

PAN, X., YANG, F., GAO, L., CHEN, Z., ZHANG, B., FAN, H. and REN, J. 2019. Building extraction from high-resolution aerial imagery using a generative adversarial network with spatial and channel attention mechanisms. *Remote sensing* [online], 11(8), article 917. Available from: <https://doi.org/10.3390/rs11080917>

# Building extraction from high-resolution aerial imagery using a generative adversarial network with spatial and channel attention mechanisms.





PAN, X., YANG, F., GAO, L., CHEN, Z., ZHANG, B., FAN, H. and REN, J.

2019

© 2019 by the authors. Licensee MDPI, Basel, Switzerland.

Article

# Building Extraction from High-Resolution Aerial Imagery Using a Generative Adversarial Network with Spatial and Channel Attention Mechanisms

Xuran Pan <sup>1,2</sup>, Fan Yang <sup>1,\*</sup>, Lianru Gao <sup>2</sup>, Zhengchao Chen <sup>2</sup>, Bing Zhang <sup>2,3</sup>, Hairui Fan <sup>1</sup> and Jinchang Ren <sup>4</sup>

<sup>1</sup> School of Electronics and Information Engineering, Hebei University of Technology, Tianjin 300401, China; 201611901006@stu.hebut.edu.cn (X.P.); 201731904001@stu.hebut.edu.cn (H.F.)

<sup>2</sup> Key Laboratory of Digital Earth Science, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100094, China; gaolr@radi.ac.cn (L.G.); chenzc@radi.ac.cn (Z.C.); zb@radi.ac.cn (B.Z.)

<sup>3</sup> College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China

<sup>4</sup> Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, G1 1XW, UK; jinchang.ren@strath.ac.uk

\* Correspondence: 201621901026@stu.hebut.edu.cn

Received: 18 March 2019; Accepted: 12 April 2019; Published: 15 April 2019



**Abstract:** Segmentation of high-resolution remote sensing images is an important challenge with wide practical applications. The increasing spatial resolution provides fine details for image segmentation but also incurs segmentation ambiguities. In this paper, we propose a generative adversarial network with spatial and channel attention mechanisms (GAN-SCA) for the robust segmentation of buildings in remote sensing images. The segmentation network (generator) of the proposed framework is composed of the well-known semantic segmentation architecture (U-Net) and the spatial and channel attention mechanisms (SCA). The adoption of SCA enables the segmentation network to selectively enhance more useful features in specific positions and channels and enables improved results closer to the ground truth. The discriminator is an adversarial network with channel attention mechanisms that can properly discriminate the outputs of the generator and the ground truth maps. The segmentation network and adversarial network are trained in an alternating fashion on the Inria aerial image labeling dataset and Massachusetts buildings dataset. Experimental results show that the proposed GAN-SCA achieves a higher score (the overall accuracy and intersection over the union of Inria aerial image labeling dataset are 96.61% and 77.75%, respectively, and the  $F_1$ -measure of the Massachusetts buildings dataset is 96.36%) and outperforms several state-of-the-art approaches.

**Keywords:** high-resolution aerial images; deep learning; generative adversarial network; semantic segmentation; Inria aerial image labeling dataset; Massachusetts buildings dataset

## 1. Introduction

With the rapid advancement of aerospace remote sensing, the amount and spatial resolution of high-resolution remote sensing images are increasing rapidly. As a result, accurate and automatic semantic labeling of high-resolution remote sensing images is of great significance and receives wide attention [1]. Large intra-class variance and small inter-class differences of higher spatial resolution remote sensing images may cause classification ambiguities, which makes semantic segmentation of high-resolution remote sensing images a challenge. Specific to the buildings in high-resolution aerial images, buildings in different regions have different characteristics. For instance, some regions have small and very dense buildings, whilst some other regions have low-density buildings. This variability

brings great challenges to the building segmentation task, and requires strong generalization capabilities of classification techniques [2,3].

Over the last few years, deep learning architectures have made breakthroughs in the image analysis field. Convolutional neural networks (CNNs) have been proposed not only to deal with object detection and whole image classification but also progress fine inference, such as semantic segmentation. Semantic segmentation can accomplish pixel-wise prediction, which is a problem to give each pixel a class label. Long et al. [4] proposed fully convolutional networks (FCNs) to accomplish pixel-wise classification. They replaced the fully connected layers of whole image classification CNNs with convolutional layers and utilized deconvolutional layers to upsample feature maps to score map each class. FCNs created a precedent for pixel-based encoder–decoder architectures. Following this paradigm, many CNN architectures have been proposed and further improved the segmentation performance. In [5], U-Net was proposed to modify FCN by concatenating feature maps of encoder and decoder. Concatenation architecture can take full advantage of both low-level and high-level features. Hence, more precise segmentation results can be obtained. After that, DeepLab V1 [6] and V2 [7] were proposed to mitigate the information loss caused by pooling operations. The authors introduced atrous convolutions to increase receptive field size while maintaining higher resolution of feature maps, and the fully connected conditional random fields (CRFs) were utilized to further improve the segmentation performance as post processor. In [8], Noh et al. proposed DeconvNet which consists of convolution and deconvolution networks. In the deconvolution network, unpooling layers were applied to upscale feature maps and decovoconvolutional layers were followed to densify the initially upscaled sparse feature maps. Badrinarayanan et al. presented SegNet [9] which also included unpooling layers in the decoder stage and with smaller parameterization when compared with DeconvNet.

Although the CNN-based segmentation methods have achieved promising results, they still have drawbacks and can be further improved. The main problem is that the pixel-wise prediction of CNN can guarantee high pixel-wise accuracy, but the relationship between pixels is prone to be ignored. This may lead to discontinuous segmentation results, and the boundaries of objects are usually not accurate enough. Therefore, post-processing methods, e.g., fully connected CRFs or Markov random fields (MRFs), were needed to further improve the raw segmentation results [10–12]. These graphical regularization models coupled both the input images and the predicted score maps of CNN to refine the predictions with the color information and pixel position of the original image. In addition, recurrent neural networks (RNN) can also refine the segmentation results by employing a feedback connection to form a directed cycle [13]. Bergado et al. [14] proposed to incorporate the recurrent approach in the semantic segmentation task (ReuseNet) to learn contextual dependencies in the label space, and further refine the segmentation results. The ReuseNet applied the semantic segmentation operations in  $R$  cycles. Each cycle takes the score map of the previous cycle concatenated with the original image as input. Moreover, Generative adversarial networks (GANs) [15] based methods can enforce spatial label contiguity to refine the segmentation results without any time consumption during the testing phase. In [16], Luc et al. first applied adversarial training strategy to semantic segmentation task. A segmentation network and an adversarial network were trained in an alternating fashion to make the generated segmentation results hard to be distinguished from the ground truth. By doing so, the joint distribution of all label variables at each pixel location can be assessed as a whole, and thus, can enforce forms of high-order consistency that cannot be enforced by pixelwise classification or pairwise terms. Xue et al. [17] presented SegAN for medical image segmentation, which is composed of a segmentor and a critic network. The multi-scale  $L_1$  loss function was minimized and maximized alternatively to train these two networks, and the SegAN received better image segmentation performance than the original GAN.

In semantic labeling of high-resolution remote sensing images, deep learning architectures also show excellent performance [18–20]. Saito et al. [21] used patch-based CNN to learn classification maps from high-resolution images and achieved good results on Massachusetts roads

and buildings datasets [22]. However the patch-based methods suffer from limited receptive field and large computational overhead, so it was soon surpassed by pixel-based methods. Maggiori et al. [23] proposed the Inria aerial image labeling dataset that covers different forms of buildings and provided a baseline segmentation result by using an FCN-based architecture combined with multi-layer perceptron. In [24], Bischke et al. introduced a new cascaded multi-task loss to mitigate the poor boundaries of the existing prediction results. Learning with the proposed loss, the performance can achieve certain improvement without any changes in the network architecture. A multi-stage multi-task CNN for building extraction was introduced in [18]. The first stage of the proposed network provided the segmentation results, while the second stage was aimed to give the precise location by two branches. In [25], Khalel et al. proposed a stack of U-Nets to automatically label the buildings from high aerial images, of which each U-Net can be regarded as the post-processor of the previous U-Net. However, the existing results usually suffered from poor boundaries, and the accuracy can be further improved.

