches from original images and feed them into the network for training. In the prediction stage, we combine predictions back into the original size. However, splitting the image into small tiles and then stitching them together later sometimes results in blocking artifacts at the borders. In some ways, predicting a pixel in the central part of the image is much easier than at the edge because the latter have limited shared contextual information. To predict the pixels in the edge area of the image, one way is to make the predictions on overlapping patches and crop the edges, but we propose a more efficient approach. Firstly, we extrapolate the missing context by mirroring the input image [32]. After that, we add a cropping layer to the output layers of the network, similar to [55], which solves two problems simultaneously: (1) in the predicting phase, it takes advantage of contextual information in the margin; (2) overlapping edges of each patch are cropped automatically in the prediction stage. Details of experimental settings is introduced in Section 3.2.

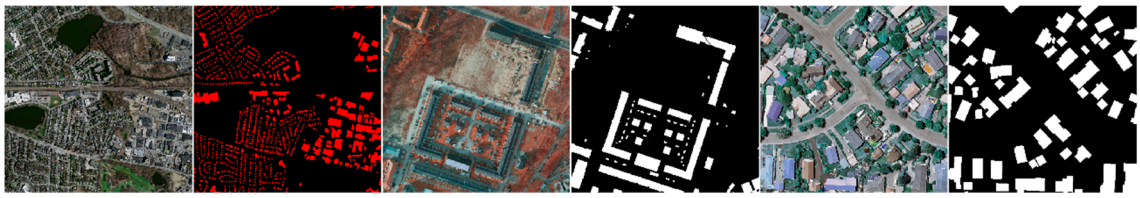
## 3. Results

### *3.1. Dataset*

Massachusetts Building Dataset (Mass. Buildings): This is proposed by Mnih [9]. It includes 151 RGB images of the Boston area with a spatial resolution of 1 m. The ground truth obtained from the OpenStreetMap project are all available. There are 137 images in the training set, 10 images in the test set, and 4 images for validation.

ISPRS Potsdam Challenge Dataset (Potsdam) [58]: This dataset contains 38 images with a spatial resolution of 5 cm. The size of each tile is $6000 \times 6000$ pixels. Among them, 24 images with available ground truth are provided for training and 14 images are remained for test. We randomly split the 24 images into 17 for training and 7 for validation. It is noted that we only use three-band IRRG images for fair comparison with existing models.

WHU Aerial Dataset (WHU) [59]: This dataset contains 8189 RGB tiles with 0.3 m ground resolution, including 187,000 samples of building in New Zealand. These $512 \times 512$ images are divided into three parts by the provider: 4736 tiles for training, 1036 tiles for validation, and 2416 tiles for test. Each tile has a corresponding Boolean raster map derived from the building vector map. Figure 4 shows some images and reference data from three datasets.



(**a**) (**b**) (**c**)

**Figure 4.** The image samples and corresponding reference images on the three datasets. The label of (**a**) Massachusetts Building includes two classes: building (red) and background (black). The label of (**b**) Potsdam Challenge, and (**c**) WHU Aerial includes two classes: building (white) and background (black).

### *3.2. Experimental Setting*

In order to train an effective deep model using a relatively small dataset, we cropped the raw images into small patches with the size of 320 × 320 pixels with the overlap of 80 pixels. Only with the WHU dataset did we maintain the original size of images at 512 × 512 pixels, following the settings of the providers. For each patch, we applied data augmentation consisting of flipping in the vertical or horizontal dimension and rotation of 90 degrees. It should be noted that for a given patch, we performed above transformations randomly rather than applying all of them each time. Table 2 shows the detailed number of patches of the three augmented datasets.

**Table 2.** Detailed information of experimental setting on three datasets.

| Dataset | Training Set | | Validation Set | | Test Set |
|---|---|---|---|---|---|
| | Images | Patches | Images | Patches | Images |
| Mass. buildings | 137 | 16,439 | 4 | 479 | 10 |
| Potsdam | 17 | 21,250 | 7 | 8750 | 14 |
| WHU | 14,208 | / | 3108 | / | 2416 |

In the prediction stage, we followed the abovementioned overlay cropping and stitching process to output the classification result of the large aerial imagery. Firstly, the size of test patch was set to 704 × 704 pixels for Mass. Buildings and Potsdam, and the overlapping pixels between adjacent patches were 204 pixels. Then, we cropped the outputs along the edge by 102 pixels and got images with size of 500 × 500 pixels, which was easily stitched into a large test image. For the WHU dataset, since the images of its test sets have been cropped into isolated tiles of 512 × 512 pixels, there is no need to do the process like the aforementioned datasets.

We implemented our models in the experiments by using the Keras framework with Tensorflow backbend. We initialized network parameters using Xavier uniform [60] and adopted Adam [61] as the weights optimization algorithm, with initial learning rate of 0.001. We set a batch size of 8 to suit the memory of graphics of the workstation used in the study. All experiments were processed on a desktop with 32GB of RAM and a 24GB Nvidia P6000 GPU.

### *3.3. Evaluation Metrics*

We used overall accuracy to evaluate the global performance of the methods. In addition, the F1-score of the positive (building) class and Intersection over Union (IoU) were used to evaluate classification performance. The F1-score and IoU metric are defined by

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{10}$$

$$\mathrm{F1} = 2\frac{\mathrm{Precison} \times \mathrm{Recall}}{\mathrm{Precison} + \mathrm{Recall}}, \tag{11}$$

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{FP} + \mathrm{TP} + \mathrm{FN}}, \tag{12}$$

where TP is the number of true positives, FP the number of false positives, and FN the number of false negatives.

### *3.4. Evaluations of Attention*

We evaluated the effect of the proposed joint attention module in UNet for building extraction in the VHR images. Furthermore, we compared our attention module with existing advanced attention modules on the three datasets. The information of these attention mechanisms is described as follows: (1) CAB [48]: A channel-wise attention block, which reweights the feature maps of low stage by using a weight vector inferred from concatenated features stage by stage. (2) GRID [43]: A grid attention gate module for medical imaging that learns to identify salient image regions of varying shapes and sizes by a grid signal conditioned to image spatial information from deep features. (3) DenseAN [52]: A spatial attention fusion module uses the high level features activation output to reweight the low level features before summation. (4) DualAN [62]: A dual attention module for scene segmentation which captures long-range contextual information in spatial dimension and channel dimension respectively. (5) RFA (Ours): A joint residual attention module consists of channel attention and spatial attention for enhancing the semantic consistency of features across layers. In the training phase, the training parameters and strategies adopted for these methods are same as ours.

