


## Article

# EANet: Edge-Aware Network for the Extraction of Buildings from Aerial Images

Guang Yang<sup>1,2</sup>, Qian Zhang<sup>1,2,\*</sup>  and Guixu Zhang<sup>1,2</sup>

<sup>1</sup> The Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai 200241, China; 51194506041@stu.ecnu.edu.cn (G.Y.); gxzhang@cs.ecnu.edu.cn (G.Z.)

<sup>2</sup> School of Computer Science and Technology, East China Normal University, Shanghai 200062, China

\* Correspondence: qzhang@cs.ecnu.edu.cn; Tel.: +86-152-2105-4245

Received: 16 June 2020; Accepted: 4 July 2020; Published: 6 July 2020



**Abstract:** Deep learning methods have been used to extract buildings from remote sensing images and have achieved state-of-the-art performance. Most previous work has emphasized the multi-scale fusion of features or the enhancement of more receptive fields to achieve global features rather than focusing on low-level details such as the edges. In this work, we propose a novel end-to-end edge-aware network, the EANet, and an edge-aware loss for getting accurate buildings from aerial images. Specifically, the architecture is composed of image segmentation networks and edge perception networks that, respectively, take charge of building prediction and edge investigation. The International Society for Photogrammetry and Remote Sensing (ISPRS) Potsdam segmentation benchmark and the Wuhan University (WHU) building benchmark were used to evaluate our approach, which, respectively, was found to achieve 90.19% and 93.33% intersection-over-union and top performance without using additional datasets, data augmentation, and post-processing. The EANet is effective in extracting buildings from aerial images, which shows that the quality of image segmentation can be improved by focusing on edge details.

**Keywords:** semantic segmentation; convolutional neural networks; building extraction; edge; multi-task learning

## 1. Introduction

As aerial and satellite remote sensing images have become convenient information sources, extracting various artificial features from image information has become a research hotspot. Buildings, as one of the main artificial features in a city, have special significance in the automatic extraction of urban areas, map updating, urban change detection, urban planning, building energy consumption assessment, and infrastructure construction. The acquisition of buildings from remote sensing images has evolved into a mature research field after decades of development [1–4]. However, many challenges persist in this domain. First, given the building materials and their very close proximity to roads, buildings are easily confused with other elements. Second, the structure and spectrum of buildings are complex and diverse. Moreover, the considerable variance that occurs within this class makes buildings difficult to identify. Third, image structure is easily affected by the shadows of buildings and trees. Thus, automatically extracting buildings from remote sensing images is challenging.

Therefore, improving the recognition ability of feature representation in pixel-level recognition is necessary to extract buildings. Early image segmentation adopted a traditional image processing method based on artificial design characteristics, such as spectral information [5], tone [6], texture [7], and geometric shape [8,9]. At present, image processing technologies based on convolutional neural networks (CNNs) have obtained remarkable development, and their accuracy and even efficiency far exceed those of traditional methods. Remote sensing image processing methods according

to artificial intelligence (AI) and machine learning (ML) provide opportunities and possibilities for the automatic acquisition of buildings from remote sensing images. Particularly key to these technologies is the development of image classification [10–13], object detection [14–16], and semantic segmentation [17–19], as represented by deep learning. Semantic segmentation for the image is the purpose of each pixel in the allocation of a single category, and it can be seen as a dense classification problem. Most object parsing issues in image segmentation can be regarded as semantic segmentation. The depth of the continuous development of CNNs has helped many remarkable achievements in the semantic segmentation field. Long et al. [18] extended the original convolutional neural network structure and proposed an end-to-end full convolutional neural network (FCN). Their FCN could achieve the intensive prediction of images without the use of fully connected layers. This kind of network is an encoder–decoder structure and enables the segmentation network to generate images of any size, which improves processing efficiency compared with the traditional image block classification method. Since then, almost all research on semantic segmentation has adopted the FCN structure [19–21].

Many improved methods of FCN are employed to extract ground objects from remote sensing images. Maggiori et al. [10] used a multi-scale structure to improve FCN to alleviate the tradeoff between increasing contexts and increasing the number of parameters. Mou et al. [19] proposed a method that combined FCN and a recurrent neural network that used the most superficial boundary perception feature map to achieve accurate object boundary inference and semantic segmentation. Marmanis et al. [20] adopted a parallel processing chain with two identical structures to improve FCN through delayed fusion with the help of a network layer in an early active deconvolution and cyclic feature graph. Xu et al. [21] applied manual characteristics and the guided filtering technique to optimize building extraction with the res-u-net network they proposed.

Compared with the traditional manual design feature model, the semantic segmentation model based on a CNN has been significantly improved. However, in this model, the CNN uses the pooling layer many times to increase the receptive field. The use of down-sampling to compress data is irreversible, resulting in information loss and thereby causing translation invariance and smooth results. At the same time, this development leads to an inaccurate image contour generated by the convolutional network and an unclear boundary. On the basis of the strong recognition ability of CNNs, Chen et al. [22] used fully connected conditional random fields (CRFs) for post-processing, which improved the quality of the object boundary in their segmentation results.

However, image segmentation still requires the precise position information of each pixel. Consequently, such a segmentation is inapplicable to pooling layers or striding convolution as boldly as a classification task to reduce computation. A mainstream method involves utilizing an encoder–decoder structure network [18,22–25]. The encoder reduces the resolution of the input image through down-sampling to generate a feature map with low resolution. Then, the decoder performs upper sampling on these feature descriptions to restore the segmentation graph with full resolution, thereby significantly reducing the necessary calculation by decreasing the size of the feature map. Zeiler et al. [26] proposed deconvolution for the first time for the reconstruction of feature maps to help them recover their original size. Tian et al. [27] believed that the use of bilinear interpolation up-sampling to restore the resolution of a feature map might lead to an unsatisfactory segmentation result, so they designed a data-dependent up-sampling method called DUpsampling to replace bilinear interpolation.

In addition to computation, the multi-scaling of objects is also a challenge for CNNs. The extraction of any target feature is conducted on a certain scale, and different scales produce dissimilar results. To enlarge the receptive field of a feature map, Yu et al. [28] proposed atrous convolution, an approach that can increase the receptive field without pooling operation. It allows each convolution operation to extract a wider range of information. Conversely, the pyramid pooling module developed by Zhao et al. [29] can maximize the global feature hierarchy's prior knowledge to understand different scenarios and aggregate the context information of various regions so as to give the feature map more

semantic information using more global information. Chen et al. [25] recommended atrous spatial pyramid pooling, which combines atrous convolution and spatial pyramid pooling. Using their scheme allows for the re-sampling of the convolutional features extracted from a single scale and the accurate and effective classification of regions at any scale. Liu et al. [30] added low-level features to adjacent upper floors and combined them into new features. At the same time, each layer of feature maps could be predicted separately to realize detection at different scales.

In the recent semantic segmentation community, researchers have also shifted their attention toward the improvement of multi-task learning in addition to the enhancement of the encoder–decoder structure and multi-scaling. Scholars have increasingly begun to pay attention to multi-task learning. That learning entails the simultaneous learning of multiple related tasks and facilitates the sharing of the learned information among tasks. Dai et al. [31] reduced the risk of model over-fitting and improved the accuracy of the results by learning the three tasks of mask estimation, instance differentiation, and object classification. Zhang et al. [32] used head posture estimation and facial attribute inference as auxiliary tasks for improving the effectiveness of facial key-point detection.

