remote sensing

MDPI

Article

Self-Attention in Reconstruction Bias U-Net for Semantic Segmentation of Building Rooftops in Optical Remote Sensing Images

Ziyi Chen [1], Dilong Li [1], Wentao Fan [1], Haiyan Guan [2], Cheng Wang [3] and Jonathan Li [4,*]

[1] Fujian Key Laboratory of Big Data Intelligence and Security, Xiamen Key Laboratory of Computer Vision and Pattern Recognition, Department of Computer Science and Technology, Huaqiao University, Xiamen 361021, China; chenziyihq@hqu.edu.cn (Z.C.); scholar.dll@hqu.edu.cn (D.L.); fwt@hqu.edu.cn (W.F.)
[2] School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China; guanhy.nj@nuist.edu.cn
[3] School of Informatics, Xiamen University, Xiamen 361005, China; cwang@xmu.edu.cn
[4] Department of Geography and Environmental Management and Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada
* Correspondence: junli@uwaterloo.ca

Abstract: Deep learning models have brought great breakthroughs in building extraction from high-resolution optical remote-sensing images. Among recent research, the self-attention module has called up a storm in many fields, including building extraction. However, most current deep learning models loading with the self-attention module still lose sight of the reconstruction bias's effectiveness. Through tipping the balance between the abilities of encoding and decoding, i.e., making the decoding network be much more complex than the encoding network, the semantic segmentation ability will be reinforced. To remedy the research weakness in combing self-attention and reconstruction-bias modules for building extraction, this paper presents a U-Net architecture that combines self-attention and reconstruction-bias modules. In the encoding part, a self-attention module is added to learn the attention weights of the inputs. Through the self-attention module, the network will pay more attention to positions where there may be salient regions. In the decoding part, multiple large convolutional up-sampling operations are used for increasing the reconstruction ability. We test our model on two open available datasets: the WHU and Massachusetts Building datasets. We achieve IoU scores of 89.39% and 73.49% for the WHU and Massachusetts Building datasets, respectively. Compared with several recently famous semantic segmentation methods and representative building extraction methods, our method's results are satisfactory.

Keywords: building extraction; U-Net; remote sensing image; building footprint

Citation: Chen, Z.; Li, D.; Fan, W.; Guan, H.; Wang, C.; Li, J. Self-Attention in Reconstruction Bias U-Net for Semantic Segmentation of Building Rooftops in Optical Remote Sensing Images. *Remote Sens.* **2021**, *13*, 2524. https://doi.org/10.3390/rs13132524

Academic Editor: Pedram Ghamisi

Received: 3 May 2021
Accepted: 25 June 2021
Published: 28 June 2021

1. Introduction

Building footprints play an important role in many applications, ranging from urban planning [1], population estimation [2], disaster management [3], land dynamic analysis, to illegal building-construction recognition [4]. Given the rapid development in automatic semantic segmentation technology in computer vision, traditional manual building extraction and labeling works have been released greatly. Automated extraction of buildings from high-resolution optical remote sensing images is a very active research topic in both computer-vision and remote-sensing communities and has made substantial progress [4–13].

Currently, the most popular approach to building extraction seems to be deep-learning based methods. In 2012, Hinton proposed a deep convolutional neural network (CNN) and won the championship in the competition in ImageNet [14]. Since then, deep CNN networks became an instant hit worldwide. It is amazing that many researchers found that deep CNNs can obtain much better performance compared with traditional manually

designed feature-based methods in many computer vision tasks, e.g., image classification [15], target detection and recognition [16–18], target tracking [19], semantic segmentation [20], retrieval [21], and so on. Following the progress and developments in computer vision, building-extraction applications also used the most popular network structures and achieved desirable results [6,22–24]. Recently, channel attention led to a revolution in network architecture [25,26]. Not surprisingly, given its very promising performance, self-attention is widely used in many computer vision tasks, including building extraction [5,6].

However, most end-to-end building extraction models utilize symmetrical structure between encoding and decoding parts. As up-sampling operations are a more challenging task than encoding down-sampling feature extraction jobs, it may lead to an imbalance between decoding ability and encoding ability. For now, the related research about reinforcing the decoding network and using an asymmetrical structure between encoding and decoding parts is fair. According to the logical deduction, the performance may be improved with a more complex decoding structure. Furthermore, the research about combining the self-attention module and reconstruction-bias module is fresh. Thus, to explore the reconstruction-bias idea for building extraction and remedy the research deficient in combining self-attention and reconstruction-bias modules, this paper proposes a model which combines self-attention and reconstruction-bias modules for building extraction from remote-sensing images.

In our model, we first use a transformer module to learn the channel weights. Then, the channel weights are multiplied with inputs as the input of the next network layer. In the decoding part, we apply multiple scales large convolutions in each up-sampling layer to increase the reconstruction ability. We test our model on two publicly open datasets: the WHU dataset [27] and Massachusetts Building dataset [28]. Compared with several recently famous semantic segmentation methods and classical building extraction methods, our results are satisfactory.

The contributions of this paper can be summarized as:

The asymmetrical network architecture has been explored for automated building textraction.

Self-attention has been combined with the reconstruction-bias strategy for automated building extraction.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 details the datasets and our network. Section 4 illustrates the results and analysis. Section 5 details the discussion. Section 6 concludes the paper.

2. Related Work

The methods for automated extraction of buildings from very high resolution (VHR) optical remote-sensing images can be categorized into three types: morphological and geometrical feature-based [29–33], manually designed feature-based [34–36] and deep learning-based [1,6,37–39].

2.1. Morphological and Geometrical Feature-Based

For morphological and geometrical feature-based methods, the morphological and geometrical features are used as the criteria for building extraction. The morphological and geometrical features usually contain shapes, lines, length and width, etc. As morphological and geometrical features are simple and easy to be modeled for buildings in visual, morphological and geometrical features have been widely used and achieved large amounts of research success. A novel adaptive morphological attribute profile based on object boundary constraint was proposed in [29] for building extraction from high-resolution remote-sensing images. Their model was tested on groups of images from different sensors and showed good results. A building extraction approach was proposed in [30], which is based on morphological attributes' relationships among different morphological features (e.g., shape, size). They assessed their method on three VHR datasets and demonstrated good results. A method that combined CNN and morphological filters was presented

in [40] for building extraction from VHR images. In their method, the morphological features were used for final extraction filtering after the extraction of CNN. The experiments proved that their method is effective. The morphological features and support vector machine (SVM) were used in [31] for building extraction from VHR images. They tested their method on WorldView-2 and Sentinal-2 images and demonstrated good F1-scores.

Although morphological and geometrical features are simple and easy for using, these kinds of features usually suffer from the problems of the rigid model and the sensitivity to image resolution, occlusions' interference, etc.

2.2. Manually Designed Feature-Based

For manually designed feature-based methods, the researchers usually use transformations to extract features and then combine the extracted features with classifiers for the final building extraction task. The typical classifiers include SVM, Hough Forest, TensorVoting, Random Forest, etc. Since the manually designed features have shown superiority to morphological features in robustness to occlusions, brightness changing, resolution changing and imaging perspective changing, etc., the manually designed feature-based methods became popular in the past 20 years. A building extraction approach from high-resolution optical satellite images was proposed in [34] and achieved quite good and impressive results. In their method, the SVM, Hough transformation and perceptual grouping were combined. The Hough transformation was used for delineating circular-shape buildings, while the perceptual grouping strategy was used for constructing the building boundaries through integrating the detected lines. A hybrid approach to building extraction was proposed in [35], which used a template matching strategy for automatically computing the relative height of buildings. After estimating the relative height of buildings, the SVM-based classifier was employed to recognize the buildings and non-buildings, thus extracting the buildings. They tested on images of WorldView-2 and achieved high building-detection accuracy.

The manually designed feature-based methods usually can extract the classical features of the buildings, and the buildings can be extracted with quite high accuracy through combining the classifiers. However, the models' extendibility is still weak due to the brightness variations, occlusions, etc. The main reason may be that the manually designed features cannot cover all the building appearance situations in the images, resulting in incomplete considerations for special situations.

2.3. Deep Learning-Based

Recently, deep learning-based building extraction methods have made great breakthroughs. The classical models usually extract the buildings with an end-to-end strategy, i.e., input a target image and output a building extraction result image. The benefits of the deep learning models lay in the great powers of automatic feature learning and representing. In addition, the deep learning-based methods also can obtain results with fast processing speed through using GPUs. The processing time of deep learning-based building extraction methods is usually only several seconds to produce the final results (sometimes even within only 1 second), while the unsupervised and manually designed methods usually take dozens of minutes (or even several hours) for processing one image.

A single path-based CNN model was proposed in [41] for simultaneously extracting roads and buildings from remote-sensing images. After the extraction of the CNN model, the low-level features of roads and buildings were also combined to improve performance. They tested their model on two challenging datasets and demonstrated good extraction results. A Building-A-Nets for building extraction was proposed in [42,43], in which the adversarial network architecture was applied and they jointly trained their model of generator and discriminator. They tested on open available datasets and achieved good results. A building extraction model of fully convolutional network (FCN) was proposed in [43]. To further improve the final results, the Conditional Random Fields were employed. They obtained high F1-scores and the intersection of union (IoU) scores in their experiments.

A new deep learning model based on ResNet was proposed in [44], which used the specially designed guided filters to improve their results and remove the salt-and-pepper noise. The method illustrated good performance in the tests. A deep CNN model was proposed in [45], which integrated activation from multiple layers and introduced a signed distance function for representing building boundary outputs. They demonstrated superior performance on test datasets. A deep learning model was proposed in [46], which aimed to conquer the problems of sensitivity to unavoidable noise and interference, and the insufficient use of structure information. They showed good results on the test datasets. A Siamese fully convolutional network was proposed in [27] for building extraction and provided an open dataset called WHU that contained multiple data sources.

Now the WHU dataset is quite famous in open available building extraction datasets and has been used in much building extraction research. An EU-Net for building extraction was proposed in [47] that designed a dense spatial pyramid pooling module. They achieved quite good results in the test datasets. In [48], a DE-Net that consisted of four modules (the inception-style down-sampling module, the encoding module, the compressing module and the densely up-sampling module). They tested the model on an open available dataset and a self-built dataset called Suzhou. The test results showed good performance of their model. Liu et al. proposed a building extraction model that used a spatial residual inception module to obtain multiscale contexts [49]. In addition, they used depthwise separable convolutions and convolution factorization to further improve the computational efficiency. In [13], a JointNet was proposed to extract both large and small targets using a wide receptive field, and it used focal loss function to further improve the road extraction performance. In [50], an FCN was proposed to use multiscale aggregation of feature pyramids to enhance the scale robustness. After the segmentation results were obtained, a polygon regularization approach was further used for vectorizing and polygonizing the segmentation results. In [40], a multifeature CNN was proposed to extract building outlines. To improve the boundary regularity, they also combined morphological filtering in the post-processing. They achieved good results in the experiments. In [51], a CNN model was proposed that used an improved boundary-aware perceptual loss for building extraction, and their experimental results were very promising. The DR-Net, a dense residual network presented in [1] showed promising results in the test datasets. The attention-gate-based encoder-decoder network was used in [5] for building extraction, and illustrated good performance in both an open available dataset and the dataset built by themselves.

Except for the methods specially designed for building extraction from remote-sensing images, the segmentation methods for natural scenes are also suitable for building extraction from remote-sensing images. Thus, we also give a brief introduction about segmentation methods for natural scenes. The classical segmentation methods based on deep learning include FCN [52], PSPNet [20], U-Net [53,54] , DANet [55] and Residual U-Net [53] etc. In addition to the classical segmentation methods, many recent semantic segmentation methods have been proposed. Chen et al. proposed a DeepLab for semantic segmentation, and they achieved good experimental results [56]. Zhong et al. proposed a Squeeze-and-Attention Network for semantic segmentation. They achieved good results on two challenging public datasets. Zhang et al. used an encoding part that extracts multiscale contextual features for semantic segmentation, and they showed good results in the experiments [57]. Yu et al. proposed a CPNet (Context Prior Network) for learning robust context features in semantic segmentation tasks, and they showed good results in the experiments [58].

In general, the deep learning-based methods have achieved great progress in building extraction from remote-sensing images. However, their main shortcoming is the requirement of a large amount of labelling work. On the other hand, the research about a more powerful decoding part of the designed models is still insufficient in our reviews.

3. Data and Method

In this section, we first illustrate the datasets used in this paper. Then, we give an introduction about our model architecture. Third, we give a detailed presentation about the transformation module. Finally, we illustrate the detailed module about the decoding in our model.

3.1. Datasets

In this paper, two publicly available datasets are employed for training and evaluating in our experiments: the WHU dataset [27] and the Massachusetts Building dataset [59].