In this paper, we propose a generative adversarial network with spatial and channel attention mechanisms (GAN-SCA) for high accurate semantic labeling of buildings in high-resolution aerial images with precise boundaries. The GAN-SCA is composed of a segmentation network and an adversarial network, in which the segmentation network is a semantic segmentation network to predict the pixel-wise labeling results, and the adversarial network is to distinguish whether the inputs are predicted results of the segmentation network or ground truth. Moreover, we embed channel and spatial attention mechanisms into the network to selectively enhance useful information, and further improve the segmentation accuracy.

The main contributions of our work can be summarized as follows:

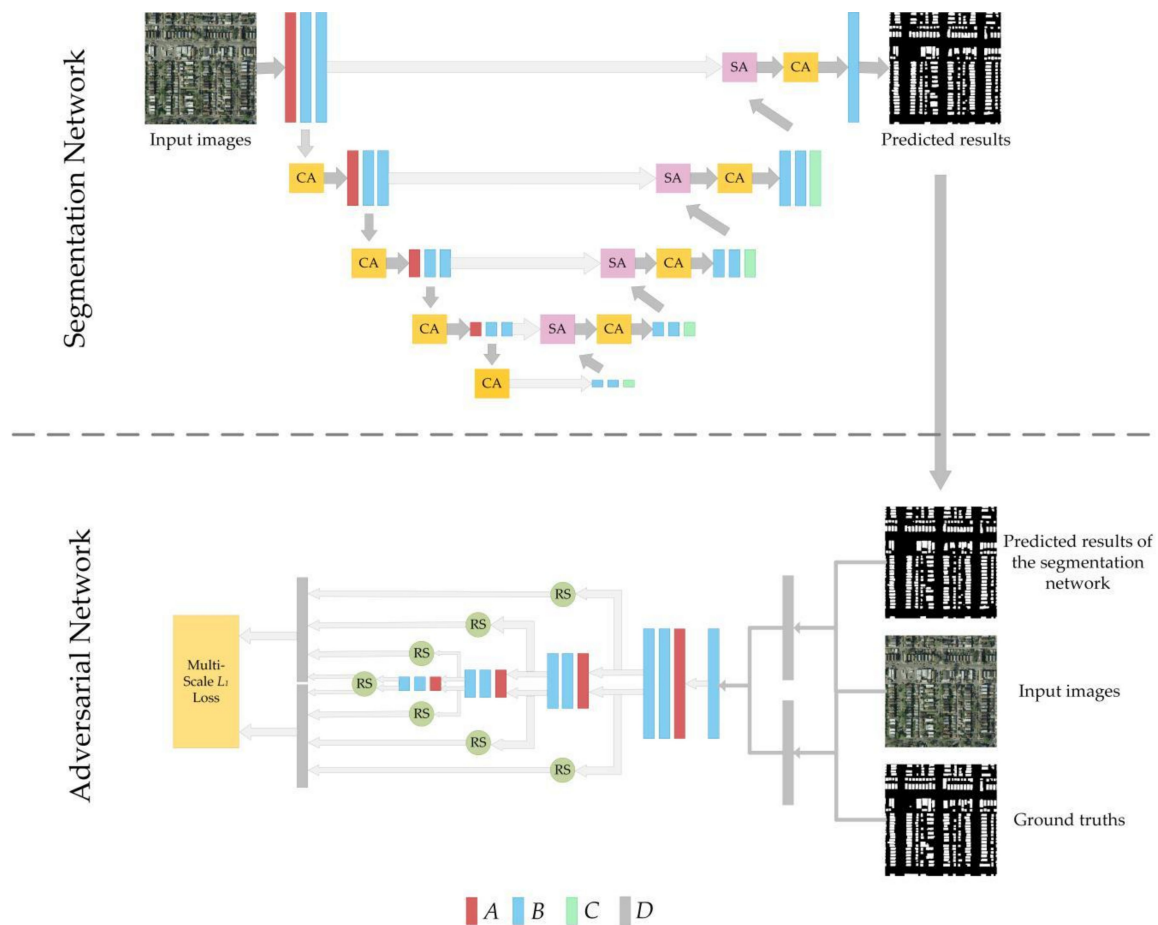
- A GAN-based network called GAN-SCA is proposed for building extraction from high-resolution aerial imagery. The architecture is composed of a segmentation network and an adversarial network. The segmentation network aims to predict pixel-wise labeling maps that are similar to ground truths, while the adversarial network is set to discriminate different characteristics of different label maps to further enhance the high-frequency continuity of the prediction maps.
- Spatial and channel attention mechanisms are embedded in the proposed GAN-SCA architecture to enable selectively attaching important features from both the spatial dimension and channel relationship.
- The adversarial network and segmentation network are trained to optimize a multi-scale  $L_1$  loss and multiple cross entropy losses combined with a multi-scale  $L_1$  loss alternatively. With no requirements for any post-processing, our proposed network improved the state-of-the-art performance on both the Inria aerial image labeling dataset and Massachusetts buildings dataset.

The rest of this paper is organized as follows. In Section 2, we introduce the architecture and training strategy of the proposed network in detail. The dataset description and experimental setting are presented in Section 3. Section 4 details the experimental results and analyses. Section 5 discusses the effectiveness of the spatial and channel attention mechanisms and the training strategy. The results are drawn in Section 6.

## 2. Methods

### 2.1. Proposed Network GAN-SCA

As shown in Figure 1, the proposed GAN-SCA is composed of two parts, i.e., the segmentation network and the adversarial network.



**Figure 1.** Architecture of the proposed generative adversarial network with spatial and channel attention mechanisms (GAN-SCA). *A* is max pooling layer; *B* are convolutional + batch normalization + rectified linear unit (ReLU) layers; *C* is upsampling layer; *D* is the concatenation operation; SA is the spatial attention mechanism; CA is the channel attention mechanism; RS is the reshape operation.

The segmentation network is a U-Net-based architecture, where spatial and channel attention mechanisms are embedded. U-Net is a powerful CNN architecture for semantic segmentation and has been widely applied in remote sensing image classification field [5]. U-Net was initially designed for binary segmentation of biomedical images with a relatively small number of training samples. As it achieves better performance than other classic semantic segmentation architecture, U-Net is a good choice for the building extraction task in this study. However, these classic deep convolutional neural network (DCNN) architectures for semantic segmentation usually produce a large number of multi-level feature maps but do not perform any feature selection operation throughout the whole process. On the one hand, fusion of the high-level and low-level features without feature selection may result in over-segmentation when the model tends to receive more information from lower layers. On the other hand, the channel-wise information combined by convolutional filters without considering channel-interdependencies might affect the segmentation performance of the network. Therefore, we propose to introduce the attention mechanisms to employ feature selection from the aspect of spatial information and channel relationship.

To mitigate the neglect of inter-pixel relationships caused by the pixel-wise loss function used in the training phase, we propose to refine the segmentation result using the adversarial training. The adversarial network can learn latent higher-order structural features which can be fed into the segmentation network in the training phase, and the segmentation results can be refined without an adversarial network in the testing phase. In contrast with graphical models and recurrent

approaches, adversarial training can achieve segmentation refinement without extra time consumption. The architecture of adversarial network we adopt in the proposed GAN-SCA shares a similar structure as the encoder of the segmentation network and is fed with the predicted maps combined with the original images and ground truth maps combined with the original images. In particular, the multi-scale features from a different stage of the adversarial network are reshaped into one-dimensional vectors and concatenated together to compute the multi-scale  $L_1$  loss.