The above studies not only developed various CNN models for semantic segmentation but also provided numerous ideas for our research. In addition, Zeiler et al. [33] proved that a feature map gains more semantic information and loses more detailed information as the layers of its neural network deepen. However, as a pixel-level prediction task, semantic segmentation requires detailed and accurate information. As part of detailed information, semantic segmentation is widely concerned with the edge in to boost the performance of neural networks. Yu et al. [34] obtained as many features as possible through inter-class differences to enhance semantic segmentation performance under precise boundary supervision. Qin et al. [35] implicitly injected precise boundary prediction targets into mixing loss to reduce false errors from the cross-propagation of information learned in other areas of the boundary and image.

Inspired by multi-task learning and edge information utilization, we hereby designed a new edge perception network according to edge information supervision to automatically extract buildings from high-resolution aerial images. The main contributions of this work are as follows.

1. The EANet, a multi-task learning network based on the encoder–decoder structure, is proposed to automatically extract buildings from high-resolution aerial images. Our proposed network EANet was trained end-to-end through a series of loss functions. The EANet consists of an image segmentation network and an edge perception network. The former predicts the segmentation of images using remote sensing images, and the latter aims to supervise the segmentation network and further enhance the accuracy of edge prediction.
2. EALoss, a new loss function, is proposed to refine the prediction results of the segmentation network and was designed to supervise the learning process of accurate prediction for the binary boundary segmentation of images.
3. Without using additional datasets, data augmentation, and post-processing, the EANet achieves top performance on two remote sensing image semantic segmentation datasets, i.e., The International Society for Photogrammetry and Remote Sensing (ISPRS) Potsdam [36] and Wuhan University (WHU) building datasets [37].

The rest of this article is organized as follows. Section 2 introduces the composition of the model in detail. Section 3 describes the experimental dataset, model evaluation methods, experimental design, and the analysis of the experimental results. Section 4 discusses the effectiveness of the EANet and future work. Section 5 summarizes the paper.

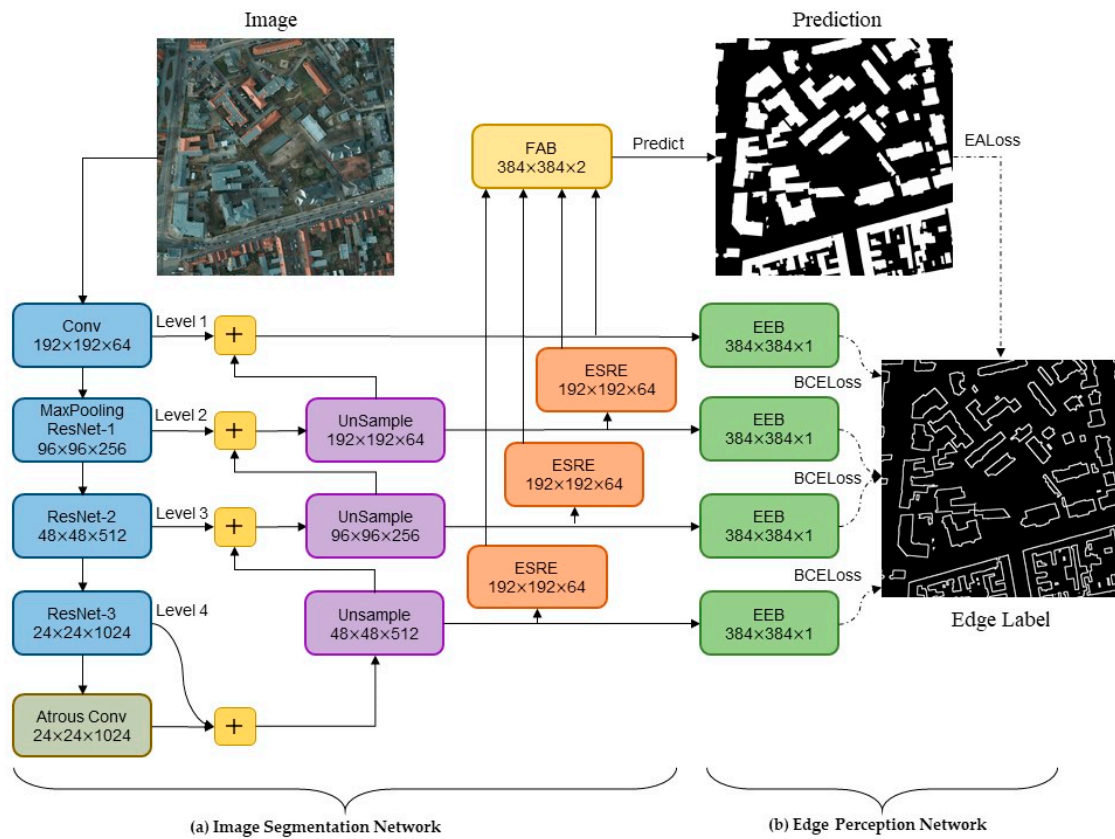
## 2. Methods

The main purpose of this article is to explore a means to overcome the incompleteness of detail information that is extracted automatically from remote sensing images. The key idea in this article involves using the edge to guide the neural network to pay greater attention to the edge and use

edge information to guide the network to refine the segmentation results. This section begins with an overview of network architecture. The image segmentation network is first illustrated in Section 2.2. Then, details of the newly designed edge perception network and EALoss are provided in Section 2.3.

### 2.1. Overview of Network Architecture

The overall structure of the EANet is shown in Figure 1. The EANet consists of two blocks—the image segmentation and the edge perception networks that are, respectively, used for image semantic segmentation and edge supervision. The EANet was developed on the basis of a rough-to-fine encoding–decoding structure. The edge perception network learns the residuals between the edge labels and the predicted edges through the supervision of the loss function to further refine the feature map of the segmentation network.

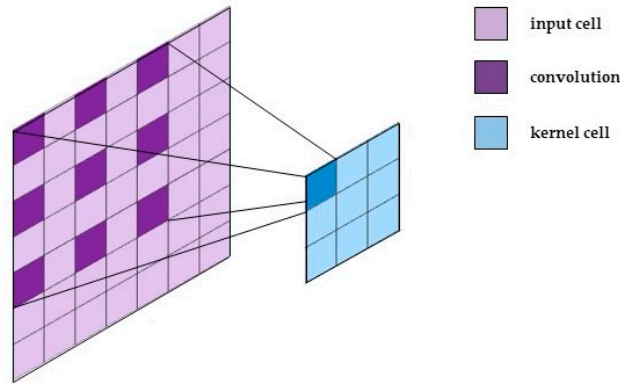


**Figure 1.** Overall architecture of our approach. Solid lines represent the direction of information flow. Dashed lines indicate the edge of supervision. The output dimensions of each block ( $H \times W \times C$ ) are indicated in the boxes. ESRE: explicit spatial resolution embedding; EEB: edge extraction branch; FAB: feature aggregation block.