The WHU dataset was publicly opened in 2019 and has become a famous and popular building extraction dataset in remote-sensing research. The dataset contains more than 2.2 million independent buildings in aerial images. The resolution of the aerial images is 7.5 cm, which is so high and it makes the features of buildings clear. The cover area of WHU is about 450 km^2, covering Christchurch, New Zealand. One major advantage of WHU is that WHU contains various and versatile architectural types of buildings in different location areas (countryside, residential, culture, and industrial areas) and with different appearances (different colors and sizes). To make the aerial image with a resolution of 7.5 cm be more suitable for building extraction, the final images are down-sampled to a resolution of 0.3 m. The original image size in WHU is 512 $\times$ 512 pixels. Thus, the WHU contains 4736, 1036 and 2416 training, validation and testing images, respectively. In our experiment, as we design our model input size as 256 $\times$ 256 pixels, we further down-sample each original WHU image into 256 $\times$ 256 pixels. Figure 1 shows several sample images for our experiments in training, validation and testing stages, respectively. As shown in Figure 1, we can see that the image quality in WHU is high and the backgrounds are complex.

The second dataset used in this paper is the Massachusetts Building dataset [59]. The dataset contains 137, 4 and 10 images for training, validation and testing, respectively. The size of the original images is 1500 $\times$ 1500 pixels. The authors did not provide the detailed resolution information about the images. It is obvious that the resolution of the Massachusetts Building dataset is much lower than the resolution of WHU. To roughly estimate the resolution of the Massachusetts Building dataset, we compared with the satellite images. After estimation, we found that the rough resolution of the Massachusetts Building dataset is about 1.5 meter. Each image in the Massachusetts Building dataset has a large image size, which is not suitable for directly using in training, validation and testing. Thus, we cut each image into small pieces with an image size of 256 $\times$ 256 pixels, which is just fit for our model. Note that we cut into small pieces without overlapping. Finally, we generated 3425, 100 and 250 images for training, validation and testing, respectively. Figure 2 shows several original images in the Massachusetts Building dataset, while Figure 3 shows several cut 256 $\times$ 256 pixels sample images used in our experiments. As shown in Figures 2 and 3, we can know that the image quality of the Massachusetts Building dataset is lower compared with the WHU dataset. Moreover, comparative brightness in the WHU and the Massachusetts Building dataset is also rather different.

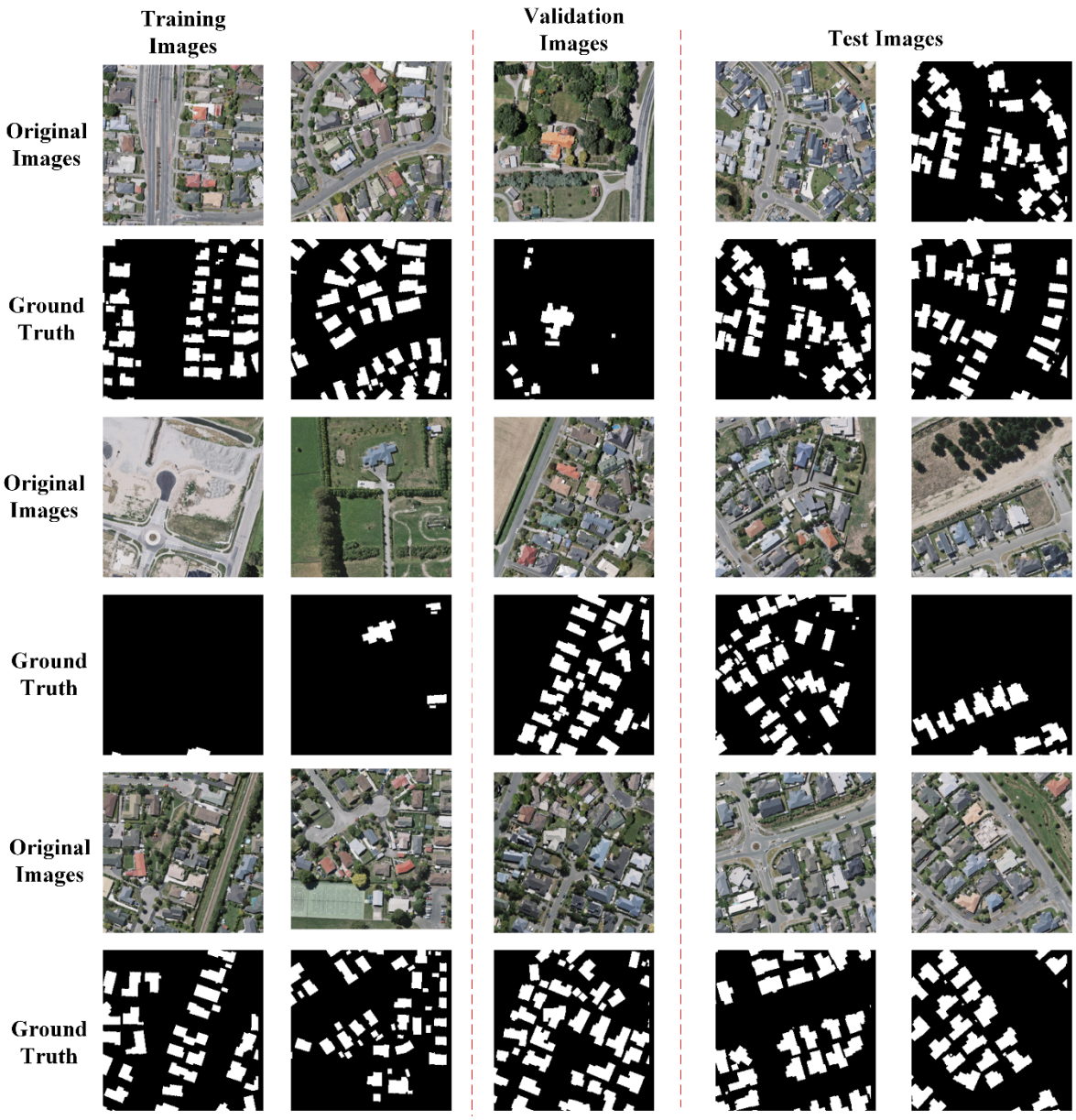

Figure 1. Several training, validation and testing sample images in the WHU building dataset.

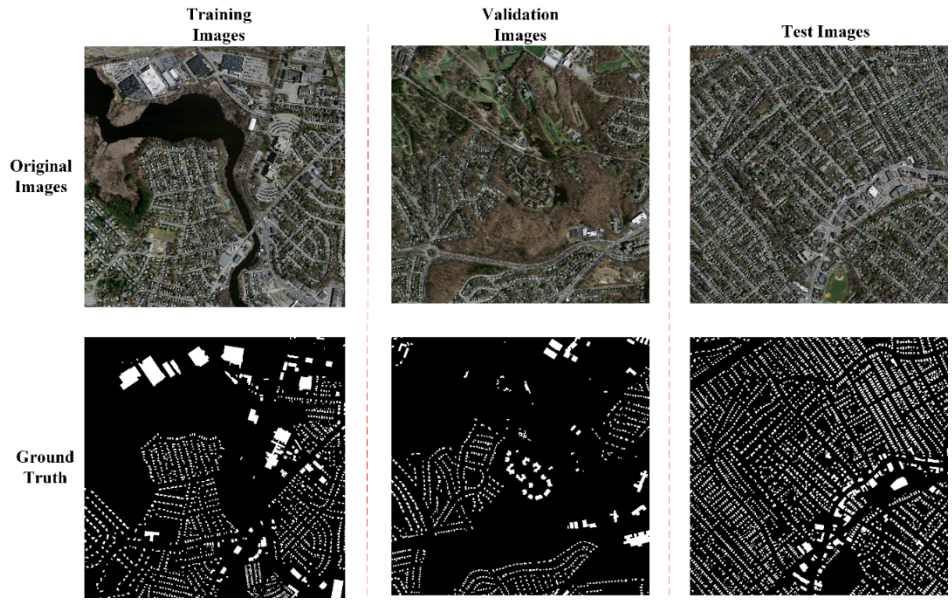

Figure 2. Several sample images in the original Massachusetts Building dataset for training, validation and testing, respectively.

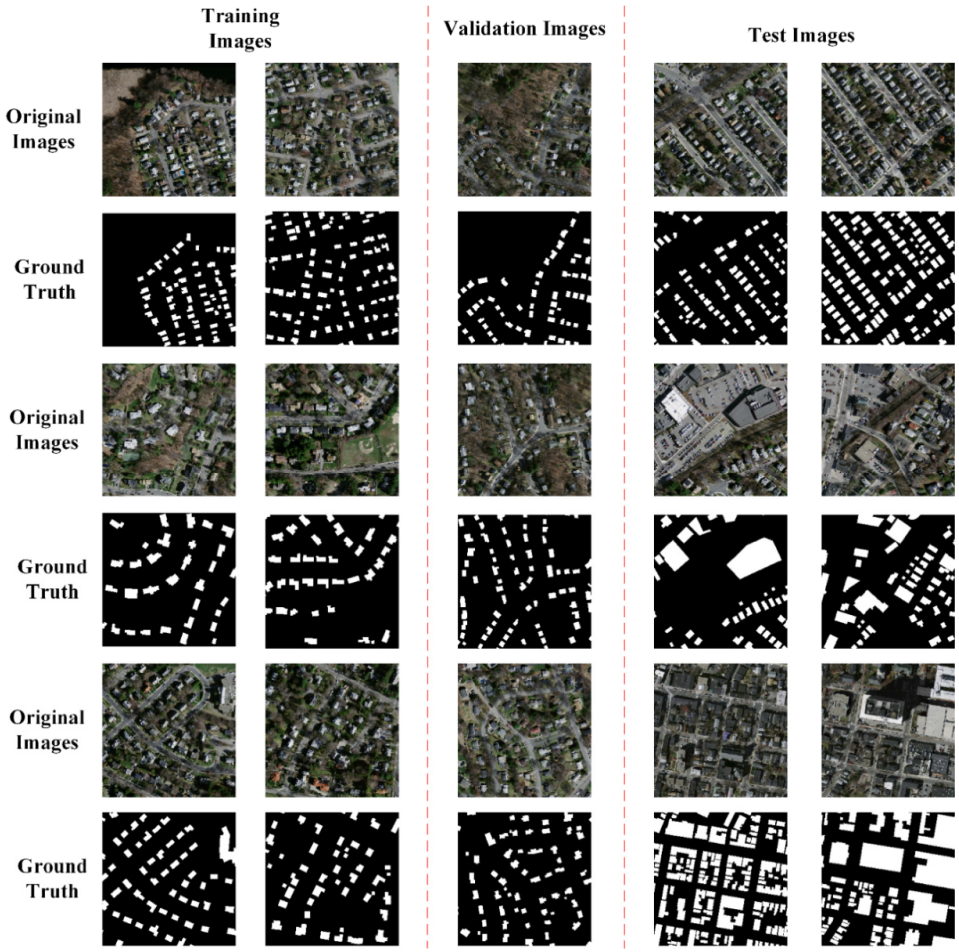

Figure 3. Several 256 × 256 sample images of Massachusetts Building dataset used in our experiments.

3.2. Model Architecture

Figure 4 shows the detail architecture of our model. The blue, pink, green, yellow, baby blue and red rectangles represent the convolution, ReLU, pooling, up-sampling, drop out and sigmoid operations, respectively. It can be seen that our model consists of two parts: the encoder and decoder. Essentially, our model is developed from the model of U-Net architecture. In the encoder part, our model includes two modules. The first module is the Self-Attention module, which is used for automatically learning the channel and position weights. The second module is the feature extraction module, which includes 4 groups of convolution, ReLU, pooling operations. The fourth group also includes a drop out operation. Through the 4 groups of operations, the feature is extracted. The detail convolutional size used in our model is 2×2. The max pooling operation with a pooling size of 2×2 is used in our model. In the decoder part, we use multi-up-sampling with multiple kernel sizes for each up-sampling layer strategy, which is similar with the strategy in [60]. The first up-sampling layer simultaneously has four up-sampling operations with up-sampling kernel sizes of 2, 4, 8 and 16, respectively. The filter numbers of up-sampling with kernel sizes of 2, 4, 8 and 16 are 512, 128, 64 and 32, respectively. Considering that large numbers of kernels with large size will put great pressure on the GPU memories, we use a strategy that the larger kernels have fewer filters. In the second up-sampling layer, we simultaneously use five up-sampling operations with up-sampling kernel sizes of 2, 4, 8, 16 and 32, respectively. The filter numbers of up-sampling with kernel sizes of 2, 4, 8, 16 and 32 are 256, 64, 32, 16 and 8, respectively. In the third up-sampling layer, we simultaneously use four up-sampling operations with up-sampling kernel sizes of 2, 4, 8 and 64, respectively. The filter numbers of up-sampling with kernel sizes of 2, 4, 8 and 64 are 128, 32, 16 and 2, respectively. In the fourth up-sampling layer, we simultaneously use three up-sampling operations with up-sampling kernel sizes of 2, 4 and 8, respectively. The filter numbers of up-sampling with kernel sizes of 2, 4 and 8 are 64, 16 and 8, respectively. In the next, we use five couples of convolution and ReLU operations. Finally, one couple of convolution and sigmoid operations is used. The final output of our model is a binary image with the same size of the input image. Note that for each up-sampling layer, multiple up-sampling operations are merged through concatenation operation instead of add operation. Figure 4 also shows the outputs after each operation layer in our model.