### 2.1.1. Segmentation Network

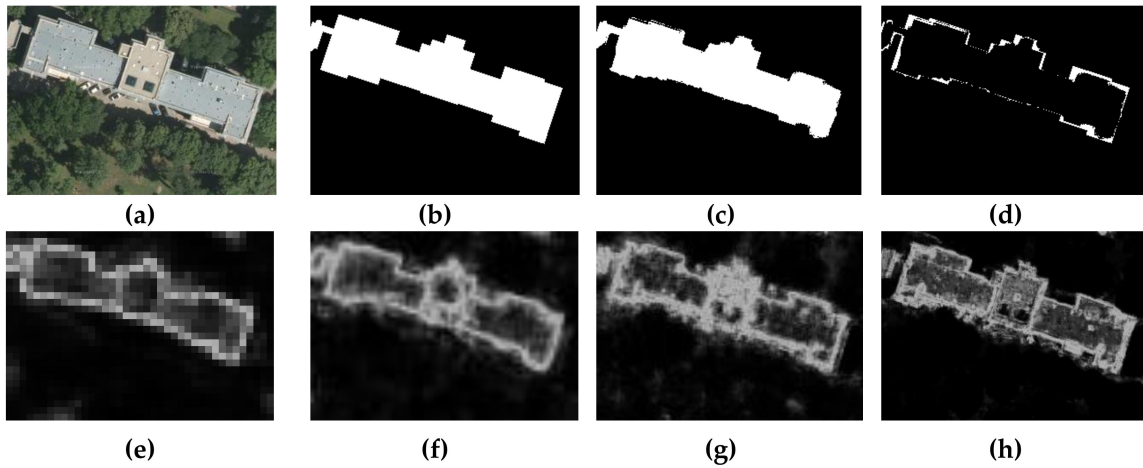
The segmentation network of GAN-SCA is based on U-Net architecture. To accomplish feature selection from the aspect of the spatial information and channel-wise relationship, we introduce two kinds of attention mechanisms into the network architecture. The attention mechanism is an effective operation to enable the network to selectively enhance more useful features and has been widely applied in the image analysis field [26]. In this work, we consider both spatial and channel-wise attention mechanisms to improve the segmentation performance. The spatial attention mechanisms are embedded between the contracting path and expanding path of the U-Net, as shown in Figure 1. The U-Net fuses low-level feature maps of the contracting path with the high-level features of the expanding path by concatenation to re-utilize fine details in the low-level features. However, the rough concatenation may result in the over-use of low-level features. Therefore, we can utilize flexible semantic information of the high-level features to assist the selection of low-level information. Usually, the low-level features contain rich details, and we prefer to enhance the hard classified information and suppress the interference information. Figure 2 shows the error map of U-Net prediction result, from which we can observe that building boundaries are prone to be mislabeled in the building extraction task. Inspired by [27], the entropy score map of high-level features has similar characteristics with the mislabeled map, as shown in Figure 2. Therefore, when we compute the entropy score map of high-level features in each decoder stage, and weight the low-level features according to the results of corresponding entropy score map before high-level and low-level feature fusion, we can selectively enhance the hard classified information while suppressing the less useful information of the low-level features. The entropy score map can be computed with Equation (1):

$$E(x) = - \sum_{i=1}^K p_i(x) \log(p_i(x)) \quad (1)$$

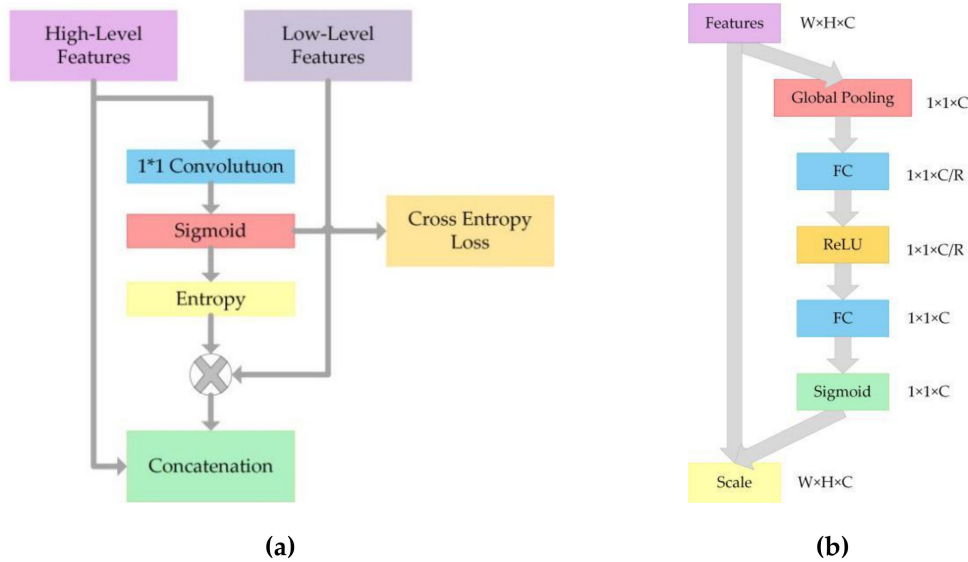
where  $p_i(x)$  denotes the score map of class  $i$ ,  $K$  means the total number of the classes. Figure 2 displays the entropy score maps of four-scale spatial attention mechanisms, from which we can see that the entropy maps have a strong relationship with the error map. Usually, building boundary pixels are prone to being mislabeled, so the entropy maps also share similar characteristics with the boundaries of buildings. Thus, with the spatial attention mechanisms, building boundaries information from lower level features will be highly weighted into the final output fusion feature.

The detailed structure of the spatial attention mechanism is shown in Figure 3a. As can be seen, high-level features are first convoluted by  $1 \times 1$  convolutions for dimensionality reduction and normalized to  $[0,1]$  by using the sigmoid function to generate the score maps. Afterward, the entropy score map is computed to element-wise conducts with low-level features. After that, the high-level features are concatenated with the weighted low-level features to further process. It is worth noting that, the entropy score map has a strong relationship with the building boundaries in the building extraction tasks so the spatial attention mechanisms can bring benefits to the building boundaries segmentation. In particular, we compute four cross entropy losses of each spatial attention mechanism to combine with the overall cross entropy loss to train the segmentation network. The detail of model optimization will be introduced in Section 2.2.





**Figure 2.** Entropy score maps of four-scale spatial attention mechanisms. (a) is the original image; (b) is the ground truth; (c) is the prediction result; (d) is the error map; (e–h) are the entropy score maps of the low-to-high scale spatial attention mechanisms.



**Figure 3.** Composition modules in the GAN-SCA. (a) Spatial attention mechanism; (b) Channel attention mechanism. FC is fully connected layer.

Apart from spatial attention, the proposed architecture also takes advantage of the channel relationship enhancement. Squeeze-and-excitation (SE) block is a computational unit that can re-scale each channel according to its importance adaptively. SE blocks can be stacked together with many existing state-of-the-art CNNs, and bring significant improvements in performance across different datasets with minimal additional computational cost [28]. So we adopt SE blocks as channel attention mechanisms at each stage in both contracting path and expanding path, as shown in Figure 1. The structure of the SE block is depicted in Figure 3b, which can model channel inter-dependencies in two steps, namely, squeeze and excitation. The input features  $x$  are first squeezed into channel-wise statistics  $s$  by performing global average pooling, and the  $c$ -th channel of  $s$  can be computed by:

$$s_c(x_c) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W x_c(i, j) \tag{2}$$

where  $x_c$  is the  $c$ -th channel of the input feature  $x$ , and  $H \times W$  denote spatial dimensions of  $x_c$ .

To properly capture the information of  $s$  to model the channel inter-dependencies, the excitation operation is followed. A fully connected layer is adopted to reduce the dimension of  $s_{1 \times 1 \times C}$  to  $s'_{1 \times 1 \times \frac{C}{R}}$  and a rectified linear unit (ReLU) layer is followed to activate. After that, another fully connected (FC) layer is performed to ascend  $s'$  back to the original dimension  $1 \times 1 \times C$ . By doing so, it can better fit the complex relationship between channels with less computational overhead. The weight of each channel is normalized to  $[0,1]$  with a sigmoid activation. The excitation operation can be written as:

$$e = \sigma(W_2 \delta(W_1 s)) \quad (3)$$

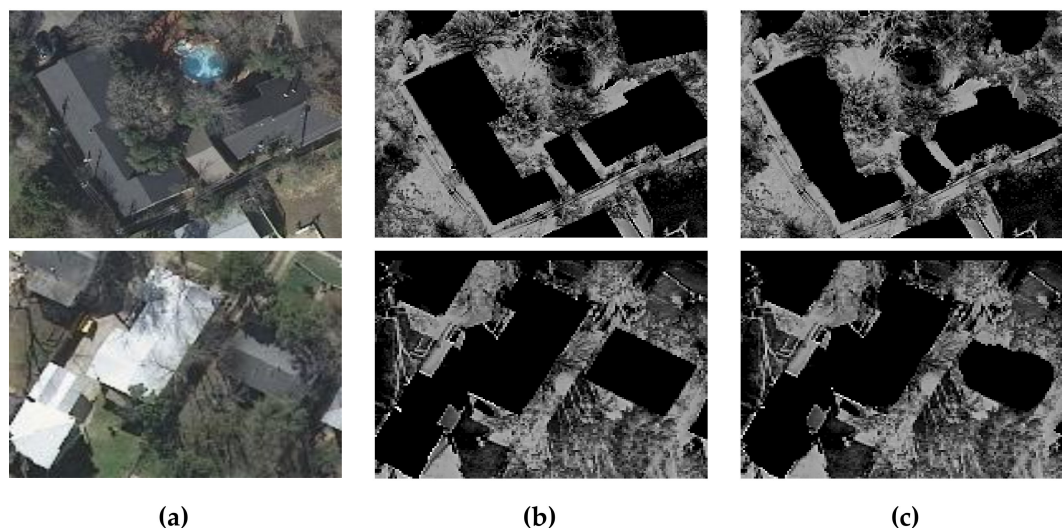
where  $\sigma$  stands for the sigmoid activation, and  $\delta$  stands for the ReLU function [29].  $W_1$  and  $W_2$  are two real matrices of size  $\frac{C}{R} \times C$  and  $C \times \frac{C}{R}$  to limit the complexity and generalization of the channel attention mechanism. This operation is implemented by two FC layers.

The final output of the channel attention mechanism is the re-scaled input features  $y_c$ . The re-scaled operation can be expressed by Equation (4) below:

$$y_c = e_c \cdot x_c \quad (4)$$

### 2.1.2. Adversarial Network

The adversarial network of GAN-SCA has a similar structure with the encoder in the segmentation network. Two inputs are fed into the adversarial network, namely original images concatenated with predicted label maps and original images concatenated with ground truths. The network starts with a  $1 \times 1$  convolutional layer to learn to fuse the input images with the predicted label maps/ground truths. Figure 4 shows two visual results of such fusion. Then the fused images are fed into the encoder-like network to extract features, respectively. To capture long- and short- range spatial relations between pixels, we extract multi-scale feature maps from multiple layers and concatenate them together to compute the multi-scale  $L_1$  loss [17], the detailed introduction of loss function will be presented in the next section.



**Figure 4.** Fusion features of input images (one channel) and the predicted label maps/ground truths. (a) Input images; (b) Fusion results of input images and ground truths; (c) Fusion results of input images and the predicted label maps (5000 iterations).

### 2.2. Training Strategy

The proposed GAN-SCA is trained in an adversarial fashion. The segmentation network aims to generate the predicted labeling map to deceive the adversarial network, and the adversarial network aims to distinguish the ground truths from the predicted labeling maps generated by the segmentation



network. Therefore, the segmentation network and adversarial network are trained alternatively in the training phase [30]. We first fix the parameter of the segmentation network (S) and train adversarial network (A) to minimize the multi-scale  $L_1$  loss (Equation (5)). Then the parameter of A is fixed, and the S is trained by minimizing the cross-entropy losses combined with the negative multi-scale  $L_1$  loss (Equation (7)).

$$L_A = -\frac{1}{N} \sum_{n=1}^N l_{mae}(f_A(x_n, S(x_n)), f_A(x_n, y_n)) \quad (5)$$

where  $(x_n, S(x_n))$  is the concatenation of input images and the predicted results of  $(x_n, y_n)$  is the concatenation of input images and ground truths,  $f_A(x)$  denotes hierarchical features extracted from  $x$ ,  $l_{mae}$  is the  $L_1$  distance or mean absolute error (mae), which is defined as:

$$l_{mae}(f_A(x), f_A(x')) = \frac{1}{L} \sum_{i=1}^L \|f_A^i(x) - f_A^i(x')\|_1 \quad (6)$$

where  $L$  is the total number of the feature scales in the adversarial network,  $f_A^i(x)$  is the features in scale  $i$ .

$$L_S = -\frac{1}{N} \sum_{n=1}^N (y(x_n) \log(S(x_n)) + (1 - y(x_n)) \log(1 - S(x_n))) - L_A + L_{fa} \quad (7)$$

where the  $L_{fa}$  is the auxiliary cross entropy loss computed in each spatial attention mechanism,  $y(x_n)$  denotes the ground truth of the  $n$ -th image in the current batch.

The parameters of the segmentation network and adversarial network are initialized by normally distributed random variables. The initial learning rate is set to  $10^{-3}$  and divided by 2 every 15 epochs. The batch size is set to 5. We crop the training images into size  $384 \times 384$  with 25% overlap, and data augmentation including flip and rotation are also implemented. In the testing phase, to meet the memory constraints, we employ a sliding window with size  $1024 \times 1024$  to accomplish the full tile prediction. We set 75% overlapping size in the testing stage to mitigate inconsistent border phenomenon since the size is proven to give the best results in previous works [11,31].

### 3. Datasets and Evaluation Metrics

#### 3.1. Datasets

The datasets we used in this work are two open buildings datasets, namely Inria aerial image labeling dataset for buildings and Massachusetts buildings dataset. These two datasets cover various building characteristics, such as shape, size, distribution, and spatial resolution, which can evaluate the generalization ability of networks.

The first dataset we used is the Inria aerial image labeling dataset for buildings [23]. The dataset consists of 360 high-resolution aerial images which over different cities including Austin, Chicago, Kitsap, Western/Eastern Tyrol, Vienna, Bellingham, Bloomington, and San Francisco. These regions cover dissimilar urban buildings, for instance, most buildings in Chicago and San Francisco are densely distributed and usually small in shape, while buildings in Kitsap are scattered. The spatial resolution of images is 30 cm with an image size of  $5000 \times 5000$  pixels, and each image covers a surface of  $1500 \times 1500 \text{ m}^2$ . Only 180 tiles are provided with ground truths, and the other 180 tiles are preserved for testing. Following a common practice [23], we choose the first five images of each region from the training set for validation.

The second dataset is the Massachusetts buildings dataset [22]. The dataset consists of 151 high-resolution aerial images of urban and suburban areas at Boston. The size of images in this dataset is  $1500 \times 1500$  pixels, and each image covers a surface of  $2250 \times 2250 \text{ m}^2$ . The dataset is randomly divided into three subsets, namely training set (137 tiles), validation set (4 tiles), and testing set (10 tiles).

### 3.2. Evaluation Metrics

To make a fair comparison, we compute the same metrics as in other literatures. For the Inria Aerial Image Labeling Dataset, the overall accuracy (Acc.) and intersection over union (IoU) are utilized for quantitative performance evaluation. Acc. is the proportion of the correctly labeled pixels (see Equation (8)). IoU is the intersection of pixels labeled as building in the predicted results and ground truths, divided by the union of pixels labeled as building in the predicted results and ground truths (see Equation (9)).

$$Acc. = \frac{tp + tn}{tp + tn + fp + fn} \quad (8)$$

$$IoU = \frac{tp}{fp + tp + fn} \quad (9)$$

where  $tp$  denotes the number of true positive pixels,  $fp$  denotes the number of false positive pixels,  $tn$  denotes the number of true negative pixels, and  $fn$  denotes the number of false negative pixels.

For the Massachusetts buildings dataset, relaxed  $F_1$ -measure is used to evaluate the segmentation performance of each network. A relaxed factor  $\rho$  is introduced when computing the confusion metrics because the tools producer of this dataset used to generate labels is only accurate up to a few pixels. Following the previous works [23–25], we compute the  $F_1$ -measure with a relaxation factor of three, and the  $F_1$ -measure without relaxation version ( $\rho = 0$ ) is also reported. The  $F_1$ -measure can be written as:

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (10)$$

$$precision = \frac{tp}{tp + fp} \quad (11)$$

$$recall = \frac{tp}{tp + fn} \quad (12)$$

## 4. Experiments

### 4.1. Ablation Study

In this section, we first evaluate whether the two attention mechanisms can bring benefit to the segmentation performance, so we compare the base architecture (i.e., the standard U-Net) with the U-Net embedded with the attention mechanisms (U-Net-SCA). It should be noted that the U-Net-SCA is the segmentation network of the proposed GAN-SCA. In addition, we employ dense CRFs as the post-processor of U-Net-SCA to further improve the segmentation results (U-Net-SCA+CRFs). We also explore the recurrent approach to achieve label refinement followed the ReuseNet in [14] (U-Net-SCA+Reuse), that applies U-Net-SCA in  $R$  cycles. We choose  $R = 3$  in this experiment because the U-Net-SCA architecture in three cycles achieves the best performance on the Inria aerial image labeling dataset. Finally, we train the U-Net-SCA combined with an adversarial network (GAN-SCA) in an alternating fashion to see how the adversarial training affects the segmentation results.