### 2.2. The Image Segmentation Network

Inspired by FPN [38] and U-Net [23], the image segmentation network was designed herein as an encoding–decoding structure with residual connection, because this structure can simultaneously extract high-level global semantic information and low-level detail information on the basis of reducing computation. In line with the choices in previous works [39–43], we also used ResNet [13] as the backbone of this work. Finally, the last full connection layer of ResNet was replaced with atrous convolution [24–28,44] to reduce the loss of detail due to pooling operations. After each convolutional layer, a ReLU function [45] is used as the activation function, and batch normalization [46] is added at the same time. Figure 2 depicts a concrete atrous convolution operation. The atrous convolution

maintains the relative spatial position of the feature map and can improve the receptive field without decreasing the spatial resolution, making it possible to aggregate a wider range of information after convolution operation. The other two methods we propose are as follows.



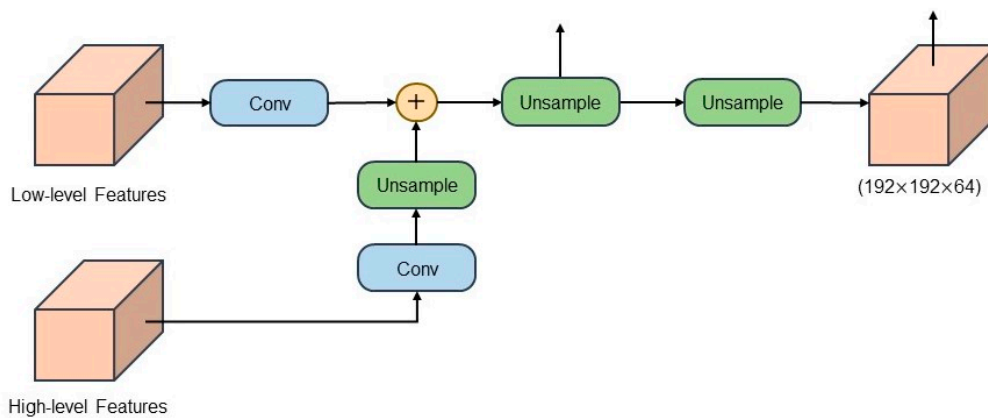
**Figure 2.** The atrous convolution operation with pooling  $p = 2$  and stride  $s = 1$ . Pooling is applied but is not indicated in this figure. The dimension of the input feature map is  $7 \times 7$ , of the kernel is  $3 \times 3$ , and of the output feature map through atrous convolution becomes  $7 \times 7$ .

### 2.2.1. Explicit Spatial Resolution Embedding

As mentioned, many semantic segmentation networks like FPN [38] and FRRN [39] adopt the feature fusion method of residual connection. This common form of residual connection is formulated as:

$$\mathbf{y}_l = \mathcal{F}(\mathbf{x}_l) + \text{Unsample}(\mathbf{x}_{l+1}) \quad (1)$$

where  $\mathbf{y}_l$  is the feature map obtained by the decoder fusion at  $l$ -th level,  $\mathbf{x}_l$  stands for the feature map obtained by the encoder at  $l$ -th level, and  $\mathbf{x}_{l+1}$  denotes the higher level of the feature map generated by the encoder. Features have more semantic information as large  $l$ , but their spatial resolution decreases and vice versa (see Figure 1). Our framework utilizes the residual connection method and further improves the feature extraction method as the explicit spatial resolution embedding (ESRE) module (Figure 3). We applied this component in Levels 2–4 to extract more detailed features at different scales. The features at the high level provide less spatial information, so a natural motivation is to extract more low-level details from the high-level feature space to help model inference. The ESRE module can be used to obtain detailed spatial information from various scales.

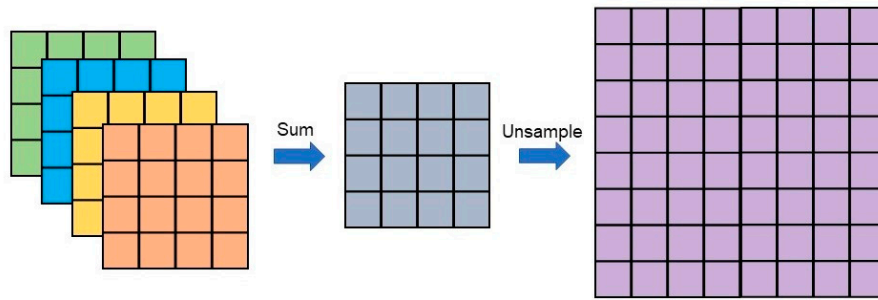


**Figure 3.** Details of the explicit spatial resolution embedding (ESRE) component in Figure 1, where “ $\oplus$ ” denotes an element-wise sum.



### 2.2.2. Feature Aggregation Block (FAB)

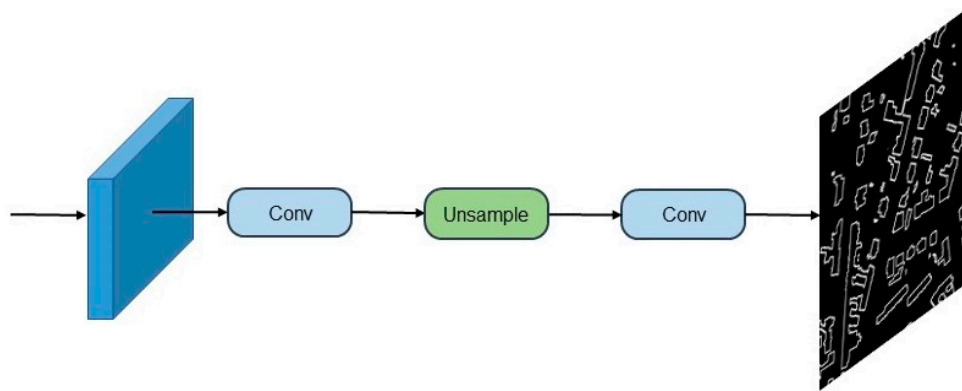
Feature fusion is of great help to the promotion of targets at different scales. As different levels have dissimilar semantic information, the spatial information extracted from varied levels can be fused to highlight the detection targets. As shown in Figure 4, we designed the feature aggregation block (FAB) module for feature fusion after extracting additional spatial information on different scales. The spatial information extracted from Levels 2–4 after passing through the ESRE module and from Level 1 is fused by the FAB module. Finally, the prediction results are obtained through up-sampling.



**Figure 4.** A description of the feature aggregation block (FAB) component in our pipeline.

### 2.3. The Edge Perception Network

As shown in Figure 5, we designed an edge perception network to guide the semantic segmentation network to learn more edge information of images by using an edge extraction branch (EEB) module, which refines image prediction results. Inspired by [47–49], we believe that learning the edge information directly from the feature map is helpful for the segmentation task. Edges belong to low-level details and are considered important spatial information. The full use of edge information in the network can make up for the loss of spatial information caused by down-sampling to a certain extent.