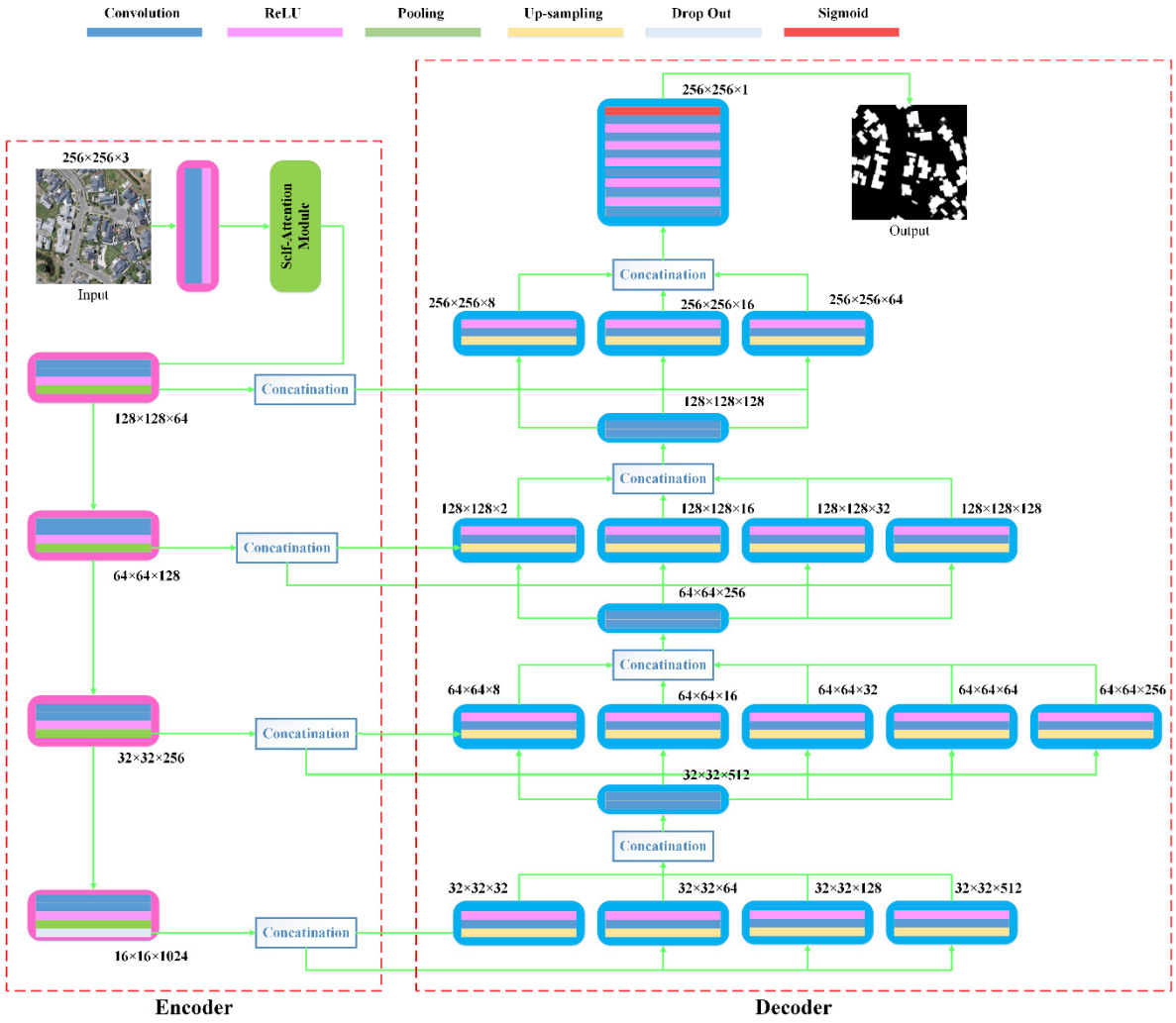

Figure 4. Detail architecture of our model.

3.3. Self-Attention Module

Figure 5 shows the detailed flowchart of the Self-Attention module. In the Self-Attention module, three convolution operations are simultaneously implemented first. After that, the outputs of two convolutions are merged by a matrix multiplication operation. Then, a softmax layer is followed by the merge operation. Fourth, the third output of convolution operation is merged with the output of softmax through matrix multiplication. Finally, the merge result is merged into the original inputs through an element-wise addition operation. In our model, the kernel size of the convolution operations is set at 1×1. Through the Self-Attention module, we can make the network learn the channel and position weights automatically.

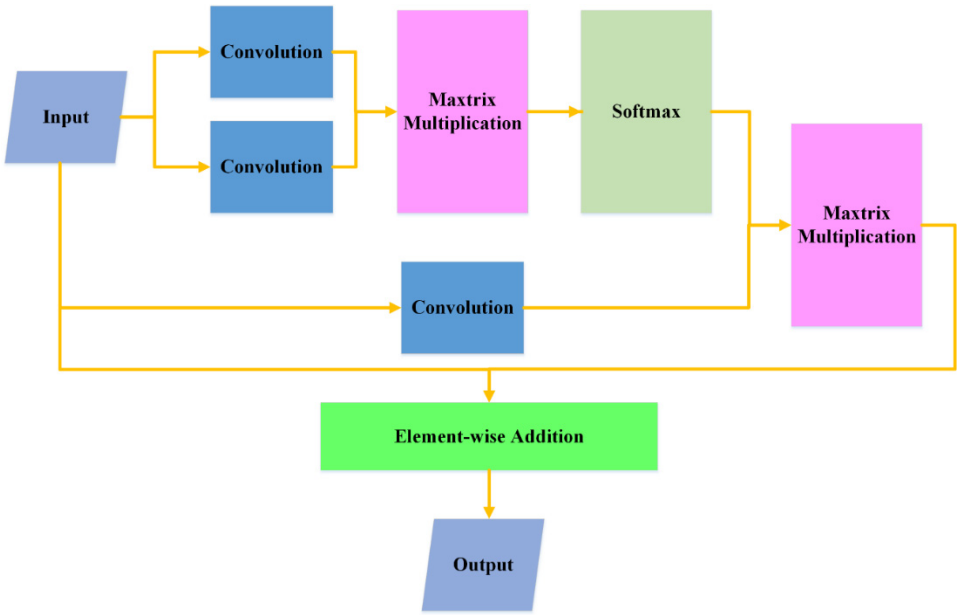

Figure 5. Flowchart of Self-Attention module.

3.4. Loss Function

In our model, as our goal is to segment the input image into two classes (building and background areas), we use binary cross entropy as our model's loss function. Given a couple of image and label (x, y), the output of the model is denoted as y_p, then the binary cross entropy loss of sample y can be represented as:

$$\text{Loss}(\text{x}, \text{y}) = -y_p * \text{logy} + (1 - y_p) \log(1 - \text{y}) \tag{1}$$

Then, the leaning goal of our network will be:

$$\text{Min}Loss = Min \sum_{i=1}^{n} (-y_{pi} * \text{logy}_i + (1 - y_{pi}) \log(1 - \text{y}_i)) \tag{2}$$

where n represents the total sample numbers.

4. Results

In this section, we first introduce the details about implementations. Then, we introduce the evaluation criteria used in the experiments. Finally, we present and analyze the experimental results on the test datasets.

4.1. Experimental Implementation Detail

We implement our experiments on a computer that has an Intel®Core™ i9-9900X 3.5 GHz CPU and 128 GB memory. The GPU type used in this computer is RTX 2080 Ti with 11 GB GPU memory. The codes of our experiments are all based on Python, TensorFlow and Keras. Due to the limitation of GPU memory, we use a small training batch size of 2 during the training and validation. Instead of using stochastic gradient descent (SGD), we use Adam optimizer for training our model. The learning rate is set at 0.0001. For both WHU and Massachusetts Building datasets, the training epoch is set at 500. The steps per epoch are set at 4000 and 1800 for WHU and Massachusetts Building datasets, respectively. To further enhance the performance and prevent the overfitting problem, we apply a data augmentation strategy that will randomly rotate the image in a range of -1 to 1 degree, width shift in a range of 0.1, height shift in a range of 0.1 and horizontal flip. During the augmentation, the fill mode is set as nearest.

4.2. Evaluation Criteria

To comprehensively evaluate the performance of models, we use four evaluation criteria that are widely used for evaluating building segmentation performance. The first to fourth criteria are recall, precision, IoU and F1-Score [61], the representations are as follows:

$$\begin{aligned} \text{recall} &= \frac{TP}{TP+FN} \\ \text{precision} &= \frac{TP}{TP+FP} \\ \text{IoU} &= \frac{TP}{TP+FN+FP}, \\ \text{F} - Score &= \frac{2*precision*recall}{precision+recall} \end{aligned} \tag{3}$$

where *TP*, *FN* and *FP* denote true positive, false negative and false positive, respectively. Note that all the following performance evaluation results are calculated based on pixels.

4.3. Experimental Results and Analysis

4.3.1. Experimental Results on the WHU Dataset

In this section, we demonstrate the performance of our model tested on the WHU dataset. For comparison, we also compared with several currently popular building extraction methods tested on the WHU dataset. As WHU is an open available dataset, we just use the reported performances for our comparisons. The compared methods for the WHU dataset include the U-Net [54], the Segnet [62], the DRNet [1], the SRI-Net [49], the DeepLabV3+ [63] and the Zhou's [6]. In Table 1, the U-Net, SegNet, DRNet, SRI-Net, DeepLabV3+, Zhou's and ours obtain the recalls of 90.67%, 89.93%, 93.3%, 93.28%, 92.2%, 92.94% and 95.56%, respectively. The precision scores of the seven methods (including ours) are 94.59%, 92.11%, 94.3%, 95.21%, 94.27%, 94.2% and 93.25%, respectively. The IoU scores of the seven methods are 86.2%, 85.56%, 88.3%, 89.09%, 87.31%, 87.97% and 89.39%, respectively. The F1-Scores of the seven methods are 92.59%, 91.01%, 93.8%, 94.23%, 93.22%, 93.55% and 94.4%, respectively. In the precision metric, SRI-Net achieves the best score. However, our approach achieves a higher score in the Recall metric compared with SRI-Net. Thus, the IoU scores and F1-Scores are higher than SRI-Net, achieving the best performances among the compared methods. Since IoU and F1-Score are both comprehensive metrics, the results prove the satisfactory performance of our model. From the aspect of F1-Score, our model obtains score results that are about 1.96%, 3.7%, 0.6%, 0.2%, 1.3% and 0.9% higher than the results of U-Net, SegNet, DRNet, SRI-Net, DeepLabV3+ and Zhou's, respectively. From the aspect of IoU, our method achieves score results about 3.7%, 4.5%, 1.2%, 0.3%, 2.4% and 1.6% higher than the results of U-Net, SegNet, DRNet, SRI-Net, DeepLabV3+ and Zhou's, respectively.

Table 1. Quantitative comparison results (%) among our method and the other six classical building extraction methods tested on the WHU dataset.

Method	Recall	Precision	IoU	F1-Score
U-Net	90.67	94.59	86.2	92.59
SegNet	89.93	92.11	85.56	91.01
DRNet	93.3	94.3	88.3	93.8
SRI-Net	93.28	**95.21**	89.09	94.23
DeepLabV3+	92.2	94.27	87.31	93.22
Zhou's	92.94	94.2	87.97	93.55
Ours	**95.56**	93.25	**89.39**	**94.4**

Figure 6 shows the sample exhibitions of the extraction results by our method. The first, second and third columns represent the original images, ground truths and our building extraction results. The selected target original images are difficult for extraction as the buildings are located densely with complex backgrounds. Even so, our model extracts the building areas with pretty good results in visual. The visual results further prove that our model is effective.