The comparisons with different attentions on the three test sets are presented as follows:

(1) Mass. Buildings: As shown in Figure 5, CAB (Figure 5c) and Ours-RFA (Figure 5g) achieved better global performance than other attention methods visually. There are many FPs and FNs in the results of both GRID (Figure 5d) and DenseAN (Figure 5e). DualAN (Figure 5f) had difficulty in recognizing the small and dense buildings. Figure 6 shows the close-ups (as marked in yellow rectangles in Figure 5a) of the results for detailed inspection. The results in Figure 6 demonstrating that most buildings were correctly identified using all five methods, but CAB, GRID, and DenseAN tended to misclassify pixels as some FPs are found in the areas covered by ground or shadows (Figure 6c–e). DualAN (Figure 6f) performed better in the large building pixels. However, the FPs and FNs in the dense residential areas indicate that DualAN does not perform well enough.

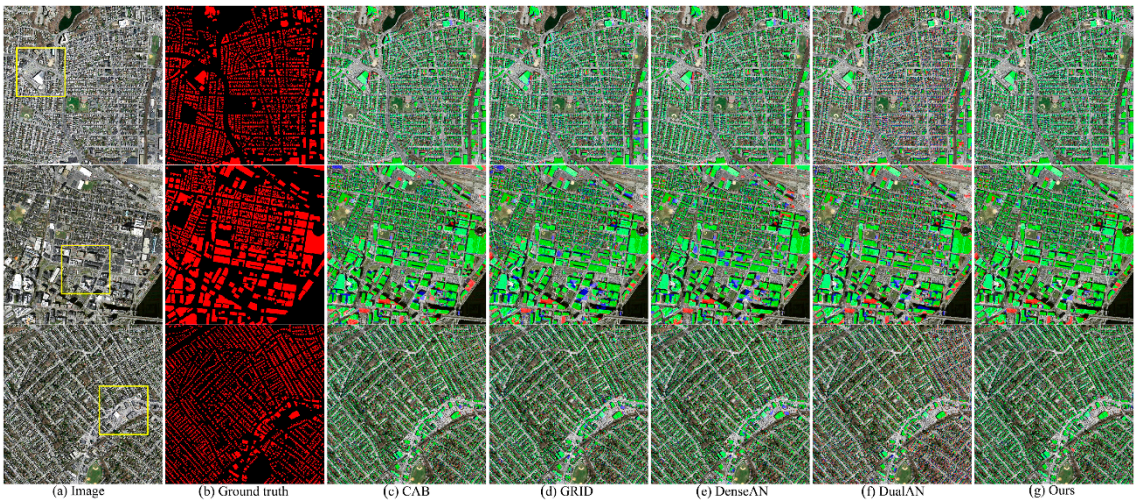



**Figure 5.** Comparison with the state-of-the-art attention methods on Mass. Buildings test sets. True positive (TP) is marked as green, false positive (FP) as blue and false negative (FN) as red.


**Figure 6.** Close-ups of comparison with the state-of-the-art attention methods on Mass. Buildings test sets. True positive (TP) is marked as green, false positive (FP), as blue and false negative (FN) as red.

(2) Potsdam: Figure 7 displays the classification results on Potsdam test sets. Ours-RFA (Figure 7g) outperforms the other four methods because there are many FPs generated in the prediction (Figure 7c–f). The result of DenseAN is slightly better, except that some FPs in the first two rows of Figure 7e. The FPs in the results of CAB, GRID, and DualAN (Figure 8c,d,f) indicate that these models are sensitive to the low vegetation and roads, and they frequently misclassify vegetation pixels similar to the color of the rooftops as building pixels. Though the result of DualAN in the last row (Figure 8f) has more TPs, it also tends to have more FPs. Compared the other attention methods, except DualAN, Ours-RFA (Figure 8g) achieves better performance which has more TPs and less FPs.

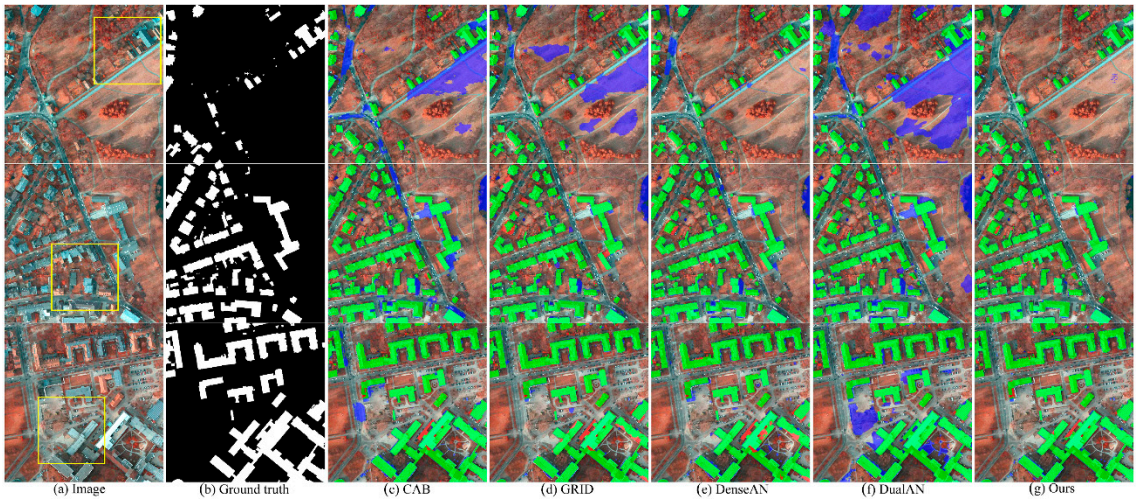



**Figure 7.** Comparison with the state-of-the-art attention methods on Potsdam test sets. True positive (TP) is marked as green, false positive (FP), as blue and false negative (FN) as red.