**Figure 5.** Design of the edge extraction branch (EEB) module in our framework.

Identifying the edge label directly from the ground truth of the image is straightforward. We used a morphological erosion operation [50] to obtain a reliable edge label. Let  $S$  be the set formed by the elements of the ground truth and  $K_n$  be a kernel whose elements are all 1 with shape  $n$ . The erosion of  $S$  by the kernel  $K_n$  is denoted by  $\phi_{K_n}(S)$  and defined as follows:

$$\phi_{K_n}(S) = \{s | K_n + s \subseteq S\} \quad (2)$$

Using the erosion result of the ground truth obtained by Equation (2), we could conveniently ascertain the edge label of the ground truth. The edge label is produced by subtracting the erosion result, defined as  $\eta(S)$ , from the ground truth. That is:

$$\eta(S) = S - \phi_{K_n}(S) \quad (3)$$

Edge prediction results and true labels are, respectively, obtained by the EEB module and morphological operation. As shown in Figure 1, each level has a side-output that generates an edge prediction through the EEB module. We used binary cross entropy (BCE) [51] as a loss function to evaluate the extraction of the edge. The BCE loss function is one of the loss functions commonly used in semantic segmentation or a binary classification task. It is also used in our network. That loss is defined as:

$$\ell_{bce}^{(k)} = - \sum_{(x,y)} [\eta(x,y) \log(P^{(k)}(x,y)) + (1 - \eta(x,y)) \log(1 - P^{(k)}(x,y))] \quad (4)$$

where  $\eta(x,y) \in \{0,1\}$  is the edge label of the pixel  $(x,y)$  and  $P^{(k)}(x,y) \in [0,1]$  is the edge prediction probability obtained by the EEM module from the feature map at the  $k$ -th layer. For accurate segmentation results and clear edges, we propose a new edge loss function, EALoss, which is expressed as:

$$\ell_{ea} = 1 - \frac{\sum_{x=1}^H \sum_{y=1}^W \eta(x,y) P(x,y)}{\sum_{x=1}^H \sum_{y=1}^W \eta(x,y)} \quad (5)$$

where  $\eta(x,y) \in \{0,1\}$  is the edge label of the pixel  $(x,y)$ .  $P(x,y) \in [0,1]$  predicts the probability that the pixel  $(x,y)$  is the edge, as obtained by the calculation result of the FAB module (Figure 1) through the Softmax function. As the most commonly used loss function in deep neural networks, the cross-entropy (CE) loss is also used to constrain the final semantic segmentation result, which is defined as:

$$\ell_{ce} = - \sum_{x=1}^H \sum_{y=1}^W [\eta_0(x,y) \log \frac{e^{a_0}}{e^{a_0} + e^{a_1}} + \eta_1(x,y) \log \frac{e^{a_1}}{e^{a_0} + e^{a_1}}] \quad (6)$$

where  $a_i$  is the probability of belonging to category  $i \in \{0,1\}$  at pixel  $(x,y)$ . During training, we defined the loss function as the sum of BCE, EALoss, and CE:

$$\mathcal{L} = \sum_{k=1}^K \ell_{bce}^{(k)} + \ell_{ea} + \ell_{ce} \quad (7)$$

As described in Figure 1, our edge perception network is supervised with four side-outputs, i.e.,  $K = 4$ .

### 3. Experiment

Our model was implemented using PyTorch 1.3.0 [52] and Cuda 10.1 with ResNet101 [13] pre-trained from ImageNet [53] as the backbone. To test the validity and the correctness of our model, we constructed an experiment of the automatic acquisition of buildings on two remote sensing image datasets and compared the outcomes against those of the CNN model developed by other researchers. This section introduces the basic situation of the two datasets, describes the experimental setup, presents the evaluation standard, and illustrates the experimental results.

#### 3.1. Datasets

The ISPRS Potsdam dataset [36] is a state-of-the-art airborne image dataset released in 2018. The said dataset was designed for urban classification, 3D reconstruction, and the semantic annotation testing tasks of high-resolution remote sensing images. The dataset reflects the large building volume, narrow streets, and dense settlements of Potsdam City, Germany. Furthermore, ISPRS Potsdam also

includes water bodies, tennis courts, swimming pools, and other semantic objects that are usually overlooked in urban scenes. Those complex scenes require high algorithm robustness. The dataset contains 38 patches (6000 × 6000 pixel RGB images, for which each channel has a spectral resolution of 8 bit and the spatial resolution is 5 cm). We selected 14 images as the test set (IDs: 2\_13, 2\_14, 3\_13, 3\_14, 4\_13, 4\_15, 5\_13, 5\_14, 5\_15, 6\_14, 6\_14, 6\_15, and 7\_13). The validation set contained 1 image (ID: 2\_10), and the rest of the images were used as training sets.

The WHU building dataset [37] was specifically designed for building extraction with high-resolution remote sensing images. The dataset was collected by aircraft over 450 m<sup>2</sup> of Christchurch, New Zealand with a resolution of 0.075, and it contains more than 220,000 individual buildings. The aerial image involved down-sampling from a 0.075 to 0.3 m ground resolution. The whole dataset was cropped into 8189 pieces of 512 × 512 pixels without overlapping. All masks were manually annotated. In addition to the difference of satellite sensors, the change of atmospheric conditions, the corrections from the atmosphere and radiation, and the change of the seasons also increase the requirements of the samples in regard to the robustness of the building extraction algorithm. We used 4736 images as the training set, 2416 images as the test set, and the remaining 1032 images as the verification set.

### 3.2. Experimental Setup

As the images in the Potsdam dataset are too large to fit into memory, each image was cropped to 384 × 384 with a 25% overlap, with padding added beyond the border. In our experiment, we did not use any data augmentation or any training tricks [54] for all models, such as a warm-up [55] or label smoothing [56]. To demonstrate the effectiveness of the proposed algorithm, we compared it with eleven state-of-the-art models, including the PSPNet [29], U-Net [23], FCN [18], Deeplabv3-plus [24], RefineNet [57], CGNet [58], BiSeNet [40], SegNet [59], SiU-Net [37], SRINet [60], and DE-Net [17]. SegNet, FCN, and U-Net are the most representative models of deep learning in semantic segmentation. PSPNet, Deeplabv3-Plus, CGnet, BiSeNet, and RefineNet are recently proposed state-of-the-art approaches that have achieved excellent performance on natural image datasets. Similar to the method proposed by us, SIU-NET, SRINet, and De-Net also use a boundary-aware approach to extract objects.

For fairness, an open source code was used, and all the backbones of the models used ResNet-101, which is pre-trained on ImageNet [53]. For the training of all models, the stochastic gradient descent [61] optimizer was adopted, starting with a learning rate of 0.001, a weight decay of 0.0005, and a momentum of 0.9. A polynomial learning rate decay strategy was also used, and the learning rate was updated after each iteration by the decay factor  $(1 - \frac{\text{iteration}}{\text{max\_iteration}})^{0.9}$ . As mentioned, the parameters of the encoder were initialized from ResNet-101, and all convolutional layer parameters in the remaining decoder were initialized by Kaiming initialization [62]. The validation set was not employed during training. All experiments were trained on 2 GeForce RTX 2080Ti GPUs for 200 epochs with eight clips in a mini-batch (a mini-batch has a total size of 16 clips).