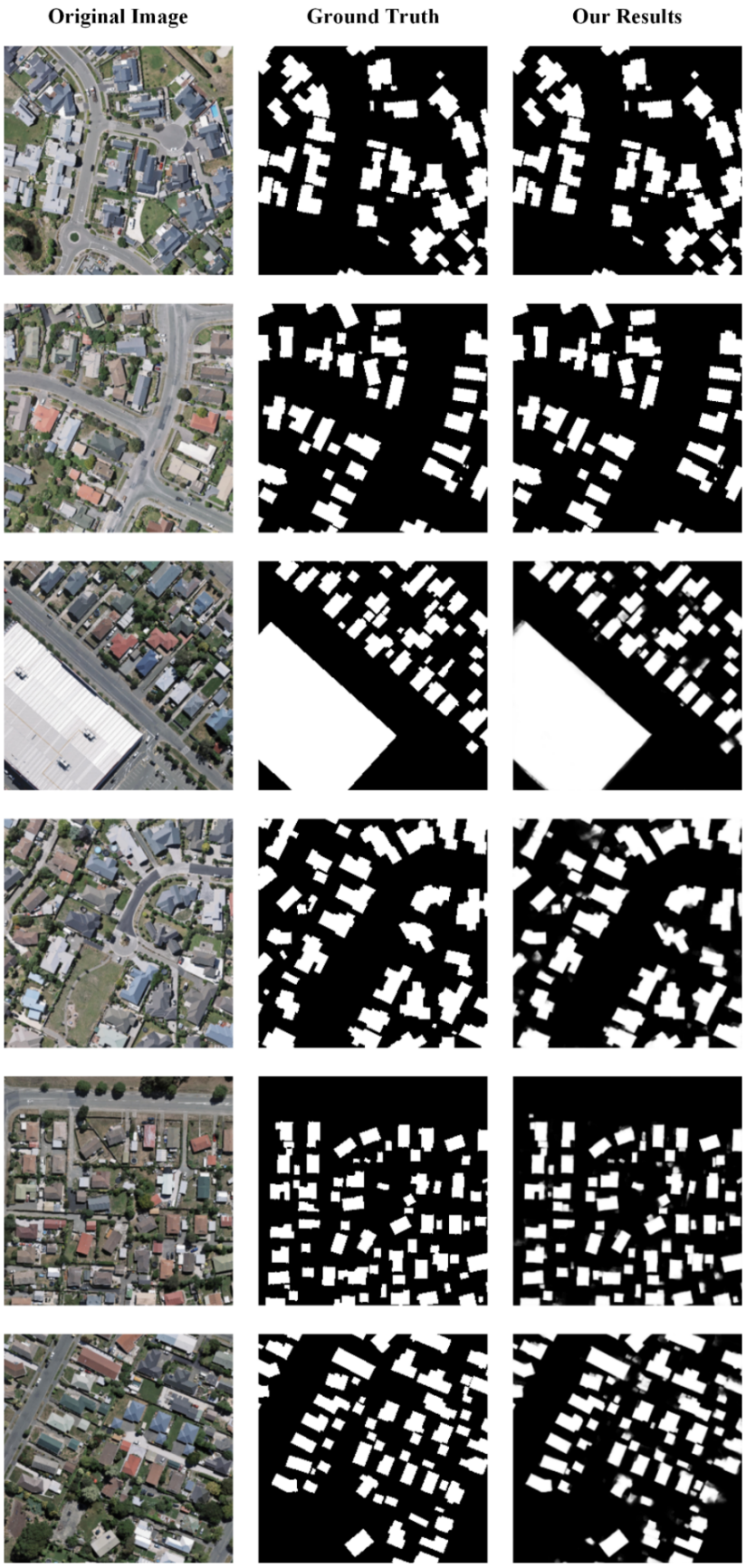

Figure 6. Building extraction results of our methods. The first, second and third columns are the original images, ground truths and our extraction results, respectively.

4.3.2. Experimental Results on the Massachusetts Building Dataset

In this part, we first show the quantitative comparison results among our method and nine recent classical and popular methods tested on the Massachusetts Building dataset. The compared methods consist of the U-Net [54], the Segnet [62], the DRNet [1], the DeepLabV3+ [63], the FCN [52], the DeepLavV3+ [63], the MSCRF [64], the D-LinkNet [65], the GMEDN [66] and the JointNet [13]. Table 2 shows the quantitative comparison results among our method and other classical building extraction methods tested on the Massachusetts Building dataset. Note that all the results of other compared methods are obtained from the standard results reported by the authors. As not all the methods have the result reports for all the recall, precision, IoU and F1-score metrics, several methods have blanks for some metrics. Fortunately, all the compared methods have their scores for the comprehensive metric IoU. In Table 2, the MSCRF obtains the highest recall score among the valid four scores. However, our method obtains a much higher score for precision. Thus, for the comprehensive IoU metric, our method achieves a much higher score compared with MSCRF, and also higher than all the compared methods. In the F1-Score, only three methods have the score reports, among which our method shows a much higher score than other two methods. For detail, our method acquires a IoU score of 73.49%, which is about 11.9%, 15.9%, 11.3%, 10.5%, 19.7%, 3.2%, 12.1%, 4.4%, 2.1% higher compared with the U-Net [54], the Segnet [62], the DRNet [1], the DeepLabV3+ [63], the FCN [52], the DeepLavV3+ [63], the MSCRF [64], the D-LinkNet [65], the GMEDN [66] and the JointNet [13], respectively. The results prove the effectiveness of our model.

Table 2. Quantitative comparison results (%) among our method and other nine classical building extraction methods tested on the Massachusetts Building dataset.

Method	Recall	Precision	IoU	F1-Score
U-Net	-	-	65.65	-
SegNet	-	-	63.43	-
DRNet	-	-	66	79.5
FCN	-	-	66.5	-
DeepLabV3+	-	-	61.4	76.1
MSCRF	89.93	80.14	71.19	-
D-LinkNet	85.58	73.36	65.54	-
GMEDN	-	-	70.39	-
JointNet	81.29	86.21	71.99	-
Ours	86.15	83.34	73.49	84.72

Figure 7 shows several sample extraction results by our method tested on the Massachusetts Building dataset. The first, second and third columns are the original images, ground truths and our extraction results, respectively. The given test samples are all quite challenging in visual. However, our method obtains quite good extraction results in all the given sample images in visual. Compared with the ground truth, our extraction results seem to be smoother. Moreover, our extraction results seem to lose the very small building area targets.

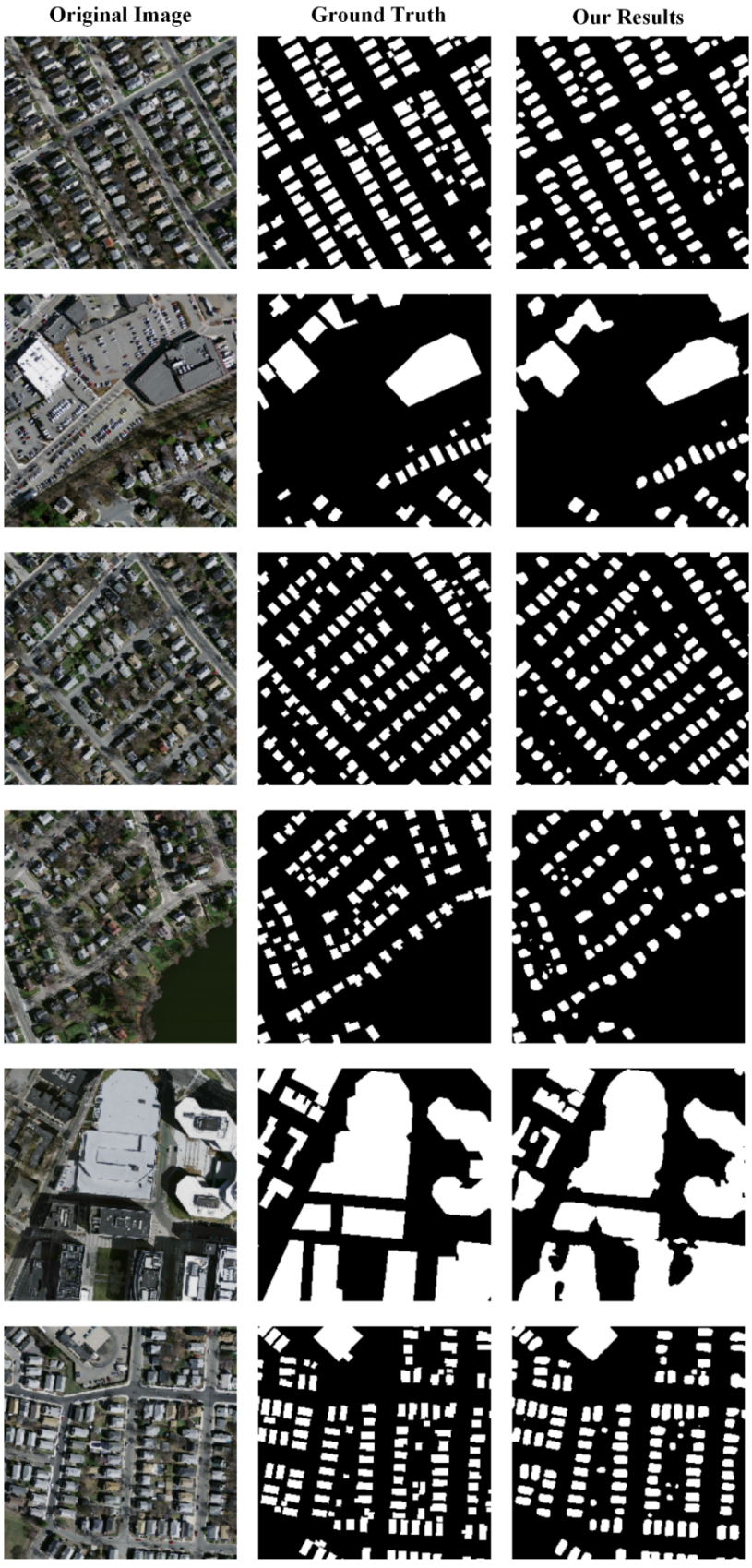

Figure 7. Building extraction results by our method tested on the Massachusetts Building dataset. The first, second and third columns are the original images, ground truths and our extraction results, respectively.

4.3.3. Major Abbreviations Used in Our Paper

In order for readers to have a better understanding of our work, we provide a detailed introduction about the major abbreviations used in our paper. The details are as shown in Table 3:

Table 3. The major abbreviations used in our paper.

CNN:	Convolutional Neural Networks	WHU:	WuHan University
FCN:	Fully Convolutional Networks	IoU:	Intersection over Union
ReLU:	Rectified Linear Unit	TP:	True Positive
SGD:	Stochastic Gradient Descent	FP:	False Positive
VHR:	Very High Resolution	FN:	False Negative
SVM:	Support Vector Machine		

5. Discussion

5.1. Effectiveness of Transformer Module

In this part, we analyze the effectiveness of the transformer module. To verify the effectiveness of the transformer module, we designed an ablation experiment. In this experiment, we removed the transformer module from our model structure and kept other parts of our model the same as our original model. We tested the model without the transformer model on the WHU dataset and compared the result with our original model. Table 4 shows the comparison results. In Table 4, the score of the recall metric of our original model was a little lower than the score of our model without the transformer. However, our original model achieved higher scores in all the precision, IoU and F1-Score metrics

Table 4. Quantitative comparison results (%) between our method with and without the transformer module tested on the WHU dataset.

Method	Recall	Precision	IoU	F1-Score
Ours with Transformer	95.56	93.25	89.39	94.4
Ours without transformer	95.59	92.02	88.2	93.77

Compared with our model without transformer module. The results prove that the transformer module can effectively enhance the attention recognition ability through the channel and position weights. Thus, the final performance is improved through the transformer module.

Figure 8 shows several sample result comparisons in visual between our model with and without the transformer module. The first, second, third and fourth columns represent the original images, ground truths, results with transformer and results without transformer, respectively. From Figure 8, we can see that the results by our whole model present clearer in the areas that may be interfered with by other complex background targets. The reason may be that the transformer module can automatically learn the position and channel weights during the training. Thus, the learned attention weights can tell the network where it needs to pay more attention when extracting buildings on the given test image. Through focusing on positions where there may be building areas with larger probabilities, the model can achieve a better performance finally.