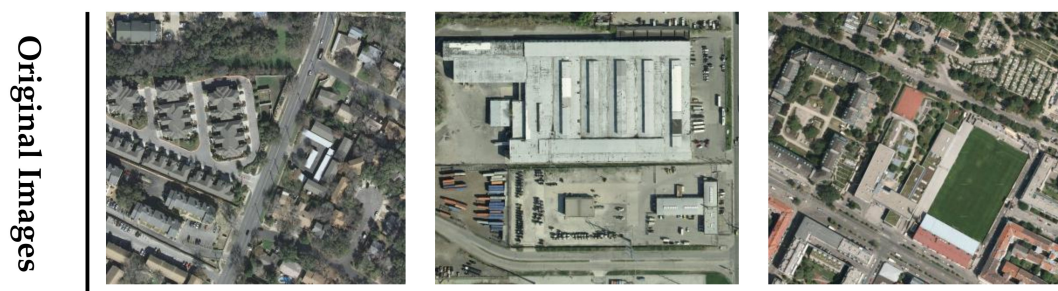
The models described above are trained over five independent runs with random initialization, and the average accuracy and IoU with the standard deviation of the experimental results on the validation set of the Inria aerial image labeling dataset are reported in Table 1. As can be observed from Table 1, the proposed U-Net-SCA achieves improvement of 0.19% and 0.72% in terms of the overall accuracy and IoU compared to the standard U-Net. For accuracy and IoU of each region, the U-Net-SCA also outperforms the standard U-Net. Especially for the regions in Chicago and Vienna, where buildings are high-densely distributed, and the proportion of building pixels in the training set is higher, the accuracy increase is more evident. This indicates that the spatial and channel attention mechanisms enable the network to selectively enhance useful features to further improve segmentation

accuracy. The U-Net-SCA+CRFs has few improvements over the U-Net-SCA, with the overall accuracy and IoU improved by 0.01% and 0.21%, respectively. By adopting the recurrent approach and adversarial network, the U-Net-SCA+Reuse and GAN-SCA have a similar small improvement of overall accuracy and IoU when compared to the U-Net-SCA. Let us recall that the adversarial training strategy adopted by the proposed GAN-SCA can learn high-order consistency without extra time consumption in the testing phase, whereas the recurrent approach of U-Net-SCA+Reuse is accompanied by the multi-fold increase of trainable weights which increases the computational complexity.

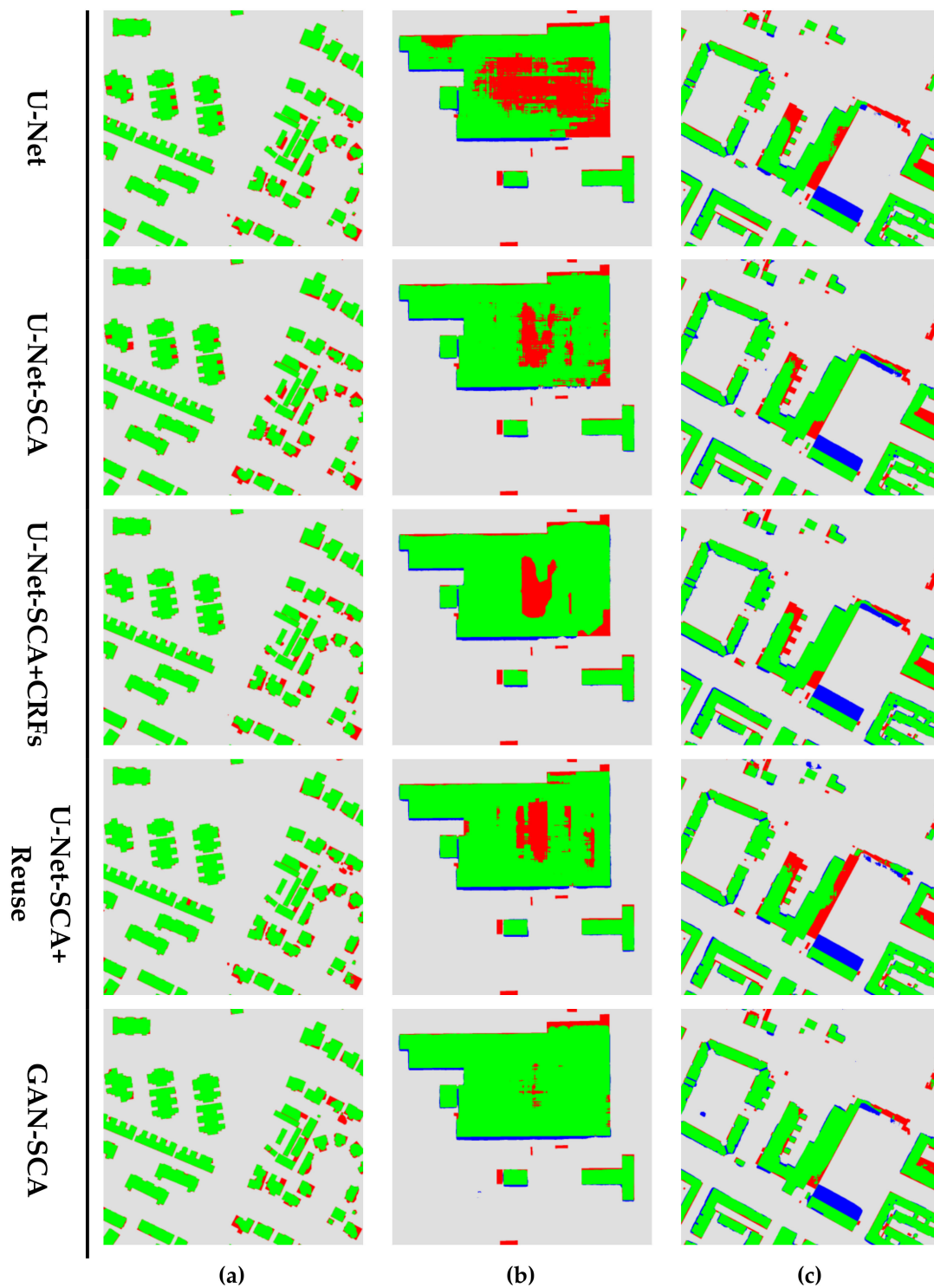
**Table 1.** Experimental results on Inria aerial image labeling dataset.

Methods	Metrics	Austin	Chicago	Kitsap	Tyrol-w	Vienna	Overall
U-Net	IoU	79.95 ± 0.81	70.18 ± 0.22	68.56 ± 1.49	76.29 ± 2.06	79.92 ± 0.39	76.16 ± 0.21
	Acc.	97.10 ± 0.11	92.67 ± 0.15	99.31 ± 0.03	98.15 ± 0.15	94.25 ± 0.18	96.31 ± 0.07
U-Net-SCA	IoU	80.40 ± 0.43	71.04 ± 0.70	68.25 ± 0.46	76.77 ± 1.86	80.55 ± 0.40	76.88 ± 0.42
	Acc.	97.22 ± 0.04	93.21 ± 0.10	99.30 ± 0.01	98.19 ± 0.15	94.57 ± 0.10	96.50 ± 0.05
U-Net-SCA +CRFs	IoU	80.36 ± 0.76	71.53 ± 0.15	68.40 ± 0.26	77.04 ± 1.99	80.83 ± 0.16	77.09 ± 0.22
	Acc.	97.17 ± 0.10	93.24 ± 0.06	<b>99.31 ± 0.01</b>	98.20 ± 0.15	94.62 ± 0.06	96.51 ± 0.05
U-Net-SCA +Reuse	IoU	80.76 ± 0.23	71.52 ± 0.30	68.08 ± 0.65	78.26 ± 0.51	81.36 ± 0.38	77.48 ± 0.30
	Acc.	97.22 ± 0.04	93.25 ± 0.10	99.30 ± 0.01	98.30 ± 0.03	94.78 ± 0.08	96.57 ± 0.07
GAN-SCA	IoU	<b>80.82 ± 0.21</b>	<b>71.37 ± 0.44</b>	<b>68.67 ± 0.18</b>	<b>78.68 ± 0.09</b>	<b>81.62 ± 0.26</b>	<b>77.52 ± 0.19</b>
	Acc.	<b>97.24 ± 0.03</b>	<b>93.32 ± 0.12</b>	<b>99.31 ± 0.01</b>	<b>98.33 ± 0.01</b>	<b>94.80 ± 0.06</b>	<b>96.60 ± 0.02</b>