### 3.3. Evaluation Metrics

To correctly evaluate the performance of the model, four commonly used indicators in traditional semantic segmentation tasks were adopted: precision, recall, the F1 score (F1), and the intersection-over-union (IoU) ratio. Precision is the percentage of all retrieved results that are correctly retrieved. Recall is the proportion of correctly retrieved results that should be detected. The F1 score is a commonly used index in machine learning and is the harmonic mean of precision and recall. The IoU represents the ratio of the intersection and union of the prediction and ground truth. The formulas are as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (8)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (9)$$



$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

$$IoU = \frac{TP}{TP + FP + FN}, \quad (11)$$

where  $TP$ ,  $TN$ ,  $FP$ , and  $FN$  represent the true positive, true negative, false positive, and false negative pixels in the prediction, respectively.

### 3.4. Result

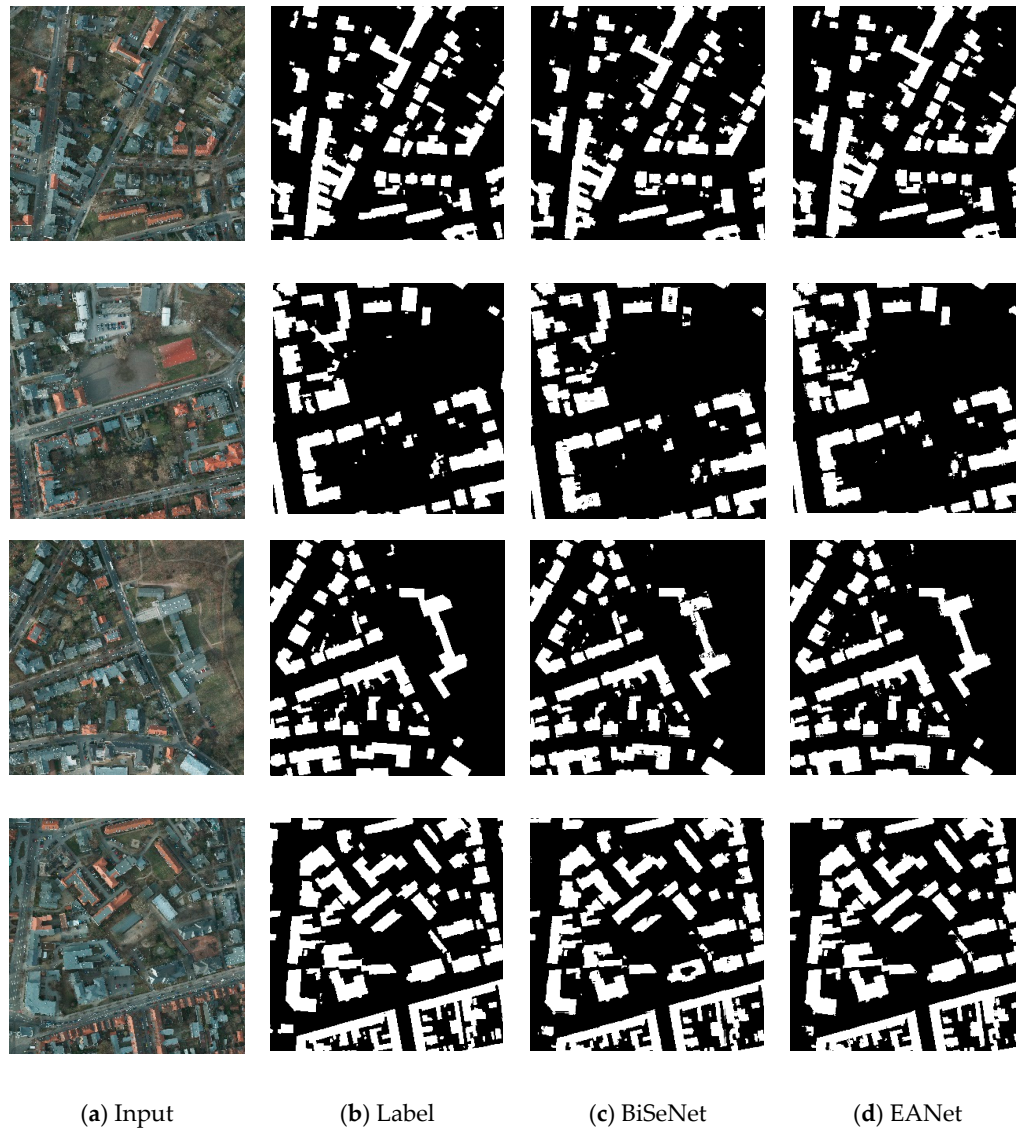
To assess the segmentation quality of our proposed model, seven state-of-the-art models, namely, FCN [18], PSPNet [29], ReFineNet [57], U-Net [23], BiSeNet [40], CGNet [58], and SegNet [59], were selected for comparison with our approach on the ISPRS Potsdam dataset. As the simplest methods, FCN, SegNet, and U-Net achieved 87.99%, 86.12%, and 86.2% performances in IoU score, respectively. The performance of PSP was very disappointing, with IoU scores about 2% lower than FCN. The results of those semantic segmentation models on the ISPRS Potsdam test dataset are summarized in Table 1. Our method outperformed existing schemes in predicting the area of a building. Though some tree shading caused the wrong extraction of the building in Figure 6, our method focused more on the edge. The edge in the predicted results was more continuous and clearer than that in the other techniques. The incorrect classification of other ground objects as buildings rarely occurred with the EANet, but some buildings were improperly classified.

**Table 1.** A comparison experiment with state-of-the-art models on the International Society for Photogrammetry and Remote Sensing (ISPRS) Potsdam test set. IoU: intersection-over-union. The bold type represents the best data under that metrics.

Method	Backbone	F1	Precision	Recall	IoU
BiSeNet [40]	ResNet-101	0.9524	0.9653	0.9451	0.8871
FCN-8s [18]	VGG-16 [63]	0.9427	0.9619	0.9297	0.8799
PSPNet [29]	ResNet-101	0.9424	0.9525	0.9370	0.8554
CGNet [58]	ResNet-101	0.9405	0.9600	0.9269	0.8729
RefineNet [57]	ResNet-101	0.9473	0.9635	0.9373	0.8849
U-Net [23]	ResNet-101	0.9386	0.9547	0.9284	0.8620
SegNet [59]	VGG-16 [63]	0.9419	0.9549	0.9351	0.8612
EANet	ResNet-101	<b>0.9548</b>	<b>0.9702</b>	<b>0.9454</b>	<b>0.9019</b>

As a dataset with larger data volume and a more complex situation, the WHU building dataset presented a greater challenge in correctly extracting buildings from remote sensing images relative to smaller and simpler datasets. To further demonstrate the effectiveness of our proposed model, we also selected seven most advanced models, namely Deeplabv3-plus [24], PSPNet [29], FCN [18], U-Net [23], SRINet [60], DeNet [17], and SiU-Net [37], for the experiment on the WHU building dataset. Deeplabv3-plus, as a carefully designed network, has always been one of the best performing models in the field of semantic segmentation. It achieved a performance of 91.96% on the IoU score. The three models of DeNet, SRINet, and SiU-Net are similar to our idea. They also use edge information to help network learning, and they achieved IoU scores of 90.12%, 89.09%, and 88.4%, respectively, on the WHU building dataset. As indicated by Table 2, our model performed well on the state-of-the-art WHU building datasets. The EANet achieved the top performance of 93.33% of the IoU in the WHU building validation set, which was better than all other models in the experiment, and even 1.37% of the IoU more than Deeplabv3-plus. In our approach, it is almost impossible to misidentify a car as a roof. As shown in Figure 7, for buildings with regular boundaries, our method had better performance and could completely extract buildings. Compared with the ISPRS Potsdam dataset, the WHU building dataset was more different from the environment, which made it easier for the model to extract buildings from aerial images. Therefore, our model performed better in the ISPRS

Potsdam dataset than in the WHU building dataset. Compared with other schemes that use more global information to improve performance, our model focuses more on spatial information such as edges. Making full use of this low level of spatial information proved to be very helpful in extracting ground objects from high-resolution aerial images.