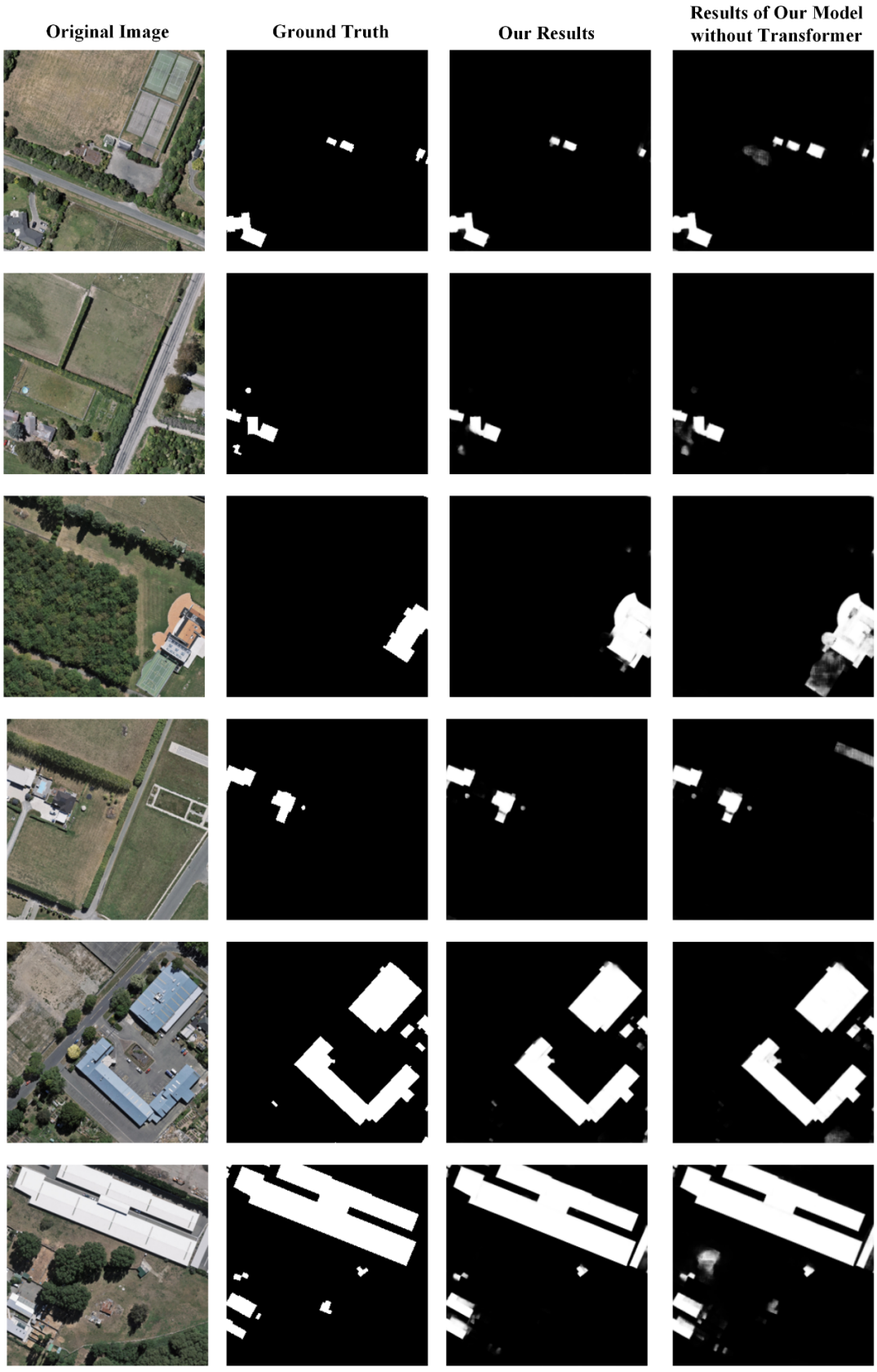

Figure 8. Several sample result comparisons in visual between our model with and without the transformer module. The first, second, third and fourth columns represent the original images, ground truths, results with transformer and results without transformer, respectively.

We also verified the effectiveness of the transformer module on the Massachusetts Building dataset. Table 5 shows the quantitative comparison results between our method with and without the transformer module tested on the Massachusetts Building dataset. In the table, our method with the transformer module achieves a higher precision, IoU and F1-Score than our method without the transformer module. The recall of our method with the transformer module is only a little lower than our method without the transformer module. The results also prove that the transformer module is effective to improve the performance of our model.

Table 5. Quantitative comparison results (%) between our method with and without the transformer module tested on the Massachusetts Building dataset.

Method	Recall	Precision	IoU	F1-Score
Ours with Transformer	86.15	**83.34**	**73.49**	**84.72**
Ours without transformer	**87.05**	80.98	72.27	83.9

5.2. False Extraction

In this section, we analyze the false extractions (including false positive and false negative) in the tests. The goal is to find out the wrong situations, from which we may find a way to further improve the model performance in our future works.

Figure 9 shows the major false positive exhibitions of our extraction results in the WHU dataset. The green, red and blue areas represent the right, false positive and false negative areas, respectively. From the visualization results of major false positive areas, we can see that the major false positive extractions occurred for four reasons: (1) the building yard appears at an unusual shape similar to a building and may lead to the hard recognition at the edge areas and cause false positives; (2) the shapes of other objects look very similar to buildings and may result in false recognition; (3) the containers seem to cause a large amount of false positives; (4) the areas that are not buildings and have shadows may also result in false positives.

Figure 10 shows the major false negative exhibitions of our extraction results in the WHU dataset. The green, red and blue areas represent the right, false positive and false negative areas, respectively. We analyzed the false negatives in the false negative areas in our test images, and found that the major false negatives may occur for the following reasons: (1) the occlusions by the trees or Other objects; (2) the building with a special roof color that is similar to the color of the ground or road; (3) our model seems to lose the consistency restrictions of a building, such as the example in the last three false negative samples in the Figure 10.

Figure 11 shows the major false positive and false negative exhibitions of our extraction results in the Massachusetts Building dataset. The green, red and blue areas represent the right, false positive and false negative areas, respectively. In Figure 11, the major false positives occur at four kinds of positions: (1) the edges of the buildings; (2) the interspace between buildings, especially where it appears with dark shadows; (3) the sports ground, such as the tennis court; (4) the areas with light grey colors, such as the beach areas. On the other hand, the major false negatives seem to occur at the following positions: (1) building areas that are occluded by shadows and trees, etc.; (2) the wrong labels in the Massachusetts Building dataset; (3) the areas that look similar to roads.

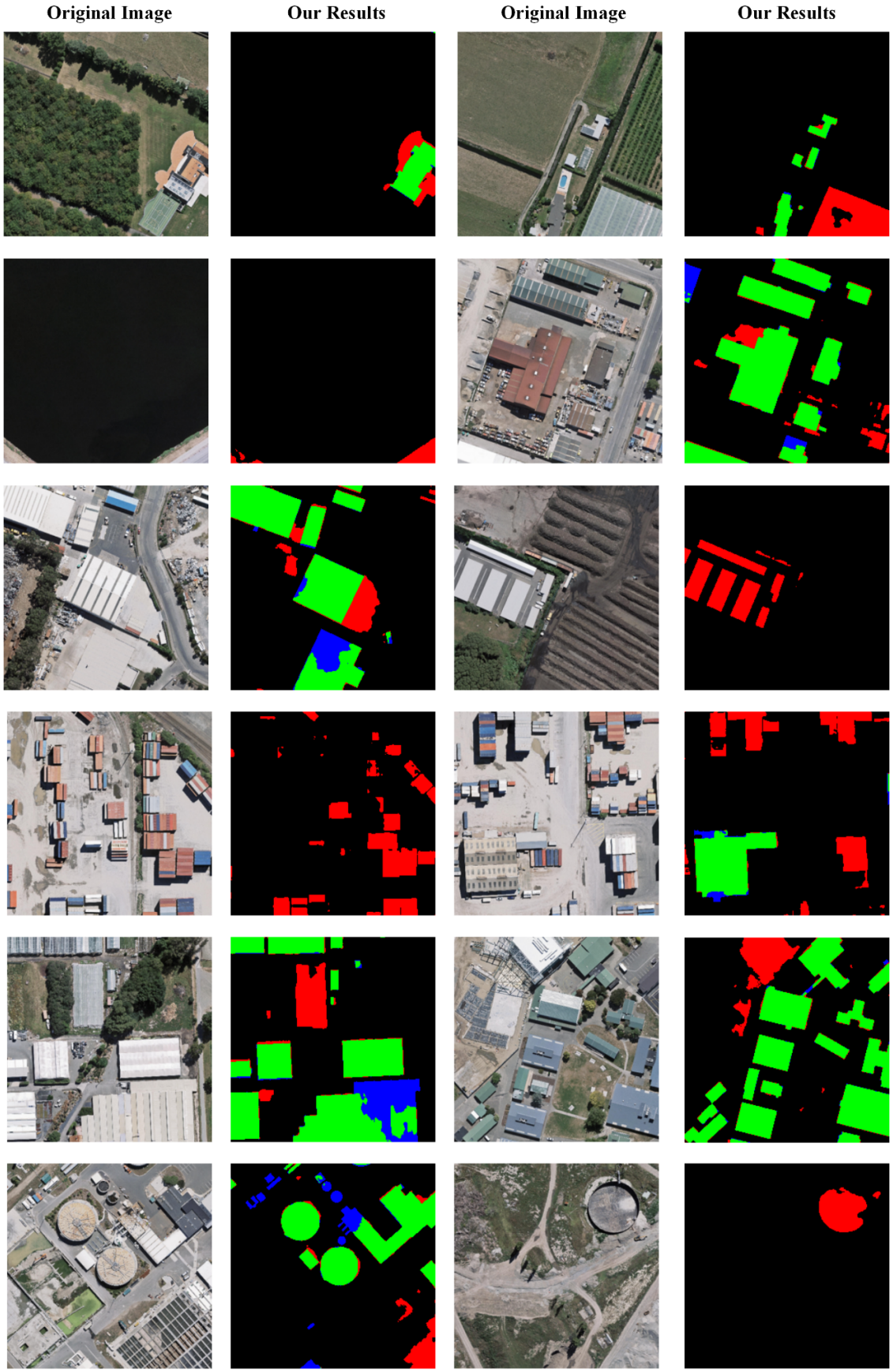

Figure 9. Major false positive exhibitions of our extraction results in the WHU dataset. The green, red and blue areas represent the right, false positive and false negative areas, respectively.

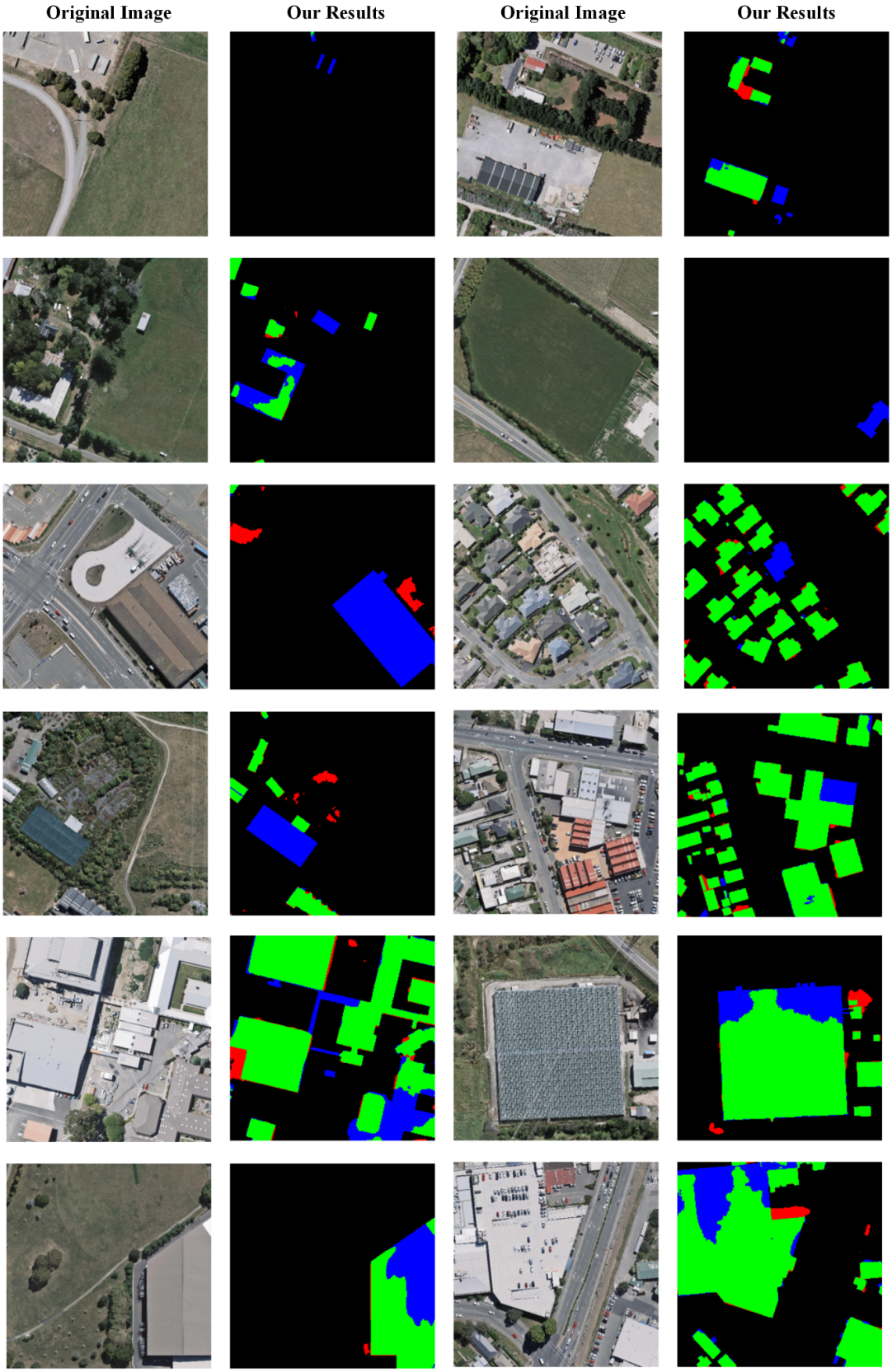

Figure 10. Major false negative exhibitions of our extraction results in the WHU dataset. The green, red and blue areas represent the right, false positive and false negative areas, respectively.