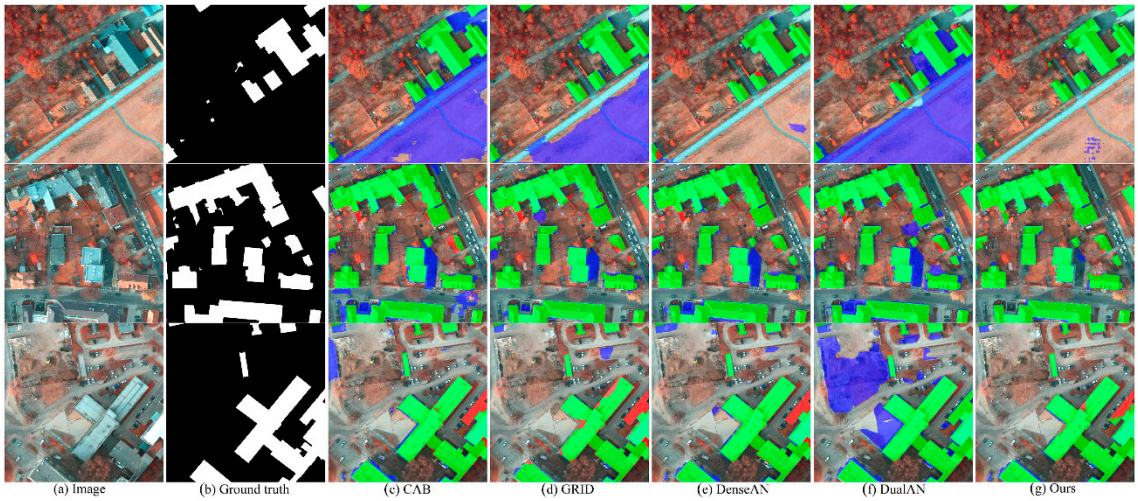



**Figure 8.** Close-ups of comparison with the state-of-the-art attention methods on Potsdam test sets. True positive (TP) is marked as green, false positive (FP), as blue and false negative (FN) as red.

(3) WHU: Figure 9 exhibits the results of proposed attention method and its comparisons. The result obtained by using Ours-RFA (Figure 9g) are more complete, which indicates that the proposed attention module improves the performance of labeling fine-structured buildings in the VHR images. Closer inspection of the 2 rd row in Figure 10 testifies our point, only Ours-RFA method identified almost building pixels and a small number of misclassified pixels, while other methods struggled with identifying the rooftop and the results of them have many FNs. The results of CAB (Figure 9c) are relatively good, but still some FPs presented in Figure 10c, indicating that CAB did not distinguish ground pixels well enough. Most FNs in the results of GRID (see the third row in Figure 10d) implies that GRID does not fully utilize the context information and lack ability of identifying rooftop pixels with complex texture.

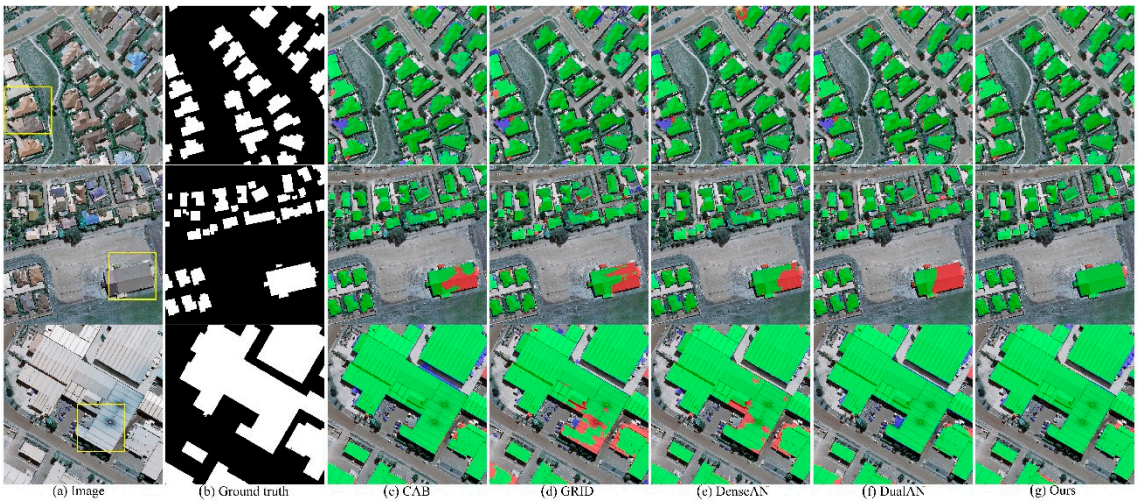



**Figure 9.** Comparison with the state-of-the-art attention methods on WHU test sets. Building (white) and background (black). True positive (TP) is marked as green, false positive (FP) as blue and false negative (FN) as red.

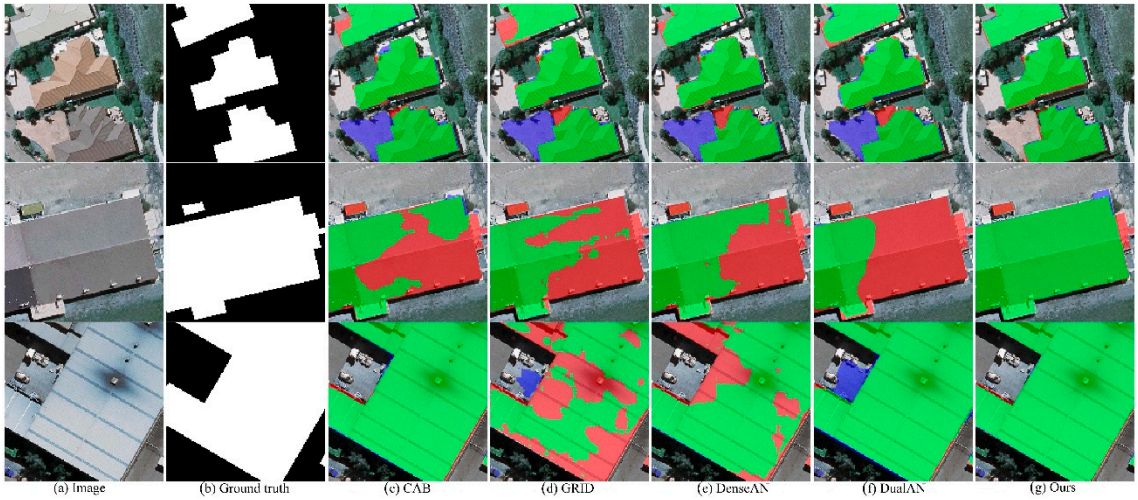



**Figure 10.** Close-ups of comparison with the state-of-the-art attention methods on WHU test sets. Building (white) and background (black). True positive (TP) is marked as green, false positive (FP) as blue and false negative (FN) as red.