**Figure 6.** Experimental results on the ISPRS Potsdam test dataset. (a) Original image selected from the dataset. (b) Image label. (c) BiSeNet processing result. (d) EANet processing result.

**Table 2.** Accuracy levels with the WHU building validation set. The bold type represents the best data under that metrics.

Method	F1	Precision	Recall	IoU
Deeplabv3-plus [24]	0.9676	<b>0.9867</b>	0.9566	0.9196
PSPNet [29]	0.9669	0.9853	0.9562	0.9182
FCN-8s [18]	0.9651	0.9813	0.9528	0.9032
U-Net [23]	0.9135	0.9542	0.8826	0.8813
SiU-Net [37]	-	0.9380	0.9390	0.8840
SRINet [60]	0.9423	0.9521	0.9328	0.8909
DeNet [17]	0.9480	0.9500	0.9460	0.9012
EANet	<b>0.9752</b>	<b>0.9867</b>	<b>0.9642</b>	<b>0.9333</b>



**Figure 7.** Experimental results with the Wuhan University (WHU) building validation set. (a) Original image selected from the dataset. (b) Image label. (c) BiSeNet processing result. (d) EANet processing result.

In summary, we conducted comparative experiments on two different datasets with eleven other SOTA models to verify whether EANet could obtain high-quality segmentation results. Experiments confirmed that our model had better results than other SOTA models in all evaluation metrics, which not only indicated that our model has a good performance in building extraction but also indicates that the increase of spatial information, especially edge information, is conducive to the automatic extraction of buildings in a network.

#### 4. Discussion

In this section, the impact of the various parts of the proposed network on the results is first examined. Then, we discuss our conducted experiments on the ISPRS Potsdam dataset to test the role of each part of our proposed model. Finally, the future work is also discussed.

##### 4.1. Ablation Study

Ablation research consists of two parts: architecture ablation and loss ablation. Our model served as a kind of u-shaped structure [38,39,60,64], and we selected U-Net [23] as the baseline.

In the architecture ablation experiment, the effectiveness of atrous convolution was first tested. Specifically, the atrous convolutional layer, whose dilation rate was equal to 2, was used to replace the last stage of ResNet, which could add more receptive field and global semantic information without reducing the spatial resolution. Furthermore, we also used our proposed ESRE and FAB modules to replace the baseline decoder in the experiment. As our decoding module contains multi-scale feature fusion, different from the up-sampling strategy with a ratio of 4 that is adopted in the last layer of the decoder in most segmentation networks, our feature map adopted an up-sampling operation with a ratio of 2 after multi-scale fusion, which makes the results more smooth. The IoU was adopted as the evaluation metric (Table 3). The IoU score was substantially increased from 89.17% to 92.75% when using the feature fusion method we proposed, which reflected the advantages of ESRE and FAB modules and atrous convolution over the decoder in the baseline.

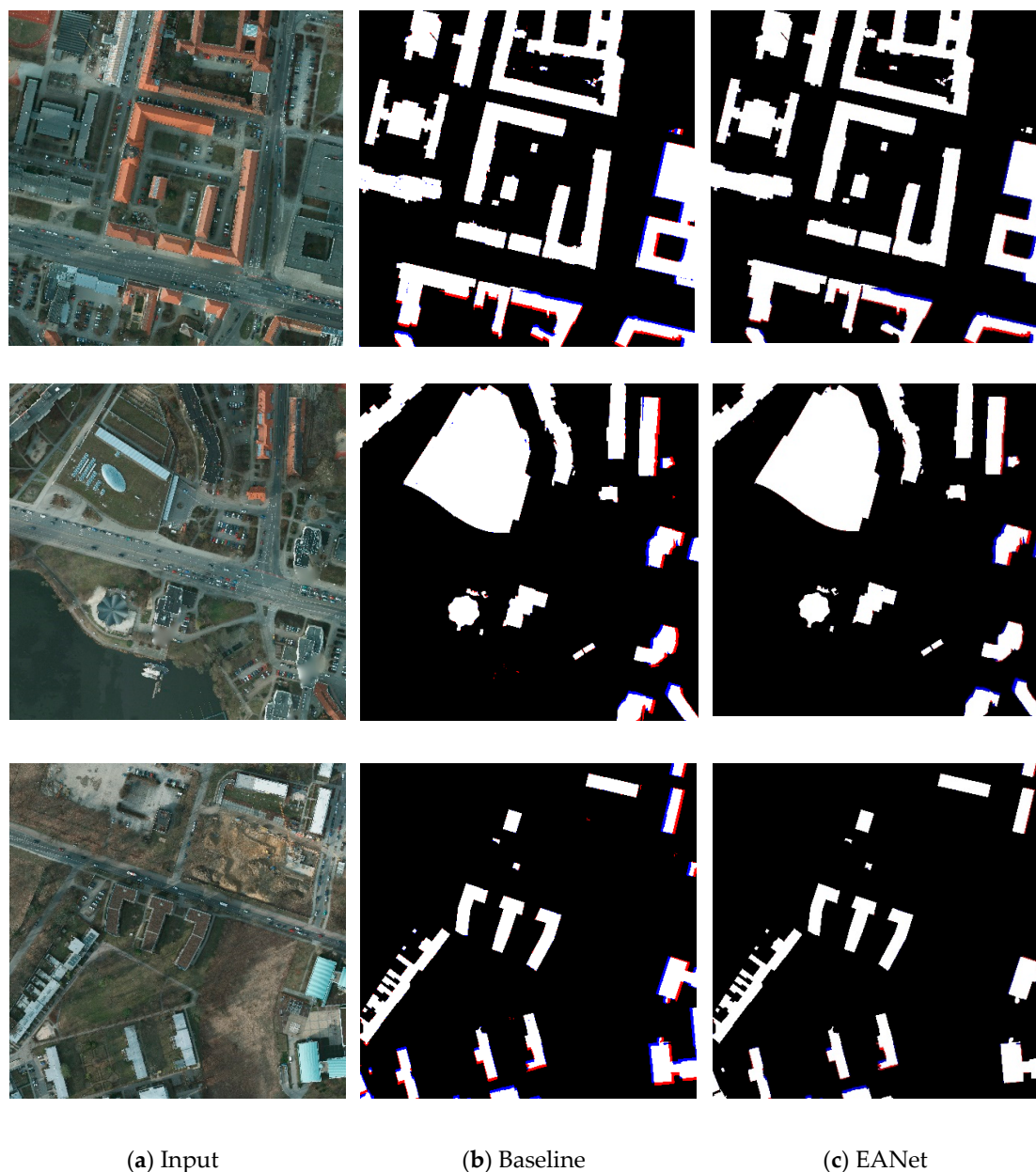
**Table 3.** Ablation experiments of the methods in Section 4.1. The IoU (%) was adopted as the evaluation metric on the ISPRS Potsdam dataset. The baseline model was the U-Net [23]. Atrous: atrous convolution layer; ESRE and FAB: explicit spatial resolution embedding and feature aggregation block; EEB: edge extraction branch. The bold type represents the best data.