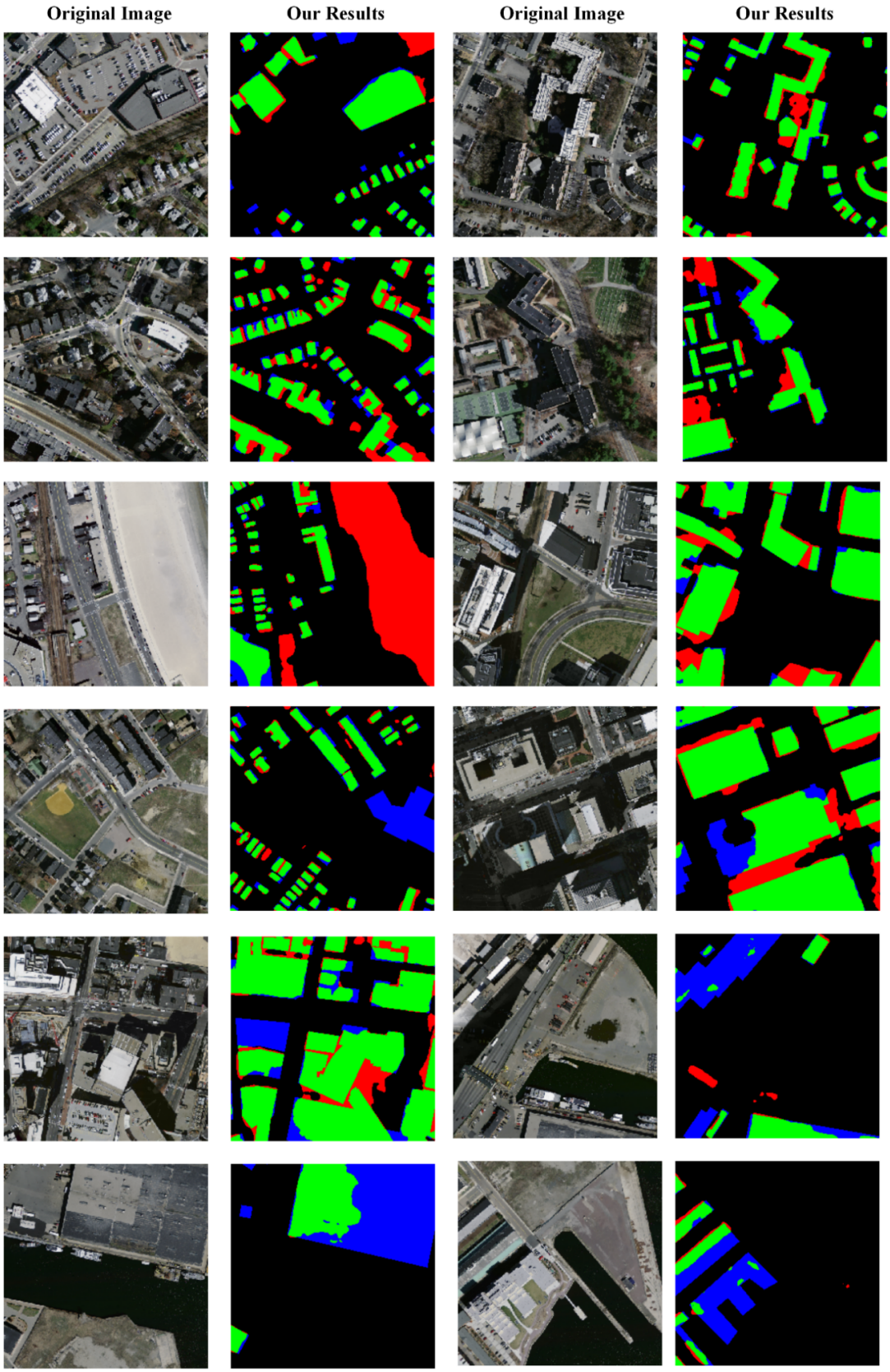

Figure 11. Major false positive and false negative exhibitions of our extraction results in the Massachusetts Building dataset. The green, red and blue areas represent the right, false positive and false negative areas, respectively.

Figures 12 and 13 show the extraction results of our model in the WHU building dataset and Massachusetts Building dataset, respectively. In Figures 12 and 13, few wrong extractions occur when the given test images have no buildings or other artificial objects. The results prove that our model is robust to images that only contain lands, grasses, vegetation and soil, etc.

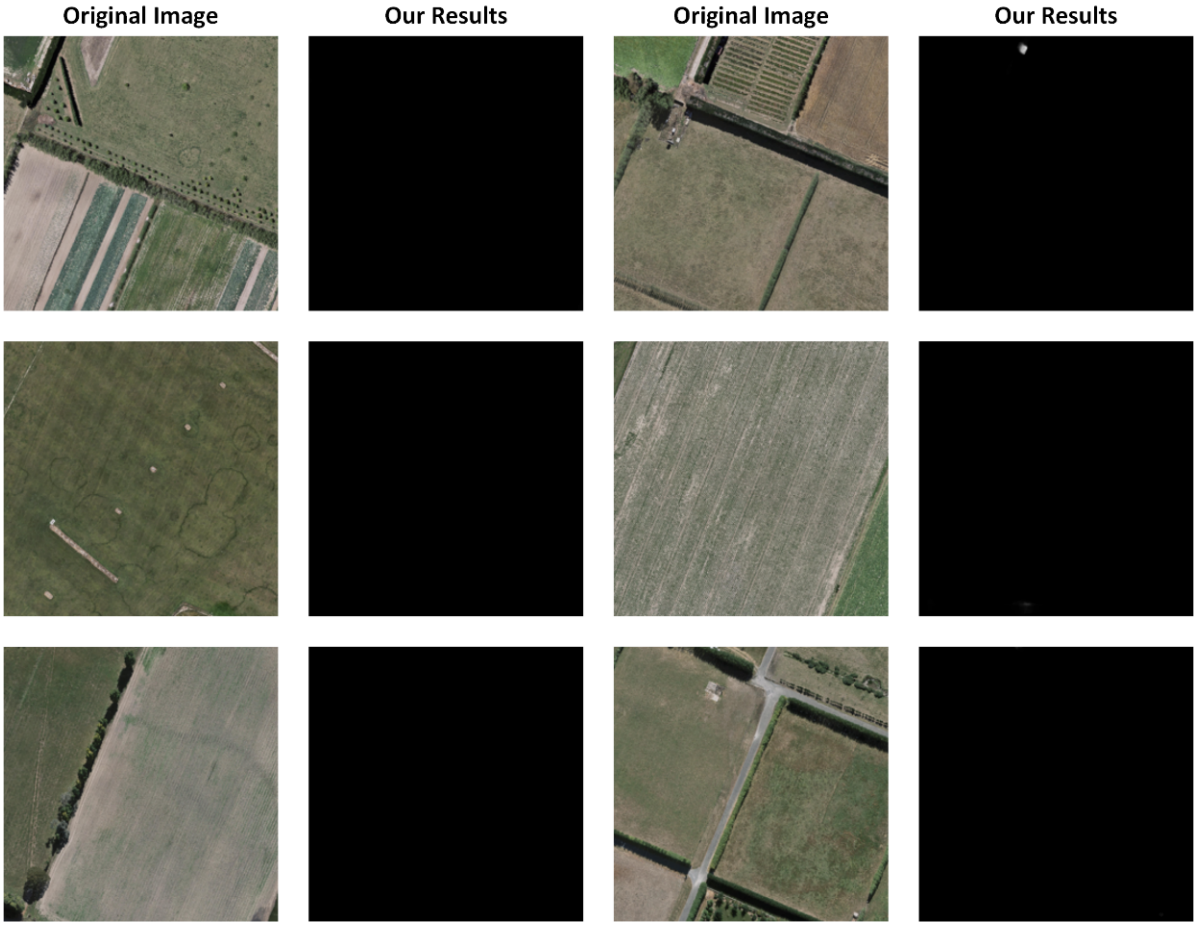

Figure 12. The extraction result exhibitions of images without buildings in the WHU building dataset.

Overall, from the visualizations of false positive and negative areas, we can clearly see what happened in the wrong extractions. After that, we can obtain helpful information through analyzing the wrong extractions. According to the analysis about the visualizations of false positive and negative areas of our method, we find that our model may partly lose the extractions within a building under several situations. We think the reason is that our model does not consider the entirety of the building structure and the local context information. Moreover, the in-painting processing is not considered after the extractions of our model. From the above analysis, the following research directions may be the breakthrough points in our future research. First, we will further study how to enhance the context information learning to help improve the performance. Second, the entirety consistence restriction may be useful for promising the completeness of an extracted building. Third, to further improve the model's generalization ability, a powerful and effective augmentation strategy for the training images is necessary. Augmenting the training images simply through rotation and shift transformation is insufficient. In our future work, we will keep studying building extraction in remote-sensing images from the above three aspects.

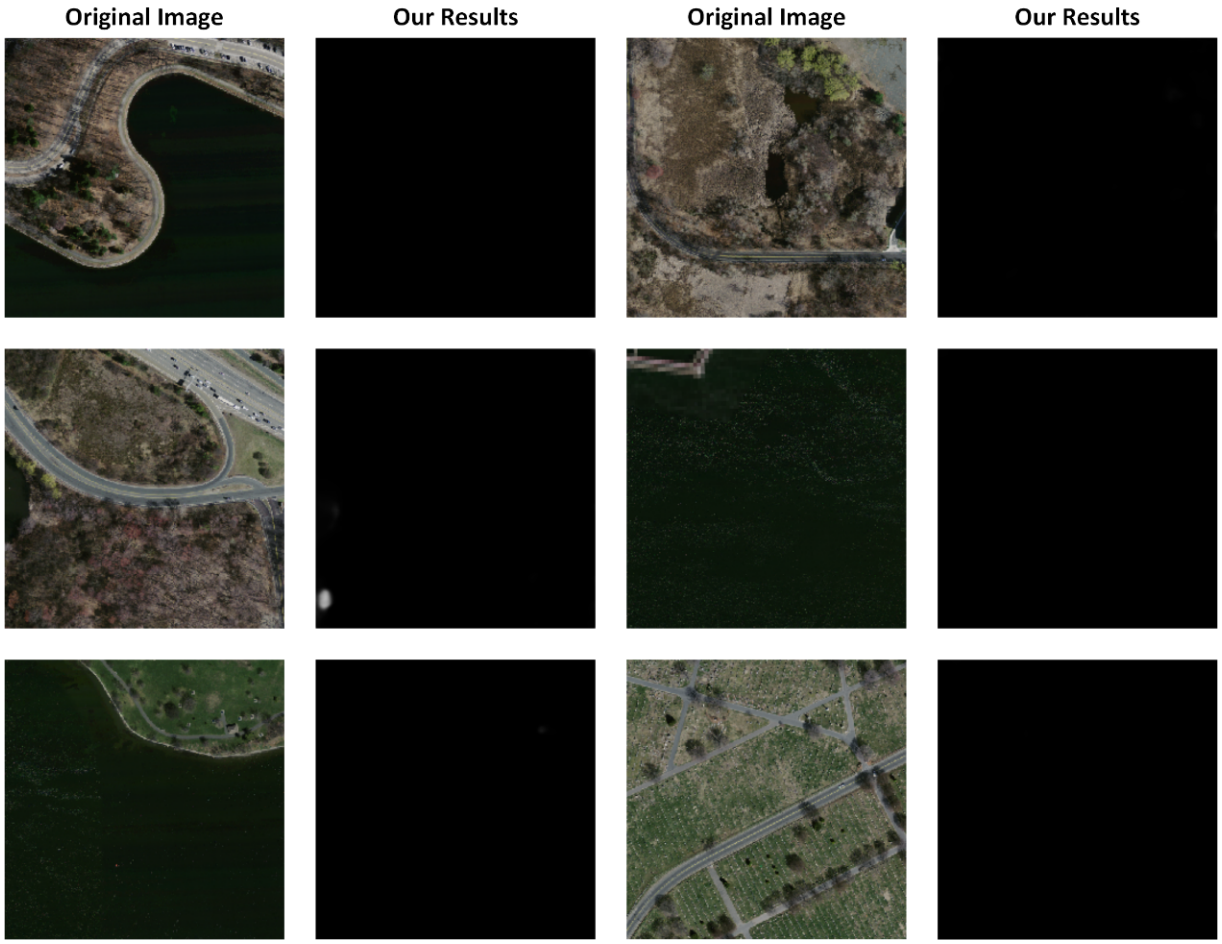

Figure 13. The extraction result exhibitions of images without buildings in the Massachusetts Building dataset.