Table 3 provides the summary statistics for quantitve comparisons of different attention modules. Ours-RFA obtained the best result with the overall accuracy, the F1-score and the IoU score among all attention methods on three datasets, and the result of quantitative comparison is consistent with visual effect. Both on the Mass. Buildings test sets and WHU test sets, CAB achieved the second best performance. DenseAN had a comparable result of the overall accuracy with the proposed RFA on the Potsdam test sets but the IoU score was 6.63% lower than that of Ours-RFA. The quantitive results of DualAN on three datasets are not satisfactory, implying that the strategy of only applying refinement on the deep feature is not stable for building extraction in the VHR images.

**Table 3.** Quantitative comparison with other attention modules (%) on three test sets, where values in bold are the best and the underlined values are the second best.

| Dataset | Attention | Overall Accuracy | F1-Score | IoU |
|---|---|---|---|---|
| Mass. Buildings | CAB | <u>93.83</u> | <u>83.00</u> | <u>70.94</u> |
| | GRID | 93.33 | 82.35 | 70.00 |
| | DenseAN | 93.54 | 82.36 | 70.01 |
| | DualAN | 89.83 | 71.17 | 55.24 |
| | RFA(Ours) | **94.71** | **85.65** | **74.91** |
| Potsdam | CAB | 92.81 | 85.89 | 75.27 |
| | GRID | 94.95 | 89.38 | 80.80 |
| | DenseAN | <u>95.97</u> | <u>91.58</u> | <u>84.48</u> |
| | DualAN | 95.62 | 90.90 | 83.33 |
| | RFA(Ours) | **97.79** | **95.35** | **91.11** |
| WHU | CAB | <u>98.34</u> | <u>92.51</u> | <u>86.07</u> |
| | GRID | 97.92 | 90.57 | 82.77 |
| | DenseAN | 98.06 | 91.28 | 83.96 |
| | DualAN | 97.89 | 90.40 | 82.49 |
| | RFA(Ours) | **98.84** | **94.75** | **90.02** |

### *3.5. Comparison with State-of-the-Art*

To evaluate the effectiveness of the proposed attention model, comparisons were made with other existing FCNs methods for building detection. The list of models includes the following: (1) RFA-UNet (Ours): an encoder–decoder style fully convolutional network with extended hybrid attention module. (2) UNet: an architecture originally proposed for segmentation of biomedical

images that captures context information at multiple scales via contracting and expansive paths. (3) SegNet [63]: an encoder–decoder architecture for scene segmentation, in which the decoder uses pooling indices computed in the max-pooling step of the corresponding encoder to perform non-linear upsampling. (4) RefineNet [64]: a multi-path refinement network that exploits multi-level features for high-resolution prediction with long-range residual connections, achieving the state-of-the-art results on several public datasets. (5) FC-DenseNet [65]: a model adopts dense connected convolutional networks into U-shape architectures to tackle the problem semantic segmentation. By using dense connections, multiple level features are concatenated iteratively to form a dense block. It should be noted that we implemented the methods above (the training parameters for these methods are same as ours) and also incorporated some advanced numerical results on each of the three datasets reported in the literatures [52,66,67].

Figures 11–13 demonstrate the close-up views of the five classification results using three subset images of three test sets, respectively. SegNet obtained comparable results on Potsdam (Figure 12c) but cannot distinguish large building objects on WHU, and obvious FNs appeared in the last two rows of Figure 13c, indicating that SegNet is not robust to identify complex manmade objects. The results of FC-DenseNet and RefineNet are relatively smooth, while they are still less accurate. As shown in Figures 11d and 12e, FC-DenseNet and RefineNet did not perform well, as many FPs and FNs appeared in the second row of their results. Similar to the results on Mass. Buildings, there were also some FPs in the results of FC-DenseNet and RefineNet (see the first row in Figure 12d,e) on Potsdam. These findings suggest that their strategies for simply reusing features densely or using long-range residual connections are not efficient enough due to the categorical ambiguity of the low level features. Our RFA-UNet model were more effective in the recognition of building objects on three test sets. Though the result of RFA-UNet have a few flaws, they still perform more precise localization and accurate labeling (see Figure 11g, Figure 12g, and Figure 13g). Meanwhile, the performance of our model also shows that the RFA module has improved the classification ability of network, as compared to the performance of the UNet method.


**Figure 11.** Comparison with the state-of-the-art attention methods on Mass. building test sets. True positive (TP) is marked as green, false positive (FP), as blue and false negative (FN) as red.