Figure 5 shows the segmentation results of methods described above on the Inria aerial image labeling dataset. Figure 5a shows the results of an image patch over Austin, from which we can observe that the standard U-Net is affected by shadows and fail to segment the boundaries of complex structural buildings (upper left part of the figure) correctly. With the help of spatial and channel attention mechanisms, U-Net-SCA achieves better performance when dealing with the same situation, but still mislabels some non-building pixels in shadows as building. The extraction results from U-Net-SCA+CRFs, U-Net-Reuse, and GAN-SCA all seem to have clearer boundaries, especially for complex structural buildings, and this is due to the adopted different label refinement method. In contrast, the extraction result of the proposed GAN-SCA achieves clearer and more accurate outlines of this kind of buildings. In addition, some large buildings are difficult to labeled correctly, due to their edges on the rooftop which have a similar color to roads. Figure 5b shows the results of a large building in the Chicago city, where the results of U-Net suffer from over-segmentation of the inner edges of the detected building, while U-Net-SCA improves the results by using the channel and spatial attention mechanisms to selectively enhance useful features. U-Net-SCA+CRFs smooths the results, yet the improvements seem insignificant. The result of U-Net-SCA+Reuse also shows a slight improvement, but the over-segmentation has not been effectively solved. In contrast, the proposed GAN-SCA labeled the large building more completely. Moreover, buildings with complex shape and multiple colors are prone to be confused by the networks, as shown in the middle of Figure 5c, and most methods mislabel this kind of building as non-buildings, while the proposed GAN-SCA can provide a relatively proper segmentation results.



**Figure 5.** Cont.



**Figure 5.** Building extraction results for three image patches of Inria aerial image labeling dataset. (a) Image patch over Austin; (b) Image patch over Chicago; (c) Image patch over Vienna. Green: true positive (tp) pixels; Gray: true negative (tn) pixels; Blue: false positive (fp) pixels; Red: false negative (fn) pixels.

## 4.2. Comparison to State-of-the-Art Methods

### 4.2.1. Inria Aerial Image Labeling Dataset

To evaluate the performance on the Inria aerial image labeling dataset, we compare the proposed GAN-SCA (best results we achieved) with some state-of-the-art methods, including the baseline method FCN [23], multi-layer perceptron (MLP) [23], Mask R-CNN [32] performed by Ohleyer et al. [33], SegNet+Multi-Task Loss [24], 2-levels U-Nets [25], and the multi-stage multi-task (MSMT) [34]. FCN and MLP are frameworks proposed by the producers of the Inria aerial image labeling dataset. MLP derived from the base FCN and introduced a multi-layer perceptron to learn how to combine features at different resolutions. Mask R-CNN consisted of a region proposed network (RPN) and an FCN, the RPN took the whole image as input and output the image with bounding box proposals. According to the proposal of RPN, the FCN then performed efficient segmentation. The SegNet+Multi-Task Loss was based on SegNet architecture and trained with an uncertainty based multi-task loss. In particular, one convolutional layer  $L$  was followed after the last layer of the decoder to generate the distance classes, and then the output of decoder's last layer was concatenated with the output of  $L$  to predict the final segmentation results. 2-Levels U-Nets was proposed in [25], where two U-Net architectures were arranged end-to-end, and the last U-Net was served as the post-processor to the first one. Moreover, the test time augmentation was applied to further improve the segmentation performance. The MSMT architecture was proposed in [34]. Authors proposed an MSMT neural network which had two stages, namely semantic segmentation and localization. The first stage was dedicated to semantic segmentation, while the second stage was designed for localization.

Table 2 presents the accuracy and IoU of different methods on the Inria aerial image labeling dataset. It is worth noting that IoU can take into account both the false alarms and the missing detections that is a more suitable metric than global accuracy on Inria dataset, because this dataset contains large areas of background pixels. It can be seen from Table 2 that MLP outperforms the base FCN [23] by introducing multi-layer perceptron to fuse multi-resolution features. Mask-RCNN is a promising architecture, but it requires very good hyperparameters tuning [33]. Therefore, it achieves better performance in Austin and Tyrol-w but lower in most regions when compared to MLP. SegNet+Multi-Task Loss improves the performance of SegNet by introducing a cascaded multi-task loss, but the improvement is still limited. Although it achieves the best accuracy in regions of Chicago and Vienna, the corresponding IoU is not ideal. 2-Levels U-Nets and MSMT achieve similar accuracy, of which the former approach outperforms the latter one in terms of IoU in all regions. This is mainly because the 2-Levels U-Nets is based on U-Net which is a deeper architecture than that of MSMT. The proposed GAN-SCA is also on top of U-Net. With the help of the attention mechanisms and adversarial training strategy, GAN-SCA outperforms 2-Levels U-Nets in most evaluation metrics and produces the highest IoU in most regions, especially the densely populated cities, such as Austin, Chicago, and Vienna. In terms of the overall accuracy and IoU, the proposed method surpasses all other methods by a considerable margin, which shows that the proposed method can accomplish accurate building segmentation. The qualitative results of the GAN-SCA are shown in Figure 6. It can be seen that the GAN-SCA achieves accurate building segmentation results in each region with smooth outlines.

**Table 2.** Experimental results on Inria aerial image labeling dataset.

Methods	Metrics	Austin	Chicago	Kitsap	Tyrol-w	Vienna	Overall
FCN [23]	IoU	47.66	53.62	33.70	46.86	60.60	53.82
	Acc.	92.22	88.59	98.58	95.83	88.72	92.79
MLP [23]	IoU	61.20	61.30	51.50	57.95	72.13	64.67
	Acc.	94.20	90.43	98.92	96.66	91.87	94.42
Mask R-CNN [32]	IoU	65.63	48.07	54.38	70.84	64.40	59.53
	Acc.	94.09	85.56	97.32	98.14	87.40	92.49



Table 2. Cont.

Methods	Metrics	Austin	Chicago	Kitsap	Tyrol-w	Vienna	Overall
SegNet+Multi-Task Loss [24]	IoU	76.76	67.06	<b>73.30</b>	66.91	76.68	73.00
	Acc.	93.21	<b>99.25</b>	97.84	91.71	<b>96.61</b>	95.73
2-Levels U-Nets [25]	IoU	77.29	68.52	72.84	75.38	78.72	74.55
	Acc.	96.69	92.40	99.25	98.11	93.79	96.05
MSMT [34]	IoU	75.39	67.93	66.35	74.07	77.12	73.31
	Acc.	95.99	92.02	99.24	97.78	92.49	96.06
GAN-SCA	IoU	<b>81.01</b>	<b>71.73</b>	68.54	<b>78.62</b>	<b>81.62</b>	<b>77.75</b>
	Acc.	<b>97.26</b>	93.32	<b>99.30</b>	<b>98.32</b>	94.84	<b>96.61</b>



**Figure 6.** Building extraction results of Inria aerial image labeling dataset. (a) Image patch over Austin; (b) Image patch over Chicago; (c) Image patch over Vienna. Green: true positive (tp) pixels; Gray: true negative (tn) pixels; Blue: false positive (fp) pixels; Red: false negative (fn) pixels.