5.3. Result Comparison Analysis

In this section, we discuss the experimental comparison results among our method and other compared methods. Although Tables 1 and 2 proved the satisfactory performance of our method, we further reproduced the results of six classical building extraction and segmentation methods tested on both the WHU and Massachusetts Building datasets. From the comparisons in quantitative results as well as visual results, we further analyzed the comparisons. Note that we only reproduced the methods for which the codes are public.

Table 6 shows the reproduced quantitative comparison results among our method, U-Net [54], SegNet [62], DANet [55], PSPNet-101 [20], and Residual U-Net [53] tested on the WHU dataset. From the results, we can see that our method obtained the highest score in both IoU and F1-Score, which further convincingly proves the good performance of our method. For the results of U-Net and SegNet, their reproducing performances were similar with the results in Table 1. Although the DANet, PSPNet-101 and Residual U-Net are all famous segmentation methods, our method performed better than those methods in the experiment.

We also illustrate the visual comparisons. Figure 14 shows the visual comparison results of our method and the compared methods tested on the WHU dataset. The first column to the eighth column are the original test images, ground truth, results of U-Net, SegNet, DANet, PSPNet-101, Residual U-Net and ours, respectively. From the visual comparison results, we can see that our results presented better performance compared with the other methods. The U-Net seems to show more false positives in the results. For example, the tennis court is extracted as a building in the first row. The SegNet shows worse results compared with ours in the extraction entirety attribution, and it seems to

have more missing positives. The same problems exist in the results of DANet. The PSPNet is sensitive to large buildings, while it shows worse accuracy in the building edge areas. The Residual U-Net's results were similar with the U-Net's. Our method showed fewer false positives and false negatives in most results. However, our method also showed bad building entirety attribution in some test areas, e.g., the results shown in the fourth row. According to the results of all the methods, we found that all the methods do not perform well in the building-entirety attribution. Thus, conquering the problem of missing part of a building may be an interesting research point in our future work.

Table 6. Quantitative comparison results (%) among our method and the other five classical building extraction and segmentation methods tested on the WHU dataset.

Method	Recall	Precision	IoU	F1-Score
U-Net	95.205	91.503	87.47	93.32
SegNet	91.48	92.49	85.15	91.98
DANet	92.02	92.72	85.82	92.37
PSPNet-101	91.94	89.39	82.9	90.65
Residual U-Net	93.24	92.08	86.31	92.65
Ours	95.56	93.25	89.39	94.4

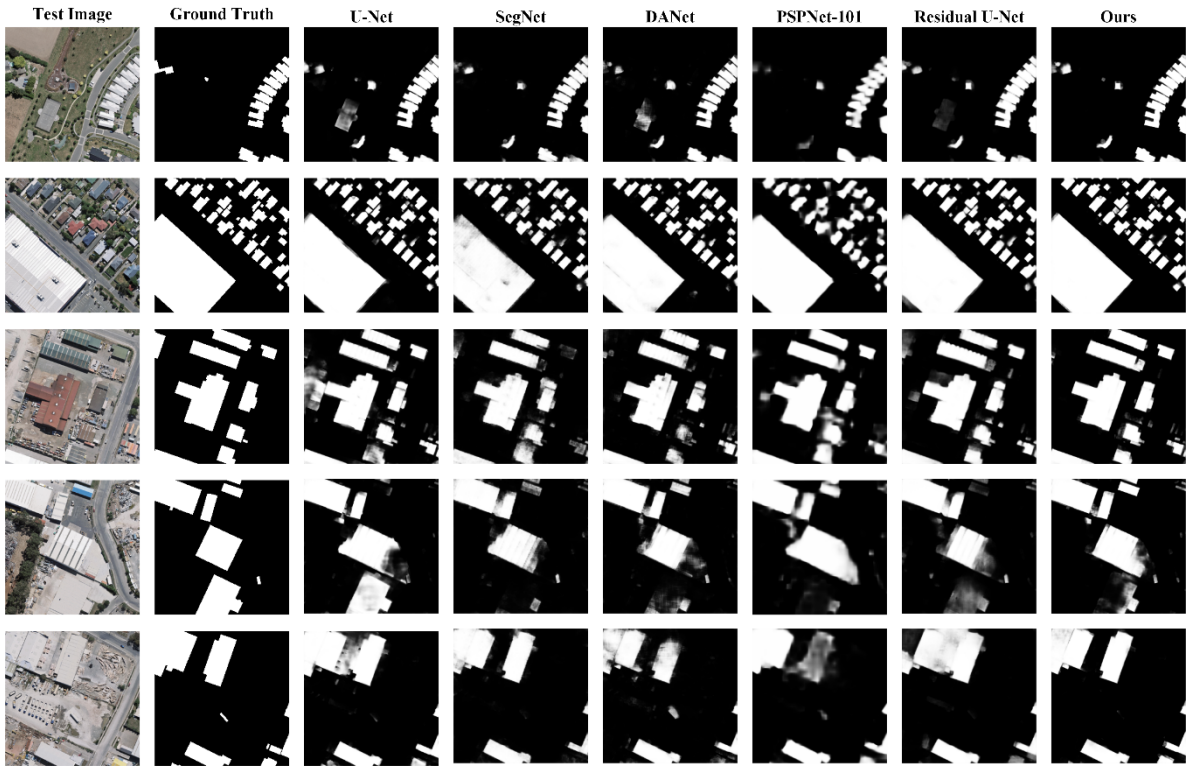

Figure 14. Several visual results of comparisons tested on the WHU dataset.

Table 7 shows the quantitative comparison results among our method and the reproducing results of the other five classical building extraction and segmentation methods tested on the Massachusetts Building dataset. The reproducing results of U-Net and SegNet seemed to obtain a little higher scores compared with the reported results. However, their result scores were still much lower ours. Although the DANet obtained the best precision score, its recall rate was lowest, resulting in lower scores in IoU and F1-Score. The PSPNet-101 showed good performance in IoU score, and the Residual U-Net showed good performance in F1-Score. From the quantitative results, our method achieved the best results in this comparison, which further proves the effectiveness of our model convincingly.

Table 7. Quantitative comparison results (%) among our method and the other five classical building extraction and segmentation methods tested on the Massachusetts Building dataset.

Method	Recall	Precision	IoU	F1-Score
U-Net	82.65	80.63	68.96	81.63
SegNet	78.85	79.54	65.55	78.85
DANet	75.31	84.56	66.21	79.67
PSPNet-101	82.52	78.38	67.22	80.39
Residual U-Net	80.75	77.43	65.36	79.05
Ours	86.15	83.34	73.49	84.72

Figure 15 shows the visual comparison of building extraction results among our method and the other five discussed methods tested on the Massachusetts Building dataset. The U-Net, SegNet, DANet and Residual U-Net performed well in small building areas, except for PSPNet-101. However, PSPNet-101 performed better than the other four methods in the large building areas. It may benefit from the pyramid pooling strategy, resulting in larger insight areas and better understanding of context information. From this phenomenon, we think that finding how to obtain both larger areas of context information and keep the detail information may be an attractive research point in the building extraction from remote-sensing images. Compared with the other five methods, our results seem to perform well in both large building areas and small building areas. In this respect, our method performed better than the other five methods in the visual results.

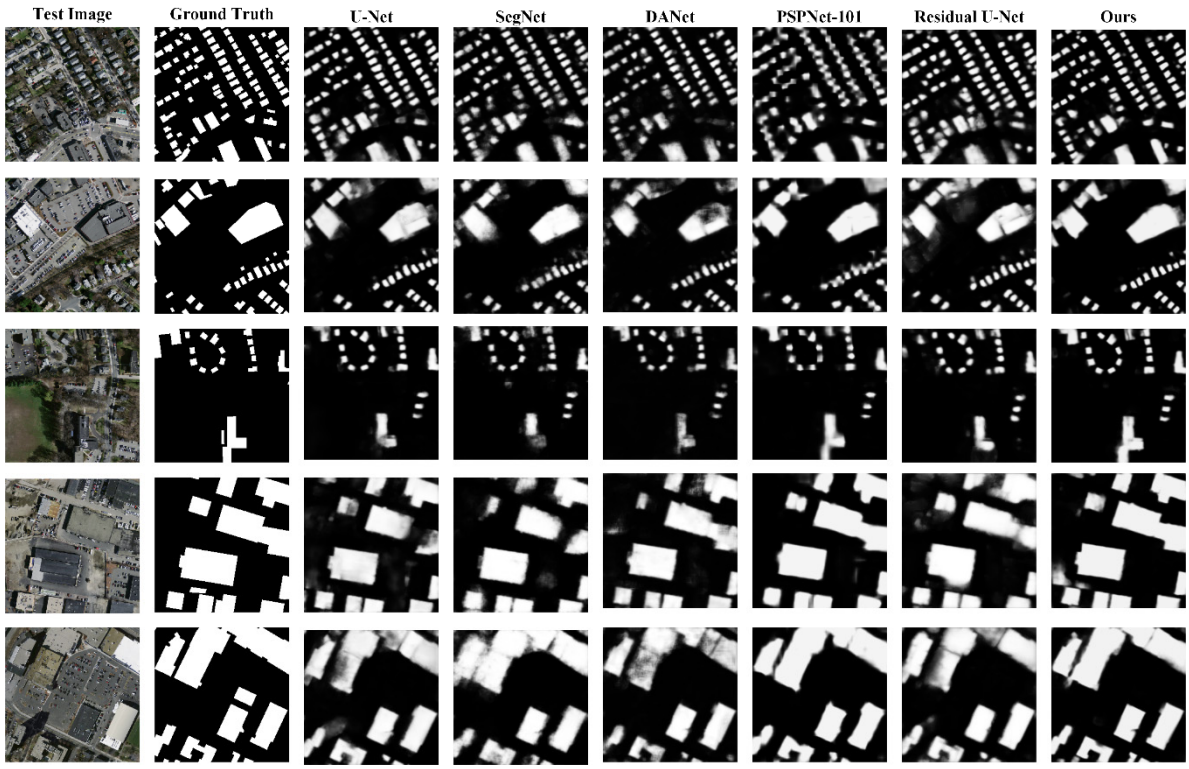

Figure 15. Several visual results of comparisons tested on the Massachusetts Building dataset.

6. Conclusions

This paper proposed an encoder–decoder model for building extraction from optical remote-sensing images. In our model, we added a transformer module in the encoder, making the model automatically learn the channel and position weights for an input image. In the decoder part, we used a reconstruction-bias structure, which reinforced the decoding ability of our model. We tested our model on two publicly available datasets: the WHU and

Massachusetts Building datasets. We also compared our performances with several other currently classical and popular methods on both datasets. The quantitative comparisons proved that our model can achieve satisfactory performance in both datasets. For detail, our model can achieve an IoU score and F1-score at 89.39% and 94.4%, respectively, tested on the WHU dataset. And our method can obtain an IoU score and F1-score at 73.49% and 84.72%, respectively, tested on the Massachusetts Building dataset. We also set up an ablation experiment to prove the effectiveness of the transformer module. In the WHU dataset, our model with the transformer module can achieve an 89.39 IoU score and a 94.4 F1-score, while our model without the transformer module can only achieve an 88.2 IoU score and a 93.77 F1-score. Finally, we visualized the wrong extractions for analyses, which is helpful for our future further study. From the false extractions analysis, we found that our model still shows its weakness in understanding the context information and the entirety structure of buildings. Thus, in future work we will continue our study and focus on catching context information and learning the entirety structure of buildings for our model.

Author Contributions: Conceptualization, Z.C. and C.W.; methodology, Z.C.; software, Z.C.; validation, Z.C.; formal analysis, Z.C., D.L. and W.F.; investigation, Z.C. and J.L.; resources, Z.C. and J.L.; data curation, Z.C.; writing—original draft preparation, Z.C.; writing—review and editing, Z.C., D.L., H.G. and J.L.; visualization, Z.C.; supervision, C.W. and H.G.; project administration, Z.C.; funding acquisition, Z.C., H.G. and C.W. All authors have read and agreed to the published version of the manuscript.

Funding: This study was financially supported by National Natural Science Foundation of China (No. 62001175), Natural Science Foundation of Fujian Province (No. 2019J01081), United National Natural Science Foundation of China (No. U1605254), National Natural Science Foundation of China (No. 6187606, 61972167, 61801121, 41971414 and 61673186), and the Project of Science and Technology Plan of Fujian Province (2020H0016).

Data Availability Statement: We do not have online result source.

Acknowledgments: This paper's work is partially supported by China Transport Telecommunications & Information Center and Guojiao Spatial Information Technology (Beijing) Co., Ltd.