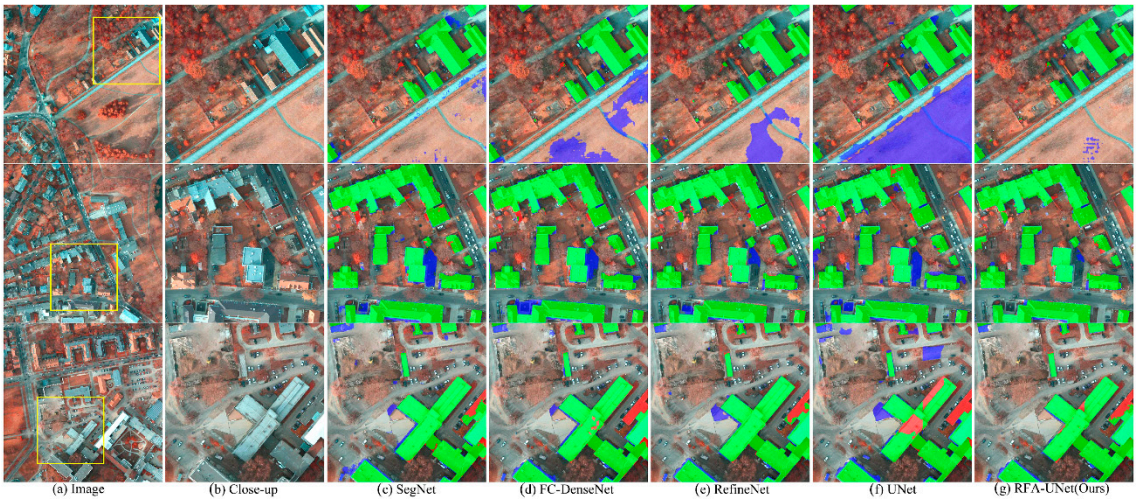



**Figure 12.** Comparison with the state-of-the-art attention methods on Potsdam building test sets. True positive (TP) is marked as green, false positive (FP), as blue and false negative (FN) as red.

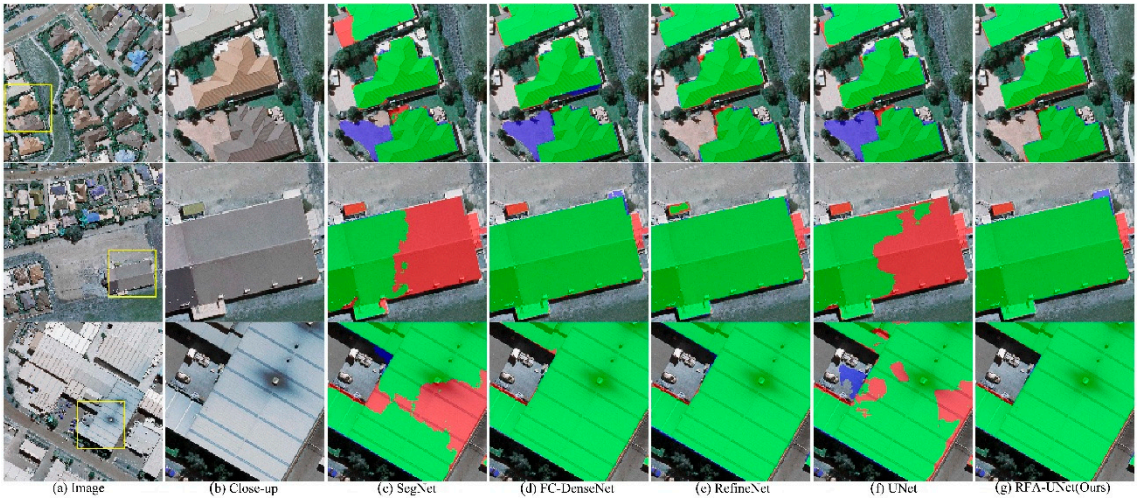



**Figure 13.** Comparison with the state-of-the-art attention methods on WHU building test sets. True positive (TP) is marked as green, false positive (FP), as blue and false negative (FN) as red.

According to Table 4, compared with abovementioned FCNs-based semantic segmentation models and recently published methods in the remote sensing fields, our method is better than them on the same datasets. On the test set of Mass. Buildings (spatial resolution of 1.0 m), our method surpasses second best model by 0.64% in terms of F1-score. For the test set with higher spatial resolution, the IoU score of our model was around 2.42% higher than that of the second best model RefineNet on the Potsdam (0.05 m), and 1.0% higher than that of pervious best model SRI-Net on the WHU (0.3 m).

**Table 4.** Quantitative comparison with other deep models (%) on three test sets, where values in bold are the best and the underlined values are the second best.

| Dataset | Methods [1] | Overall Accuracy | F1-Score | IoU |
|---|---|---|---|---|
| Mass. Buildings | UNet (baseline) | 93.33 | 81.69 | 69.04 |
| | SegNet | 92.84 | 80.25 | 67.00 |
| | RefineNet | 91.54 | 75.55 | 60.71 |
| | FC-DenseNet | 93.57 | 82.90 | 70.79 |
| | Deeplab-v3plus | 93.83 | 82.97 | <u>70.89</u> |
| | MFRN [66] | <u>94.51</u> | <u>85.01</u> | / |
| | RFA-UNet (Ours) | **94.71** | **85.65** | **74.91** |
| Potsdam | UNet (baseline) | 91.84 | 84.19 | 72.69 |
| | SegNet | 95.92 | 91.40 | 84.16 |
| | RefineNet | <u>97.10</u> | <u>94.00</u> | <u>88.69</u> |
| | FC-DenseNet | 94.33 | 87.73 | 78.14 |
| | DAN [52] | 96.16 | 92.56 | 86.71 |
| | RFA-UNet (Ours) | **97.79** | **95.35** | **91.11** |
| WHU | UNet (baseline) | 98.07 | 91.35 | 84.08 |
| | SegNet | 98.18 | 91.18 | 84.88 |
| | RefineNet | 98.64 | 93.87 | 88.45 |
| | FC-DenseNet | <u>98.44</u> | 93.11 | 87.10 |
| | Deeplab-V3+ [67] | / | 93.22 | 87.31 |
| | SRI-Net [67] | / | <u>94.23</u> | <u>89.09</u> |
| | RFA-UNet (Ours) | **98.84** | **94.75** | **90.02** |

[1] This table incorporates the results by ourselves and numerical results reported by other authors.