#### 4.2.2. Massachusetts Buildings Dataset

We tested the performance of the proposed GAN-SCA on the Massachusetts buildings dataset by using the same metrics as the compared methods. We compared the performance of GAN-SCA with several state-of-the-art methods including Mnih-CNN+CRFs [22], Satio-multi-MA&CIS [21], LG-Seg-ResNet-IL [35], and MTMS [34]. The Mnih-CNN+CRF was proposed by the producers of the Massachusetts building dataset, which belonged to the patch-based category, and CRFs was included as a post-processor. Satio-multi-MA&CIS was based on Mnih-CNN architecture, in which channel-wise inhibited softmax (CIS) loss function and modeled averaging (MA) techniques were used to further enhance the extraction performance. LG-Seg-ResNet-IL is a dual local-global semantic segmentation architecture with residual connections and an intermediate contextual loss (IL), which learned to combine local appearance and global contextual information simultaneously in a complementary way. MTMS is the same method described in Section 4.2.1.

Table 3 compares the  $F_1$ -measure of each method, in which  $\rho$  denotes the relaxed factor when computing the corresponding recall and precision measures. As shown in Table 3, our GAN-SCA obtains a superior performance than all other methods. With the help of the (CIS) loss function and (MA) with spatial displacement, the Satio-multi-MA&CIS achieves a slight improvement compared to the baseline method Mnih-CNN+CRFs. LG-Seg-ResNet-IL effectively combines the local and global information, which mitigates the problem of the limited receptive field of the patch-based method. So LG-Seg-ResNet-IL achieves a remarkable improvement compared to the first two methods. MTMS is an FCN-based method that introduces a multi-stage multi-task training strategy to enhance segmentation performance. MTMS and GAN-SCA achieve better performance compared to the patch-based methods, which indicates the superiority of the pixel-based method. Thanks to the deeper architecture and the feature selection by adopting attention mechanisms, the GAN-SCA exhibits better performance when compared to MTMS, which further indicates the rationality of the proposed method. Figure 7 exhibits the prediction results of the proposed model for three image patches. It can be seen that our proposed model presents a satisfying performance in challenging areas.

**Table 3.** Experimental results on Massachusetts buildings dataset.

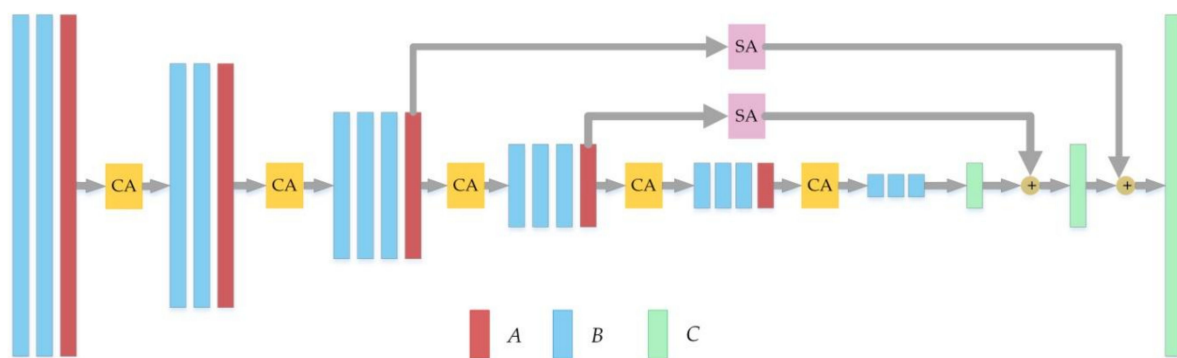
Method	F <sub>1</sub> -Measure	
	$\rho = 0$	$\rho = 3$
Mnih-CNN+CRFs [22]	-	92.11%
Satio-multi-MA&CIS [21]	-	92.30%
LG-Seg-ResNet-IL [35]	-	94.30%
MTMS [34]	83.39%	96.04%
GAN-SCA	84.79%	96.36%



**Figure 7.** Building extraction results on the Massachusetts buildings dataset. (a–c) prediction results of three image patches in Massachusetts buildings dataset. Green: true positive (tp) pixels; Gray: true negative (tn) pixels; Blue: false positive (fp) pixels; Red: false negative (fn) pixels.

### 4.3. Experiments on FCN based GAN-SCA

The experiments above adopted U-Net as the baseline of the segmentation network for the proposed GAN-SCA, and achieve a certain improvement when compared with the standard U-Net. In fact, our proposed GAN-SCA can be realized on the top of many other semantic segmentation architectures. In this section, we will explore the GAN-SCA on top of FCN-8s version to further demonstrate the effectiveness of the attention mechanisms and adversarial training in building extraction from high-resolution remote sensing images. Figure 8 shows the architecture of the segmentation network (FCN-8s-SCA), where the channel and spatial attention mechanisms are embedded into the FCN-8s architecture with the VGG-16 [36] architecture as an encoder. Same as the U-Net based GAN-SCA described above, the adversarial network of this version is followed by the encoder of its segmentation network.



**Figure 8.** Architecture of FCN-8s-SCA. A is max pooling layer; B are convolutional + Rectified Linear Unit (ReLU) layers; C is the transpose convolutional layer; SA is the spatial attention mechanism; CA is the channel attention mechanism; RS is the reshape operation.

We train the FCN-8s, FCN-8s-SCA, and GAN-SCA on the Inria aerial image labeling dataset using the same training strategy as introduced in Section 2.2. The experimental results are reported in Table 4. Compared with the FCN-8s, the FCN-8s-SCA improved the overall accuracy and IoU by 0.49% and 3.71%, respectively. For the adversarial training strategy, the FCN-8s based GAN-SCA further improved the extraction performance by 0.48% and 3.16% for the overall accuracy and IoU, respectively. We can conclude that the attention mechanisms can improve the segmentation performance by feature selection, and adversarial training can further refine the segmentation result by learning high-order consistency. In addition, the improvement of FCN-8s based GAN-SCA is more significant than the aforementioned U-Net based GAN-SCA. This is because the standard U-Net architecture has already achieved remarkable segmentation performance, as it fused high-level and low-level feature by first concatenating features together and then performing convolutions for dimensionality reduction. The convolutional layers in this process enable the network to learn how to fuse multi-scale features which can be regarded as feature selection to some extent. While FCN-8s has lower segmentation accuracy on this dataset when compared to the standard U-Net, it fused features by adopting element-wise addition, which seems unsuitable without any feature selection. Therefore, FCN-8s can take more advantage of the attention mechanisms.



**Table 4.** Experimental results on Inria aerial image labeling dataset.

Methods	Metrics	Austin	Chicago	Kitsap	Tyrol-w	Vienna	Overall
FCN-8s	IoU	67.98	63.43	53.17	68.13	72.03	68.05
	Acc.	95.47	91.41	99.01	97.52	92.38	95.16
FCN-8s-SCA	IoU	72.85	69.61	64.97	73.20	73.26	71.76
	Acc.	96.19	92.36	99.25	97.94	92.49	95.65
GAN-SCA	IoU	<b>78.51</b>	<b>70.10</b>	<b>66.42</b>	<b>76.84</b>	<b>77.24</b>	<b>74.92</b>
	Acc.	<b>96.90</b>	<b>92.86</b>	<b>99.27</b>	<b>98.14</b>	<b>93.46</b>	<b>96.13</b>

## 5. Discussion

The experimental results reported in Section 4 prove that the proposed approach achieved state-of-the-art performance on both Inria and Massachusetts buildings datasets. Furthermore, the GAN-SCA can also be employed on top of other semantic segmentation architectures with better performance. The effectiveness of our proposed method comes from the feature selection in spatial and channel dimensions, and the label refinement by learning high-order structural features. First, the adoption of spatial and channel attention mechanisms helps with enhancing the useful features while suppressing the interference information, improving the segmentation performance around building borders, and mitigating over-segmentation. Second, the adversarial training strategy learns the latent high-order structural information in the training phase and achieves label refinement in the testing phase without extra time consumption. Especially, the segmentation network and adversarial network of our architecture were optimized by multi-scale feature loss to better capture multi-range spatial relationships between pixels. These factors make the proposed GAN-SCA have a better feature extraction capability and better segmentation performance.