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

References

1. Chen, M.; Wu, J.; Liu, L.; Zhao, W.; Tian, F.; Shen, Q.; Zhao, B.; Du, R. DR-Net: An Improved Network for Building Extraction from High Resolution Remote Sensing Image. *Remote. Sens.* **2021**, *13*, 294. [CrossRef]
2. Li, W.; Wang, S.; Li, J. Object based building extraction by QuickBird image for population estimation: A case study of the City of Waterloo. In Proceedings of the 2014 IEEE Geoscience and Remote Sensing Symposium, Quebec City, QC, Canada, 13–18 July 2014; pp. 3176–3179.
3. Gupta, A.; Watson, S.; Yin, H. Deep Learning-based Aerial Image Segmentation with Open Data for Disaster Impact Assessment. *Neurocomputing* **2020**, *439*, 22–33. [CrossRef]
4. Zhang, Z.; Guo, W.; Li, M.; Yu, W. GIS-Supervised Building Extraction with Label Noise-Adaptive Fully Convolutional Neural Network. *IEEE Geosci. Remote. Sens. Lett.* **2020**, *17*, 2135–2139. [CrossRef]
5. Deng, W.; Shi, Q.; Li, J. Attention-Gate-Based Encode' Decoder Network for Automatical Building Extraction. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2021**, *14*, 2611–2620. [CrossRef]
6. Zhou, D.; Wang, G.; He, G.; Long, T.; Luo, B. Robust Building Extraction for High Spatial Resolution Remote Sensing Images with Self-Attention Network. *Sensors* **2020**, *20*, 7241. [CrossRef] [PubMed]
7. Zhao, W.; Ivanov, I.; Persello, C.; Stein, A. Building outline delineation: From aerial images to polygons with an improved end-to-end learning framework. *ISPRS Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci.* **2020**, *175*, 731–735. [CrossRef]
8. Wu, T.; Hu, Y.; Peng, L.; Chen, R. Improved Anchor-Free Instance Segmentation for Building Extraction from High-Resolution Remote Sensing Images. *Remote. Sens.* **2020**, *12*, 2910. [CrossRef]
9. Shi, Y.; Li, Q.; Zhu, X. Building segmentation through a gated graph convolutional neural network with deep structured feature embedding. *ISPRS J. Photogramm. Remote. Sens.* **2020**, *159*, 184–197. [CrossRef] [PubMed]
10. Qi, X.; Zhu, P.; Wang, Y.; Zhang, L.; Peng, J.-h.; Wu, M.; Chen, J.; Zhao, X.; Zang, N.; Mathiopoulos, P.T. MLRSNet: A Multi-label High Spatial Resolution Remote Sensing Dataset for Semantic Scene Understanding. *ISPRS J. Photogramm. Remote. Sens.* **2020**, *169*, 337–350. [CrossRef]
11. Li, Q.; Shi, Y.; Huang, X.; Zhu, X. Building Footprint Generation by Integrating Convolution Neural Network with Feature Pairwise Conditional Random Field (FPCRF). *IEEE Trans. Geosci. Remote. Sens.* **2020**, *58*, 7502–7519. [CrossRef]

12. Chen, Q.; Wang, L.; Waslander, S.L.; Liu, X. An end-to-end shape modeling framework for vectorized building outline generation from aerial images. *ISPRS J. Photogramm. Remote. Sens.* **2020**, *170*, 114–126. [CrossRef]
13. Zhang, Z.; Wang, Y. JointNet: A Common Neural Network for Road and Building Extraction. *Remote. Sens.* **2019**, *11*, 696. [CrossRef]
14. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2012**, *60*, 84–90. [CrossRef]
15. Sun, Y.; Xue, B.; Zhang, M.; Yen, G.; Lv, J. Automatically Designing CNN Architectures Using the Genetic Algorithm for Image Classification. *IEEE Trans. Cybern.* **2020**, *50*, 3840–3854. [CrossRef]
16. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
17. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
18. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
19. Zhu, Z.; Wang, Q.; Li, B.; Wu, W.; Yan, J.; Hu, W. Distractor-aware siamese networks for visual object tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 101–117.
20. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; Volume 1, pp. 6230–6239.
21. Liu, X.; Hu, Z.; Ling, H.; Cheung, Y.-M. MTFH: A Matrix Tri-Factorization Hashing Framework for Efficient Cross-Modal Retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 964–981. [CrossRef]
22. Bittner, K.; Cui, S.; Reinartz, P. Building extraction from remote sensing data using fully convolutional networks. *ISPRS Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci.* **2017**, *42*, 481–486. [CrossRef]
23. Hui, J.; Du, M.; Ye, X.; Qin, Q.; Sui, J. Effective Building Extraction From High-Resolution Remote Sensing Images With Multitask Driven Deep Neural Network. *IEEE Geosci. Remote. Sens. Lett.* **2019**, *16*, 786–790. [CrossRef]
24. Pan, X.; Yang, F.; Gao, L.; Chen, Z.; Zhang, B.; Fan, H.; Ren, J. Building Extraction from High-Resolution Aerial Imagery Using a Generative Adversarial Network with Spatial and Channel Attention Mechanisms. *Remote. Sens.* **2019**, *11*, 917. [CrossRef]
25. Choi, M.; Kim, H.-W.; Han, B.; Xu, N.; Lee, K.M. Channel Attention Is All You Need for Video Frame Interpolation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.
26. Zhang, H.; Goodfellow, I.J.; Metaxas, D.N.; Odena, A. Self-Attention Generative Adversarial Networks. In Proceedings of theInternational conference on machine learning, Long Beach, CA, USA, 9–15 June 2019.
27. Ji, S.; Wei, S.; Lu, M. Fully Convolutional Networks for Multisource Building Extraction From an Open Aerial and Satellite Imagery Data Set. *IEEE Trans. Geosci. Remote. Sens.* **2019**, *57*, 574–586. [CrossRef]
28. Mnih, V. *Machine Learning for Aerial Image Labeling*; University of Toronto: Toronto, ON, USA, 2013.
29. Wang, C.; Shen, Y.; Liu, H.; Zhao, K.; Xing, H.; Qiu, X. Building Extraction from High Resolution Remote Sensing Images by Adaptive Morphological Attribute Profile under Object Boundary Constraint. *Sensors* **2019**, *19*, 3737. [CrossRef]
30. Ma, W.; Wan, Y.; Li, J.; Zhu, S.; Wang, M. An Automatic Morphological Attribute Building Extraction Approach for Satellite High Spatial Resolution Imagery. *Remote. Sens.* **2019**, *11*, 337. [CrossRef]
31. Avudaiammal, R.; Elaveni, P.; Selvan, S.; Rajangam, V. Extraction of Buildings in Urban Area for Surface Area Assessment from Satellite Imagery based on Morphological Building Index using SVM Classifier. *J. Indian Soc. Remote. Sens.* **2020**, *48*, 1325–1344. [CrossRef]
32. Parape, C.D.; Premachandra, H.C.N.; Tamura, M. Optimization of structure elements for morphological hit-or-miss transform for building extraction from VHR airborne imagery in natural hazard areas. *Int. J. Mach. Learn. Cybern.* **2015**, *6*, 641–650. [CrossRef]
33. Niveetha, M.A.; Vidhya, R. Automatic Building Extraction Using Advanced Morphological Operations and Texture Enhancing. *Procedia Eng.* **2012**, *38*, 3573–3578. [CrossRef]
34. Turker, M.; Koc-San, D. Building extraction from high-resolution optical spaceborne images using the integration of support vector machine (SVM) classification, Hough transformation and perceptual grouping. *Int. J. Appl. Earth Obs. Geoinf.* **2015**, *34*, 58–69. [CrossRef]
35. Turlapaty, A.; Gokaraju, B.; Du, Q.; Younan, N.; Aanstoos, J. A Hybrid Approach for Building Extraction from Spaceborne Multi-Angular Optical Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2012**, *5*, 89–100. [CrossRef]
36. Hao, L.; Zhang, Y.; Cao, Z. Robust building boundary extraction method based on dual-scale feature classification and decision fusion with satellite image. *Int. J. Remote. Sens.* **2019**, *40*, 5497–5529. [CrossRef]
37. He, H.; Zhou, J.; Chen, M.; Chen, T.; Li, D.; Cheng, P. Building Extraction from UAV Images Jointly Using 6D-SLIC and Multiscale Siamese Convolutional Networks. *Remote. Sens.* **2019**, *11*, 1040. [CrossRef]
38. Zhang, Y.; Gong, W.; Sun, J.; Li, W. Web-Net: A Novel Nest Networks with Ultra-Hierarchical Sampling for Building Extraction from Aerial Imageries. *Remote. Sens.* **2019**, *11*, 1897. [CrossRef]
39. Yang, H.; Wu, P.; Yao, X.; Wu, Y.; Wang, B.; Xu, Y. Building Extraction in Very High Resolution Imagery by Dense-Attention Networks. *Remote. Sens.* **2018**, *10*, 1768. [CrossRef]

40. Xie, Y.; Zhu, J.; Cao, Y.; Feng, D.-J.; Hu, M.-j.; Li, W.; Zhang, Y.; Fu, L. Refined Extraction Of Building Outlines From High-Resolution Remote Sensing Imagery Based on a Multifeature Convolutional Neural Network and Morphological Filtering. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2020**, *13*, 1842–1855. [CrossRef]
41. Alshehhi, R.; Marpu, P.R.; Woon, W.L.; Dalla Mura, M. Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks. *ISPRS J. Photogramm. Remote. Sens.* **2017**, *130*, 139–149. [CrossRef]
42. Li, X.; Yao, X.; Fang, Y. Building-A-Nets: Robust Building Extraction From High-Resolution Remote Sensing Images With Adversarial Networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 3680–3687. [CrossRef]
43. Shrestha, S.; Vanneschi, L. Improved Fully Convolutional Network with Conditional Random Fields for Building Extraction. *Remote. Sens.* **2018**, *10*, 1135. [CrossRef]
44. 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]
45. Yuan, J. Learning Building Extraction in Aerial Scenes with Convolutional Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 2793–2798. [CrossRef]
46. Hao, L.; Zhang, Y.; Cao, Z. Active Cues Collection and Integration for Building Extraction with High-Resolution Color Remote Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2019**, *12*, 2675–2694. [CrossRef]
47. Kang, W.; Xiang, Y.; Wang, F.; You, H. EU-Net: An Efficient Fully Convolutional Network for Building Extraction from Optical Remote Sensing Images. *Remote. Sens.* **2019**, *11*, 2813. [CrossRef]
48. Liu, H.; Luo, J.; Huang, B.; Hu, X.; Sun, Y.; Yang, Y.; Xu, N.; Zhou, Y.N. DE-Net: Deep Encoding Network for Building Extraction from High-Resolution Remote Sensing Imagery. *Remote. Sens.* **2019**, *11*, 2380. [CrossRef]
49. 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]
50. Wei, S.; Ji, S.; Lu, M. toward Automatic Building Footprint Delineation from Aerial Images Using CNN and Regularization. *IEEE Trans. Geosci. Remote. Sens.* **2020**, *58*, 2178–2189. [CrossRef]
51. Zhang, Y.; Li, W.; Gong, W.; Wang, Z.; Sun, J. An Improved Boundary-Aware Perceptual Loss for Building Extraction from VHR Images. *Remote. Sens.* **2020**, *12*, 1195. [CrossRef]
52. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651.
53. Zhang, Z.; Liu, Q.; Wang, Y. Road extraction by deep residual u-net. *IEEE Geosci. Remote. Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
54. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Springer: Munich, Germany, 5–9 October 2015; pp. 234–241.
55. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 3146–3154.
56. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A. 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]
57. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context Encoding for Semantic Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7151–7160.
58. Zhou, Y.; Sun, X.; Zha, Z.; Zeng, W. Context-Reinforced Semantic Segmentation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 4041–4050.
59. Mnih, V. Machine Learning for Aerial Image Labeling. Available online: http://www.cs.toronto.edu/~{}vmnih/data/ (accessed on 3 May 2021).
60. Chen, Z.; Wang, C.; Li, J.; Xie, N.; Han, Y.; Du, J.-X. Reconstruction Bias U-Net for Road Extraction from Optical Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2021**, *14*, 2284–2294. [CrossRef]
61. Zang, Y.; Wang, C.; Yu, Y.; Luo, L.; Yang, K.; Li, J. Joint Enhancing Filtering for Road Network Extraction. *IEEE Trans. Geosci. Remote. Sens.* **2017**, *55*, 1511–1525. [CrossRef]
62. 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]
63. 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.
64. Zhu, Q.; Li, Z.-S.; Zhang, Y.; Guan, Q. Building Extraction from High Spatial Resolution Remote Sensing Images via Multiscale-Aware and Segmentation-Prior Conditional Random Fields. *Remote. Sens.* **2020**, *12*, 3983. [CrossRef]
65. Zhang, L.; Wu, J.; Fan, Y.; Gao, H.; Shao, Y. An Efficient Building Extraction Method from High Spatial Resolution Remote Sensing Images Based on Improved Mask R-CNN. *Sensors* **2020**, *20*, 1465. [CrossRef] [PubMed]
66. Ma, J.; Wu, L.; Tang, X.; Liu, F.; Zhang, X.; Jiao, L. Building Extraction of Aerial Images by a Global and Multi-Scale Encoder-Decoder Network. *Remote. Sens.* **2020**, *12*, 2350. [CrossRef]