## 4. Discussion

Applying the attention mechanism to the segmentation model UNet, we observe that our joint attention module improves the performance of existing architecture for the task of building extraction in VHR images. The reason why the proposed attention improves the performance might be related to the inherent attributes of CNNs and the flaw of the plain skip connection in the encoder–decoder architectures. Generally, CNNs increase the receptive field by stacking convolution layers, which means the receptive field of a given layer only focus on a local region, especially at the shallow of the network. Therefore, the difference between deep layer and shallow layer in the use of context information leads to the variation of classification capacities. On the other hand, the spatial information of low level features is important to localize the classified objects, but these low level features also bring debatable noisy information that results in categorical errors [68]. In this paper, we rethink the relationship between shallow and corresponding deep layers in the skip connection at the feature level. In order to leverage the spatial information from shallow layers and the context information from deep layers, we employ the attention mechanism that highlights advantageous features and suppress features making less contribution. The channel-wise attention part of the proposed module applies global average pooling to the concatenated features, which extracts global categorical information of two input features. Two subsequent fully connected layers play an important role in capturing feature dependencies in the channel dimension. This way ensures the cross-layers information exchange. Thus, the rescaled low level output activated by sigmoid is more dynamically consistent with high level features. Furthermore, the spatial attention part uses additive attention to refine the low level features with the aid from the high level features that with larger receptive fields, which introduces more elaborate context to improve the classifying ability of the features.

Compared to other existing attention method, flexibility is an advantage of our proposed attention module. The experimental results on three different datasets demonstrate that RFA module can better deal with the task of building extraction with different sources of aerial images. Taking channel and spatial dimensions into account successively allows for a more robust interaction of context information between the feature layers in the segmentation model. Meanwhile, the residual mapping branch of

RFA alleviates the gradient vanishing in the training process. These are two reasons why the proposed RFA attention module outperforms other single attention methods in this study. With respect to DualAN that also uses two kinds of attentions in the comparison, our approach is quite different from it. In particular, DualAN applies attention mechanisms in parallel to the bottleneck of network, which focuses on employing self-attention to enhance representation of deep features, rather than reducing the semantic discrepancy between different level features. Moreover, because of the high cost of intermediate matrix multiplication in the DualAN, the authors [62] just place it for the bottleneck features with low spatial resolution. The experimental results imply this strategy is not effective enough for building extraction in the aerial images. However, our practice has shown that the proposed joint attention only increases small cost of additional model parameters (see Table 5, about 0.4 million) and computation (about 1.53 MB), even when applied at every level of the network. This flexibility implies the possibility of embedding RFA in other architectures in the future.

**Table 5.** Comparison with baseline.

| Methods | Parameters | Dataset | Training Time | Inference Time [1] |
|---|---|---|---|---|
| UNet (baseline) | 25.71 million | Mass. Buildings | ~2 h | 0 m 5 s |
| | | Potsdam | ~6 h | 3 m 4 s |
| | | WHU | ~11 h | 2 m 20 s |
| RFA-UNet (Ours) | 26.11 million | Mass. Buildings | ~2 h | 0 m 6 s |
| | | Potsdam | ~12 h | 6 m 15 s |
| | | WHU | ~8 h | 4 m 10 s |

[1] Time consumed by the method to predict the complete test set.

There is abundant room for further studies. First, the proposed RFA module does not validate the possible improvements it might bring on the other encoder–decoder models. At present, the reason we do not apply the RFA module to other models is that many factors need to be considered, such as the computational resource consumption of the models, the applicability of models themselves to different data sets and hyperparameter settings of models. The comparison with other methods in training time also means further hyperparameter optimization of the proposed module is possible (see Tables 5 and A1). Therefore, it is needed to provide a more comprehensive comparison of these methods in the future. Second, we have conducted the experiments on three datasets of urban buildings in the public domain (e.g., Mass. Buildings, Potsdam, and WHU). It is promising to develop the RFA applied models on multi-source data and rural residential buildings. Finally, we only focus on the task of building extraction in this paper. Since the proposed RFA-UNet can be easily transformed into a multi-class semantic segmentation models, we plan to extend our model with extra geometric constraints and to multiple classes.

## 5. Conclusions

In this paper, an end-to-end attention FCN model was proposed for building extraction in very high resolution aerial imagery. We have implemented a re-weighting technique based attention mechanism to adjust the response of features dynamically in channel-wise and spatial dimensions. With the aid of the context information from high level features, the proposed joint attention module can effectively enhance the semantic consistency of features across layers so as to improve the discrimination power of the UNet model for the building extraction task. Experiments on three different high resolution building datasets verified the effectiveness of attention mechanism, and the proposed RFA-UNet model achieved state-of-the-art performance on these popular benchmarks.

**Author Contributions:** All the authors made great contributions to this work. Funding acquisition, M.G., J.D., and K.W.; Investigation, M.G.; Methodology, Z.Y.; Resources, K.W.; Supervision, A.C. and K.W.; Validation, Y.F.; Writing—Original draft, Z.Y.; Writing—Review and editing, Y.F. and A.C.

**Funding:** This work was supported by the National Natural Science Foundation of China, grant no. 41701171 and grant no. 41971236; the Basic Public Welfare Research Program of Zhejiang Province, grant no. LGJ19D010001; and the Natural Science Foundation of Zhejiang Province, grant no. LY18G030006.

**Acknowledgments:** The authors are grateful for the help from the International Cooperation Regional Development Project between Zhejiang University and Leeds University. They would also like to thank anonymous reviewers and the editor for their constructive comments and suggestions.

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

## Appendix A

**Table A1.** Complexity comparison with other attention modules.

| Methods | Dataset | Training Time | Inference Time [1] |
|---|---|---|---|
| CAB | Mass. Buildings | ~3 h | 0 m 6 s |
| | Potsdam | ~7 h | 2 m 40 s |
| | WHU | ~13 h | 4 m 35 s |
| GRID | Mass. Buildings | ~2 h | 0 m 5 s |
| | Potsdam | ~8 h | 2 m 14 s |
| | WHU | ~12 h | 3 m 49 s |
| DenseAN | Mass. Buildings | ~2 h | 0 m 5 s |
| | Potsdam | ~10 h | 6 m 10 s |
| | WHU | ~13 h | 3 m 45 s |
| DualAN | Mass. Buildings | ~4 h | 0 m 6 s |
| | Potsdam | ~19 h | 6 m 19 s |
| | WHU | ~15 h | 3 m 12 s |
| RFA(Ours) | Mass. Buildings | ~2 h | 0 m 6 s |
| | Potsdam | ~12 h | 6 m 15 s |
| | WHU | ~8 h | 4 m 10 s |

[1] the time consumed by the method to predict the complete test set.