Although the proposed approach performs well as a fully supervised method, it relies on a large number of manual labeling samples. Further researches are needed to alleviate the task of manual annotation. Possible directions that can be explored include data augmentation techniques and adversarial learning for semi-supervised semantic segmentation. Data augmentation techniques can increase the number of training samples and improve the generalization ability of models. We have explored some standard data augmentation techniques including flip and rotation in this work and previous works to mitigate overfitting. More data augmentation strategies will be explored in our future work. In addition, adversarial learning for semi-supervised semantic segmentation is also an interesting research direction, which can take advantage of unlabeled data to generate self-taught signal to refine the segmentation network. These approaches will be highly relevant in fields, such as remote sensing images analysis, in which large datasets are expensive to obtain.

## 6. Conclusions

This paper presented an effective GAN-based approach for building extraction from high-resolution remote sensing images. The adopted architecture consists of two parts: the segmentation network and the adversarial network, which are, in turn, used to generate segmentation maps of buildings and to discriminate the ground truths and the predicted results of the segmentation network, respectively. To enable the segmentation network to focus on more useful information, spatial and channel attention mechanisms are embedded into the standard U-Net. The adversarial network architecture is similar to the encoder of the segmentation network, where the extracted multi-layer features are considered when computing the multi-scale  $L_1$  loss in the adversarial training phase.

The experiments were conducted on the Inria aerial image labeling dataset for buildings as well as the Massachusetts buildings dataset. The experimental results show that the spatial and channel attention mechanisms can selectively enhance useful features to improve the segmentation performance, while adversarial training can further refine the segmentation results with little time consumption during the testing stage. Compared with the state-of-the-art methods on both the datasets,

the proposed GAN-SCA achieved higher overall accuracy (96.61%), IoU (77.75%) for the Inria aerial image labeling dataset and  $F_1$ -Measure (96.36%) for the Massachusetts buildings dataset. Especially for samples with dense-distributed buildings, the improvement was more evident.

In future studies, we will explore adversarial network architecture optimization and loss function improvement to take full advantage of the adversarial training. We will also research the data augmentation techniques and semi-supervised semantic segmentation.

**Author Contributions:** Conceptualization, F.Y.; Methodology, X.P. and L.G.; Software, H.F.; Supervision, B.Z.; Validation, Z.C. and H.F.; Writing-Original Draft Preparation, X.P.; Writing-Review & Editing, L.G., J.R. and Z.C.

**Funding:** This research was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant No. XDA19080302, and by the National Natural Science Foundation of China under Grant No. 91638201.

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

## References

1. Rees, W.G. *Physical Principles of Remote Sensing*; Cambridge University Press: Cambridge, UK, 2013.
2. Alshehhi, R.; Marpu, P.R.; Wei, L.W.; 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](#)]
3. Yang, H.L.; Yuan, J.; Lunga, D.; Laverdiere, M.; Rose, A.; Bhaduri, B. Building extraction at scale using convolutional neural network: Mapping of the United States. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 2600–2614. [[CrossRef](#)]
4. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
5. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015; pp. 234–241.
6. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *Computer Sci.* **2014**, *4*, 357–361.
7. 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.* **2016**, *40*, 834–848. [[CrossRef](#)] [[PubMed](#)]
8. Noh, H.; Hong, S.; Han, B. Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 13–16 December 2015; pp. 1520–1528.
9. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [[CrossRef](#)] [[PubMed](#)]
10. Liu, Y.; Piramanayagam, S.; Monteiro, S.; Saber, E. Dense semantic labeling of very-high-resolution aerial imagery and LiDAR with fully-convolutional neural networks and higher-order crfs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 1561–1570.
11. Liu, Y.; Minh Nguyen, D.; Deligiannis, N.; Ding, W.; Munteanu, A. Hourglass-shape network based semantic segmentation for high resolution aerial imagery. *Remote Sens.* **2017**, *9*, 522. [[CrossRef](#)]
12. Pan, X.; Gao, L.; Marinoni, A.; Zhang, B.; Yang, F.; Gamba, P. Semantic labeling of high resolution aerial imagery and LiDAR data with fine segmentation network. *Remote Sens.* **2018**, *10*, 743. [[CrossRef](#)]
13. Maggiori, E.; Charpiat, G.; Tarabalka, Y.; Alliez, P. Recurrent Neural Networks to Correct Satellite Image Classification Maps. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 4962–4971. [[CrossRef](#)]
14. Bergado, J.R.; Persello, C.; Stein, A. “Recurrent Multiresolution Convolutional Networks for VHR Image Classification,”. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 6361–6374. [[CrossRef](#)]
15. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, Qc, Canada, 8–13 December 2014; pp. 2672–2680.



16. Luc, P.; Couprie, C.; Chintala, S.; Verbeek, J. Semantic segmentation using adversarial networks. *arXiv*, 2016; arXiv:1611.08408. Available online: <https://arxiv.org/abs/1611.08408>(accessed on 1 April 2018).
17. Xue, Y.; Xu, T.; Zhang, H.; Long, L.R.; Huang, X. SegAN: Adversarial network with multi-scale  $L_1$  loss for medical image segmentation. *Neuroinformatics* **2017**, *6*, 1–10. [[CrossRef](#)] [[PubMed](#)]
18. Pan, X.; Gao, L.; Zhang, B.; Yang, F.; Liao, W. High-resolution aerial imagery semantic labeling with dense pyramid network. *Sensors* **2018**, *18*, 3774. [[CrossRef](#)] [[PubMed](#)]
19. Sherrah, J. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *arXiv*, 2016; arXiv:1606.02585v1. Available online: <https://arxiv.org/abs/1606.02585>(accessed on 1 April 2018).
20. Volpi, M.; Tuia, D. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. *IEEE Trans. on Geosci. Remote Sens.* **2017**, *55*, 881–893. [[CrossRef](#)]
21. Saito, S.; Yamashita, T.; Aoki, Y. Multiple Object Extraction from Aerial Imagery with Convolutional Neural Networks. *Electronic Imag.* **2016**, *60*, 10401–10402.
22. Mnih, V. Machine Learning for Aerial Image Labeling. Ph.D. Thesis, University of Toronto, Toronto, ON, Canada, 2013.
23. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Can semantic labeling methods generalize to any city? The Inria aerial image labeling benchmark. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 3226–3229.
24. Bischke, B.; Helber, P.; Folz, J.; Borth, D.; Dengel, A. Multi-task learning for segmentation of building footprints with deep neural networks. *arXiv*, 2017; arXiv:1709.05932. Available online: <https://arxiv.org/abs/1709.05932>(accessed on 27 April 2018).
25. Khalel, A.; El-Saban, M. Automatic pixelwise object labeling for aerial imagery using stacked u-nets, arXiv 2018, arXiv:1803.04953. Available online: <https://arxiv.org/abs/1803.04953> (accessed on 27 April 2018).
26. Itti, L.; Koch, C. Computational modelling of visual attention. *Nat. Rev. Neurosci.* **2001**, *2*, 194–203. [[CrossRef](#)] [[PubMed](#)]
27. 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](#)]
28. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
29. Glorot, X.; Bordes, A.; Bengio, Y. Deep Sparse Rectifier Neural Networks. In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), Fort Lauderdale, FL, USA, 11–13 April 2011; pp. 315–323.
30. Kingma, D.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations (ICLR), Banff, AB, Canada, 14–16 April 2014.
31. Audebert, N.; Le Saux, B.; Lefèvre, S. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In Proceedings of the Asian Conference on Computer Vision (ACCV), Taipei, Taiwan, 21–23 November 2016; pp. 180–196.
32. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–3288.
33. Building segmentation on satellite images. Available online: [https://project.inria.fr/aerialimagelabeling/files/2018/01/fp\\_ohleyer\\_compressed.pdf](https://project.inria.fr/aerialimagelabeling/files/2018/01/fp_ohleyer_compressed.pdf) (accessed on 1 April 2018).
34. Marcu, A.; Costea, D.; Slusanschi, E.; Leordeanu, M. A Multi-stage Multi-task neural network for aerial scene interpretation and geolocalization. *arXiv*, 2018; arXiv:1804.01322v1. Available online: <https://arxiv.org/abs/1804.01322>(accessed on 27 April 2018).
35. Marcu, A.; Leordeanu, M. Object contra context: Dual local-global semantic segmentation in aerial images. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17), San Francisco, CA, USA, 4–9 February 2017; pp. 146–152.
36. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Machine Learning (ICML), San Diego, CA, USA, 7–9 May 2015.