Index	Baseline	Atrous	ESRE and FAB	EEB	EALoss	IoU (%)
1	✓					89.17
2	✓	✓				91.80
3	✓		✓			92.51
4	✓	✓	✓			92.75
5	✓	✓	✓	✓		93.58
6	✓	✓	✓	✓	✓	<b>94.09</b>

To further demonstrate the generality of the EANet, different loss functions were used in the loss ablation experiment to guide the model in learning the building extraction task. We first added the EEB module to extract image edges based on a baseline and used the BCE loss to render the predicted results closer to the ground-truth label. With the constraint of BCE loss function, the feature map of each stage in the decoder pays more attention to the learning of edge information. Finally, to verify the effectiveness of our proposed EALoss, we added it to the baseline to enhance edge prediction of segmentation results. The experimental results in Table 3 prove that our suggested approach was significantly better than the baseline on the IoU and the ablation results reported in Table 3. As shown in Figure 8, the number of false-positive pixels in our approach was significantly reduced compared to



the baseline, possibly because the model effectively separated the background from the building using edge information. At the same time, the number of false-negative pixels was not significantly reduced from baseline, possibly because the difference between the environment and the building was too small for the model to accurately distinguish. The EANet focuses more on the edge than on the baseline and has fewer misdetections and omissions. Thus, the EANet can learn low-level spatial information well.



**Figure 8.** Results of ablation experiments on the ISPRS Potsdam dataset. White indicates true positive pixels, blue indicates false positive pixels, black indicates true negative pixels, and red indicates false negative pixels. (a) Original image selected from the dataset. (b) The processing result of U-Net as the baseline. (c) The processing result of our proposed method, the EANet.

#### 4.2. Future Work

In today's semantic segmentation community, a technique called non-local [65] is widely used by researchers in neural networks to capture the long-distance dependence of two non-adjacent pixels in an image. As can be seen from Figure 7, our network generates voids for extracting some particularly



large buildings. We speculate that this situation is caused by insufficient semantic information rather than the lack of low-level detailed information. Adopting non-local technology to obtain more complete semantic information may greatly improve this situation.

Though our proposed network presents some improvement regarding the loss of spatial information, that information loss due to down-sampling is irreversible [66]. Therefore, using a high-resolution representation [67–69] for calculation may be useful. Meanwhile, in our suggested method, the edge is only used as a constraint to supervise the learning of the model. Perhaps there is a way to incorporate the extracted edges into the network's stream of traffic rather than just using it as a guide. Additionally, the artificial building boundary presented in the remote sensing image is uncertain, and errors occur in artificial labeling. Moreover, the labeling of the boundary is affected by noise. Future work can emphasize the investigation of robust algorithms free from noise interference.

## 5. Conclusions

This work proposed the EANet, a novel encoder–decoder edge-aware network with an edge-aware loss for accurate building extraction from remote sensing images. The EANet presents an end-to-end architecture consisting of two components: an image segmentation network and an edge perception network. The image segmentation network aims to obtain high-quality segmentation results from images. Conversely, the edge perception network guides the segmentation network toward paying more attention to edge information and restores lost low-level details as much as possible. The ISPRS Potsdam and the WHU building datasets, respectively, cover two different cities. Both datasets contain civil and industrial buildings that fully demonstrate the complexity of urban buildings. Compared with the existing eleven state-of-the-art methods, our network was found to have the best performance for the extraction of buildings according to experiments with the ISPRS Potsdam and WHU building datasets, with the proposed EANet achieving the highest F1 and IoU (97.52% and 93.33%, respectively) compared with Deeplabv3-plus (96.76% and 91.96%, respectively), PSPNet (96.69% and 91.82%, respectively), U-Net (91.35% and 88.13%, respectively), SRINet (94.23% and 89.09%, respectively), DeNet (94.80% and 90.12%, respectively) for the WHU buildings dataset. For the extraction of dense buildings, the results showed that our method performed better. Meanwhile, our network is simple and efficient, and it can not only be applied to the extraction of buildings in other cities or regions but can also be easily extended to the extraction of other ground objects of remote sensing images.

**Author Contributions:** Q.Z. conceived of the presented idea and designed the study. G.Y. derived the models and performed the experiments. The manuscript was drafted by G.Y. with support from Q.Z. and G.Z. All authors discussed the results and contributed to the final manuscript. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the National Nature Science Foundation of China (Grant Nos. 61731009 and 41301472) and the Science and Technology Commission of Shanghai Municipality (Grant No. 19511120600).

**Acknowledgments:** The authors would like to thank Xiangyu Lei for his advice on models.

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

## References

1. Ahmadi, S.; Zoj, M.J.V.; Ebadi, H.; Abrishami, H.; Mohammadzadeh, A. Automatic urban building boundary extraction from high resolution aerial images using an innovative model of active contours. *Int. J. Appl. Earth Obs. Geoinf.* **2010**, *12*, 150–157. [\[CrossRef\]](#)
2. Liu, C.; Huang, X.; Zhu, Z.; Chen, H.; Tang, X.; Gong, J. Automatic extraction of built-up area from ZY3 multi-view satellite imagery: Analysis of 45 global cities. *Remote Sens. Environ.* **2019**, *226*, 51–73. [\[CrossRef\]](#)
3. Li, J.; Huang, X.; Gong, J. Deep neural network for remote sensing image interpretation: Status and perspectives. *Natl. Sci. Rev.* **2019**, *6*, 1082–1086. [\[CrossRef\]](#)
4. Huang, X.; Cao, Y.; Li, J. An automatic change detection method for monitoring newly constructed building areas using time-series multi-view high-resolution optical satellite images. *Remote Sens. Environ.* **2020**, *244*, 111802. [\[CrossRef\]](#)