## References

1. Tong, X.; Lin, X.; Feng, T.; Xie, H.; Liu, S.; Hong, Z.; Chen, P. Use of shadows for detection of earthquake-induced collapsed buildings in high-resolution satellite imagery. *ISPRS J. Photogramm. Remote Sens.* **2013**, *79*, 53–67. [CrossRef]
2. Tuia, D.; Volpi, M.; Moser, G. Decision Fusion with Multiple Spatial Supports by Conditional Random Fields. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 3277–3289. [CrossRef]
3. Zhang, C.; Pan, X.; Li, H.; Gardiner, A.; Sargent, I.; Hare, J.; Atkinson, P.M. A hybrid MLP-CNN classifier for very fine resolution remotely sensed image classification. *ISPRS J. Photogramm. Remote Sens.* **2018**, *140*, 133–144. [CrossRef]
4. LeCun, Y.; Boser, B.E.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.E.; Jackel, L.D. Handwritten Digit Recognition with a Back-Propagation Network. *Adv. Neural Inf. Process. Syst.* **1990**, *2*, 396–404.
5. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
6. Fu, G.; Liu, C.; Zhou, R.; Sun, T.; Zhang, Q. Classification for High Resolution Remote Sensing Imagery Using a Fully Convolutional Network. *Remote Sens.* **2017**, *9*, 498. [CrossRef]
7. Hamaguchi, R.; Fujita, A.; Nemoto, K.; Imaizumi, T.; Hikosaka, S. Effective Use of Dilated Convolutions for Segmenting Small Object Instances in Remote Sensing Imagery. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1442–1450.
8. Zhao, W.; Du, S.; Wang, Q.; Emery, W.J. Contextually guided very-high-resolution imagery classification with semantic segments. *ISPRS J. Photogramm. Remote Sens.* **2017**, *132*, 48–60. [CrossRef]

9. Mnih, V. Machine Learning for Aerial Image Labeling. Ph.D. Thesis, University of Toronto, Toronto, ON, Canada, 2013.
10. Tao, Y.; Xu, M.; Lu, Z.; Zhong, Y. DenseNet-Based Depth-Width Double Reinforced Deep Learning Neural Network for High-Resolution Remote Sensing Image Per-Pixel Classification. *Remote Sens.* **2018**, *10*, 779. [CrossRef]
11. Xu, X.; Li, W.; Ran, Q.; Du, Q.; Gao, L.; Zhang, B. Multisource Remote Sensing Data Classification Based on Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 937–949. [CrossRef]
12. Zhao, W.; Du, S. Learning multiscale and deep representations for classifying remotely sensed imagery. *ISPRS J. Photogramm. Remote Sens.* **2016**, *113*, 155–165. [CrossRef]
13. Paisitkriangkrai, S.; Sherrah, J.; Janney, P.; Hengel, A. van den Semantic Labeling of Aerial and Satellite Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 2868–2881. [CrossRef]
14. Alshehhi, R.; Marpu, P.R.; Woon, W.L.; Mura, M.D. Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2017**, *130*, 139–149. [CrossRef]
15. Zhang, C.; Sargent, I.; Pan, X.; Li, H.; Gardiner, A.; Hare, J.; Atkinson, P.M. An object-based convolutional neural network (OCNN) for urban land use classification. *Remote Sens. Environ.* **2018**, *216*, 57–70. [CrossRef]
16. 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.
17. 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 (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
18. Kamnitsas, K.; Bai, W.; Ferrante, E.; McDonagh, S.; Sinclair, M.; Pawlowski, N.; Rajchl, M.; Lee, M.; Kainz, B.; Rueckert, D.; et al. Ensembles of Multiple Models and Architectures for Robust Brain Tumour Segmentation. In Proceedings of the Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Granada, Spain, 16 September 2018; Crimi, A., Bakas, S., Kuijf, H., Menze, B., Reyes, M., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 450–462.
19. Li, Y.; He, B.; Long, T.; Bai, X. Evaluation the performance of fully convolutional networks for building extraction compared with shallow models. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 850–853.
20. Wu, G.; Shao, X.; Guo, Z.; Chen, Q.; Yuan, W.; Shi, X.; Xu, Y.; Shibasaki, R. Automatic Building Segmentation of Aerial Imagery Using Multi-Constraint Fully Convolutional Networks. *Remote Sens.* **2018**, *10*, 407. [CrossRef]
21. Chen, L.C.; Barron, J.T.; Papandreou, G.; Murphy, K.; Yuille, A.L. Semantic Image Segmentation with Task-Specific Edge Detection Using CNNs and a Discriminatively Trained Domain Transform. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4545–4554.
22. Cheng, D.; Meng, G.; Cheng, G.; Pan, C. SeNet: Structured Edge Network for Sea-Land Segmentation. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 247–251. [CrossRef]
23. Kampffmeyer, M.; Salberg, A.B.; Jenssen, R. Semantic Segmentation of Small Objects and Modeling of Uncertainty in Urban Remote Sensing Images Using Deep Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 680–688.
24. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. High-Resolution Aerial Image Labeling With Convolutional Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 7092–7103. [CrossRef]
25. 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]
26. Sun, Y.; Zhang, X.; Xin, Q.; Huang, J. Developing a multi-filter convolutional neural network for semantic segmentation using high-resolution aerial imagery and LiDAR data. *ISPRS J. Photogramm. Remote Sens.* **2018**, *143*, 3–14. [CrossRef]
27. Sun, Y.; Zhang, X.; Zhao, X.; Xin, Q. Extracting Building Boundaries from High Resolution Optical Images and LiDAR Data by Integrating the Convolutional Neural Network and the Active Contour Model. *Remote Sens.* **2018**, *10*, 1459. [CrossRef]

28. Wu, G.; Guo, Z.; Shi, X.; Chen, Q.; Xu, Y.; Shibasaki, R.; Shao, X. A Boundary Regulated Network for Accurate Roof Segmentation and Outline Extraction. *Remote Sens.* **2018**, *10*, 1195. [CrossRef]
29. Xu, Y.; Wu, L.; Xie, Z.; Chen, Z. Building Extraction in Very High Resolution Remote Sensing Imagery Using Deep Learning and Guided Filters. *Remote Sens.* **2018**, *10*, 144. [CrossRef]
30. Chen, G.; Zhang, X.; Wang, Q.; Dai, F.; Gong, Y.; Zhu, K. Symmetrical Dense-Shortcut Deep Fully Convolutional Networks for Semantic Segmentation of Very-High-Resolution Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 1633–1644. [CrossRef]
31. Volpi, M.; Tuia, D. Dense Semantic Labeling of Subdecimeter Resolution Images with Convolutional Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 881–893. [CrossRef]
32. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Munich, Germany, 5–9 October 2015; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
33. Chen, L.; Yang, Y.; Wang, J.; Xu, W.; Yuille, A.L. Attention to Scale: Scale-Aware Semantic Image Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3640–3649.
34. 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]
35. Islam, M.A.; Rochan, M.; Bruce, N.D.B.; Wang, Y. Gated Feedback Refinement Network for Dense Image Labeling. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4877–4885.
36. Huang, J.; Zhang, X.; Xin, Q.; Sun, Y.; Zhang, P. Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network. *ISPRS J. Photogramm. Remote Sens.* **2019**, *151*, 91–105. [CrossRef]
37. Zhou, Z.; Rahman Siddiquee, M.M.; 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*; Stoyanov, D., Taylor, Z., Carneiro, G., Syeda-Mahmood, T., Martel, A., Maier-Hein, L., Tavares, J.M.R.S., Bradley, A., Papa, J.P., Belagiannis, V., et al., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 3–11.
38. Jetley, S.; Lord, N.A.; Lee, N.; Torr, P.H.S. Learn to Pay Attention. In Proceedings of the Sixth International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
39. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T. SCA-CNN: Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6298–6306.
40. Harley, A.W.; Derpanis, K.G.; Kokkinos, I. Segmentation-Aware Convolutional Networks Using Local Attention Masks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5038–5047.
41. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual Attention Network for Image Classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3156–3164.
42. Schlemper, J.; Oktay, O.; Schaap, M.; Heinrich, M.; Kainz, B.; Glocker, B.; Rueckert, D. Attention gated networks: Learning to leverage salient regions in medical images. *Med Image Anal.* **2019**, *53*, 197–207. [CrossRef]
43. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning Where to Look for the Pancreas. *arXiv* **2018**, arXiv:1804.03999.
44. Zagoruyko, S.; Komodakis, N. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. *arXiv* **2016**, arXiv:1612.03928.
45. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**. [CrossRef]

46. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. *arXiv* **2018**, arXiv:1807.06521.
47. Li, H.; Xiong, P.; An, J.; Wang, L. Pyramid Attention Network for Semantic Segmentation. *arXiv* **2018**, arXiv:1805.10180.
48. 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/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–20 June 2018; pp. 1857–1866.
49. Cheng, W.; Yang, W.; Wang, M.; Wang, G.; Chen, J. Context Aggregation Network for Semantic Labeling in Aerial Images. *Remote Sens.* **2019**, *11*, 1158. [CrossRef]
50. Pan, X.; Yang, F.; Gao, L.; Chen, Z.; Zhang, B.; Fan, H.; Ren, J. Building Extraction from High-Resolution Aerial Imagery Using a Generative Adversarial Network with Spatial and Channel Attention Mechanisms. *Remote Sens.* **2019**, *11*, 917. [CrossRef]
51. Xu, R.; Tao, Y.; Lu, Z.; Zhong, Y. Attention-Mechanism-Containing Neural Networks for High-Resolution Remote Sensing Image Classification. *Remote Sens.* **2018**, *10*, 1602. [CrossRef]
52. Yang, H.; Wu, P.; Yao, X.; Wu, Y.; Wang, B.; Xu, Y. Building Extraction in Very High Resolution Imagery by Dense-Attention Networks. *Remote Sens.* **2018**, *10*, 1768. [CrossRef]
53. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
54. Baby, M.; Jereesh, A.S. Automatic nerve segmentation of ultrasound images. In Proceedings of the International conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 20–22 April 2017; pp. 107–112.
55. Iglovikov, V.; Mushinskiy, S.; Osin, V. Satellite Imagery Feature Detection using Deep Convolutional Neural Network: A Kaggle Competition. *arXiv* **2017**, arXiv:1706.06169.
56. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity Mappings in Deep Residual Networks. In Proceedings of the Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 630–645.
57. Tobler, W.R. A Computer Movie Simulating Urban Growth in the Detroit Region. *Econ. Geogr.* **1970**, *46*, 234. [CrossRef]
58. ISPRS. Available online: http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html (accessed on 10 December 2019).
59. Ji, S.; Wei, S.; Lu, M. Fully Convolutional Networks for Multisource Building Extraction from an Open Aerial and Satellite Imagery Data Set. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 574–586. [CrossRef]
60. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Sardinia, Italy, 13–15 May 2010; pp. 249–256.
61. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
62. 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 (CVPR), Long Beach, CA, USA, 15–21 June 2019; pp. 3146–3154.
63. 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]
64. 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 (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5168–5177.
65. Jégou, S.; Drozdzal, M.; Vazquez, D.; Romero, A.; Bengio, Y. The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 1175–1183.
66. Li, L.; Liang, J.; Weng, M.; Zhu, H.; Li, L.; Liang, J.; Weng, M.; Zhu, H. A Multiple-Feature Reuse Network to Extract Buildings from Remote Sensing Imagery. *Remote Sens.* **2018**, *10*, 1350. [CrossRef]

67. Liu, P.; Liu, X.; Liu, M.; Shi, Q.; Yang, J.; Xu, X.; Zhang, Y. Building Footprint Extraction from High-Resolution Images via Spatial Residual Inception Convolutional Neural Network. *Remote Sens.* **2019**, *11*, 830. [CrossRef]
68. Ding, H.; Jiang, X.; Shuai, B.; Liu, A.Q.; Wang, G. Semantic Correlation Promoted Shape-Variant Context for Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–21 June 2019.