5. Peng, J.; Zhang, D.; Liu, Y. An improved snake model for building detection from urban aerial images. *Pattern Recognit. Lett.* **2005**, *26*, 587–595. [[CrossRef](#)]
6. Müller, S.; Zaum, D. Robust Building Detection in Aerial Images. In Proceedings of the International Archives of Photogrammetry and Remote Sensing, Vienna, Austria, 29–30 August 2005; pp. 143–148.
7. Liu, Z.; Cui, S.; Yan, Q. Building extraction from high resolution satellite imagery based on multi-scale image segmentation and model matching. In Proceedings of the International Workshop on Earth Observation and Remote Sensing Applications, Beijing, China, 30 June–2 July 2008.
8. Shackelford, A.K.; Davis, C.H.; Wang, X. Automated 2-D Building Footprint Extraction from High-Resolution Satellite Multispectral Imagery. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Honolulu, HI, USA, 20–24 September 2004; pp. 1996–1999.
9. Zhang, Q.; Huang, X.; Zhang, G. Urban Area Extraction by Regional and Line Segment Feature Fusion and Urban Morphology Analysis. *Remote Sens.* **2017**, *9*, 663. [[CrossRef](#)]
10. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Convolutional Neural Networks for Large-Scale Remote-Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 645–657. [[CrossRef](#)]
11. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
12. 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.
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, 26 June–1 July 2016; pp. 770–778.
14. Vakalopoulou, M.; Karantzas, K.; Komodakis, N.; Paragios, N. Building detection in very high resolution multispectral data with deep learning features. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium, Milan, Italy, 26–31 July 2015; pp. 1873–1876.
15. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 13–16 December 2015; pp. 1440–1448.
16. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.; Berg, A.C. SSD: Single Shot Multibox Detector. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37.
17. Liu, H.; Luo, J.; Huang, B.; Hu, X.; Sun, Y.; Yang, Y.; Xu, N.; Zhou, N. DE-Net: Deep Encoding Network for Building Extraction from High-Resolution Remote Sensing Imagery. *Remote Sens.* **2019**, *11*, 2380. [[CrossRef](#)]
18. 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, 8–10 June 2015; pp. 3431–3440.
19. 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. Available online: <https://arxiv.org/abs/1805.02091> (accessed on 5 May 2018).
20. Marmanis, D.; Wegner, J.D.; Galliani, S.; Schindler, K.; Datcu, M.; Stilla, U. Semantic Segmentation of Aerial Images with an Ensemble of CNNs. *ISPRS Annals Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *3*, 473–480.
21. 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](#)]
22. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [[CrossRef](#)] [[PubMed](#)]
23. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
24. 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.

25. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:abs/1706.05587. Available online: <https://arxiv.org/abs/1706.05587> (accessed on 17 June 2017).
26. Zeiler, M.D.; Taylor, G.W.; Fergus, R. Adaptive deconvolutional networks for mid and high level feature learning. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 2018–2025.
27. 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, 16–20 June 2019; pp. 3126–3135.
28. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. In Proceedings of the International Conference on Learning Representations, San Juan, Puerto Rico, 2–4 May 2016.
29. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
30. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–20 June 2018; pp. 8759–8768.
31. Dai, J.; He, K.; Sun, J. Instance-aware semantic segmentation via multi-task network cascades. In Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
32. Zhang, Z.; Luo, P.; Loy, C.C.; Tang, X. Facial landmark detection by deep multi-task learning. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 94–108.
33. Zeiler, M.; Fergus, R. Visualizing and Understanding Convolutional Networks. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–7 September 2014; pp. 818–833.
34. 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, 18–20 June 2018; pp. 1857–1866.
35. Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; Jagersand, M. BASNet: Boundary-Aware Salient Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 7479–7489.
36. ISPRS 2D Semantic Labeling Contest. July 2018. Available online: <http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html> (accessed on 2 July 2018).
37. Ji, S.P.; Wei, S.Q.; 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]
38. Lin, T.; Dollar, 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.
39. Pohlen, T.; Hermans, A.; Mathias, M.; Leibe, B. Full-resolution residual networks for semantic segmentation in street scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3309–3318.
40. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 334–349.
41. Zhao, H.; Zhang, Y.; Liu, S.; Shi, J.; Loy, C.; Lin, D.; Jia, J. Psanet: Pointwise spatial attention network for scene parsing. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 270–286.
42. Zhang, Z.; Zhang, X.; Peng, C.; Xue, X.; Sun, J. Exfuse: Enhancing feature fusion for semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 269–284.
43. Zhu, Z.; Xu, M.; Bai, S.; Huang, T.; Bai, X. Asymmetric non-local neural networks for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 593–602.

44. Wang, P.; Chen, P.; Yuan, Y.; Liu, D.; Huang, Z.; Hou, X.; Cottrell, G. Understanding convolution for semantic segmentation. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1451–1460.
45. Nair, V.; Hinton, G. Rectified Linear Units Improve Restricted Boltzmann Machines Vinod Nair. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel, 21–24 June 2010.
46. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* **2015**, arXiv:1502.03167. Available online: <https://arxiv.org/abs/1502.03167> (accessed on 11 February 2015).
47. Liu, Y.; Cheng, M.; Hu, X.; Wang, K.; Bai, X. Richer Convolutional Features for Edge Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5872–5881.
48. Xie, S.; Tu, Z. Holistically-nested edge detection. In Proceedings of the IEEE Conference on Computer Vision, Santiago, Chile, 7–9 December 2015; pp. 1395–1403.
49. Liu, Y.; Lew, M. Learning relaxed deep supervision for better edge detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 231–240.
50. Haralick, R.; Sternberg, S.; Zhuang, X. Image Analysis Using Mathematical Morphology. *IEEE Trans. Pattern Anal. Mach. Intell.* **1987**, *9*, 532–550. [[CrossRef](#)] [[PubMed](#)]
51. Boer, P.; Kroese, D.; Mannor, S.; Rubinstein, R. A tutorial on the cross-entropy method. *Ann. Oper. Res.* **2005**, *134*, 19–67. [[CrossRef](#)]
52. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in pytorch. In Proceedings of the NIPS 2017 Workshop, Long Beach, CA, USA, 4–9 December 2017.
53. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, F. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009.
54. He, T.; Zhang, Z.; Zhang, H.; Zhang, Z.; Xie, J.; Li, M. Bag of tricks for image classification with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 558–567.
55. Goyal, P.; Dollar, P.; Girshick, R.; Noordhuis, P.; Wesolowski, L.; Kyrola, A.; Tulloch, A.; Jia, Y.; He, K. Accurate, large minibatch SGD: Training imagenet in 1 hour. *arXiv:1706.02677*. Available online: <https://arxiv.org/abs/1706.02677> (accessed on 8 June 2017).
56. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2818–2826.
57. Lin, G.; Anton, M.; Shen, C. 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.
58. Wu, T.; Tang, S.; Zhang, R.; Zhang, Y. CGNet: A light-weight context guided network for semantic segmentation. *arXiv* **2018**, arXiv:1811.08201. Available online: <http://arxiv.org/abs/1811.08201> (accessed on 20 November 2018).
59. Vijay, B.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder–decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495.
60. 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](#)]
61. Robbins, H.; Monro, S. A stochastic approximation method. *Ann. Math. Stat.* **1951**, *22*, 400–407. [[CrossRef](#)]
62. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1026–1034.
63. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the ICLR, San Diego, CA, USA, 7–9 May 2015.

64. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. *arXiv* **2016**, arXiv:1606.02147. Available online: <https://arxiv.org/abs/1606.02147> (accessed on 7 June 2016).
65. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–20 June 2018; pp. 7794–7803.
66. Zhang, R. Making convolutional networks shift-invariant again. In Proceedings of the ICML, Long Beach, CA, USA, 9–15 June 2019; pp. 7324–7334.
67. Huang, X.; Wang, Y.; Li, J.; Chang, X.; Cao, Y.; Xie, J.; Gong, J. High-resolution urban land-cover mapping and landscape analysis of the 42 major cities in China using ZY-3 satellite images. *Sci. Bull.* **2020**, *65*, 1039–1048. [[CrossRef](#)]
68. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5693–5703.
69. Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Xiao, B.; Liu, D.; Mu, Y.; Wang, X.; Liu, W.; Wang, J. High-resolution representations for labeling pixels and regions. *arXiv* **2019**, arXiv:1904.04514. (accessed on 9 April 2019).



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).