Article

Joint Learning of Contour and Structure for Boundary-Preserved Building Extraction

Cheng Liao [1], Han Hu [1,*], Haifeng Li [2], Xuming Ge [1], Min Chen [1], Chuangnong Li [1] and Qing Zhu [1]

[1] Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu 611756, China; liaocheng@my.swjtu.edu.cn (C.L.); xuming.ge@swjtu.edu.cn (X.G.); minchen@home.swjtu.edu.cn (M.C.); lichuangnong123@sina.com (C.L.); zhuq66@263.net (Q.Z.)

[2] School of Geosciences and Info-Physics, Central South University, Changsha 410012, China; lihaifeng@csu.edu.cn

[*] Correspondence: han.hu@swjtu.edu.cn

Abstract: Most of the existing approaches to the extraction of buildings from high-resolution orthoimages consider the problem as semantic segmentation, which extracts a pixel-wise mask for buildings and trains end-to-end with manually labeled building maps. However, as buildings are highly structured, such a strategy suffers several problems, such as blurred boundaries and the adhesion to close objects. To alleviate the above problems, we proposed a new strategy that also considers the contours of the buildings. Both the contours and structures of the buildings are jointly learned in the same network. The contours are learnable because the boundary of the mask labels of buildings implicitly represents the contours of buildings. We utilized the building contour information embedded in the labels to optimize the representation of building boundaries, then combined the contour information with multi-scale semantic features to enhance the robustness to image spatial resolution. The experimental results showed that the proposed method achieved 91.64%, 81.34%, and 74.51% intersection over union (IoU) on the WHU, Aerial, and Massachusetts building datasets, and outperformed the state-of-the-art (SOTA) methods. It significantly improved the accuracy of building boundaries, especially for the edges of adjacent buildings. The code is made publicly available.

Keywords: structural features; building extraction; remote sensing image; semantic segmentation; deep learning

1. Introduction

Building extraction from high-resolution orthoimages is of great significance for applications ranging from urban land use to three-dimensional reconstruction [1,2]. Due to the complex background and diversity of building styles, the problem of building extraction suffers severe intra-class variability and small inter-class differences in remote sensing images [3,4]. Thus, it is still a great challenge to extract buildings efficiently and accurately in a complex urban environment.

With the advent of deep learning, building extraction methods that benefit from deep convolutional neural networks (DCNNs) have been improved substantially. Most of the existing approaches [5–7] use an encoder–decoder architecture to extract a pixel-wise building mask. The network is learned end-to-end: one end is the orthoimage and the other end is a pixel-wise building mask. Although the end-to-end paradigm has significantly improved the ease of use of supervised deep learning approaches, there are still some issues remaining to be resolved:

1. Contradiction of feature representation and spatial resolution. The encoder–decoder structure will produce more representative features in deeper layers, however, at the cost of coarser spatial resolution. This issue is widely acknowledged in the computer vision community. Although the decrease of spatial resolution may only cause issues for small objects in terrestrial or natural images, it causes blurry or zigzag

effects on building boundaries when applied to orthoimages, which is undesirable for subsequent applications. As column (b) of Figure 1 shows, the boundary of the buildings are inaccurately extracted, and the smallest buildings cannot be recognized.

2. Ignorance of building contours. Buildings have clear outlines, as opposed to the background, which is typically represented as the vector boundary in applications. However, the existing end-to-end paradigm generally ignores intuitive and important prior knowledge, even if the junction of the building and the background of manually labeled masks contains such vector outline information of the building. As a consequence, the extracted buildings are generally not regularized and adhesive in a complex urban environment. As Figure 1 shows, indicated by the yellow circle, column (b) extracted by MAP-Net [8], which proposed an independently parallel network that preserves multiscale spatial details and rich high-level semantics features, is inaccurate at the building boundary, specifically in adjacent areas. Even if the pixel-wise metrics such as the intersection over union (IoU) scores (as explicitly optimized in the end-to-end objective function) are relatively high, the above flaw still renders the results less useful.
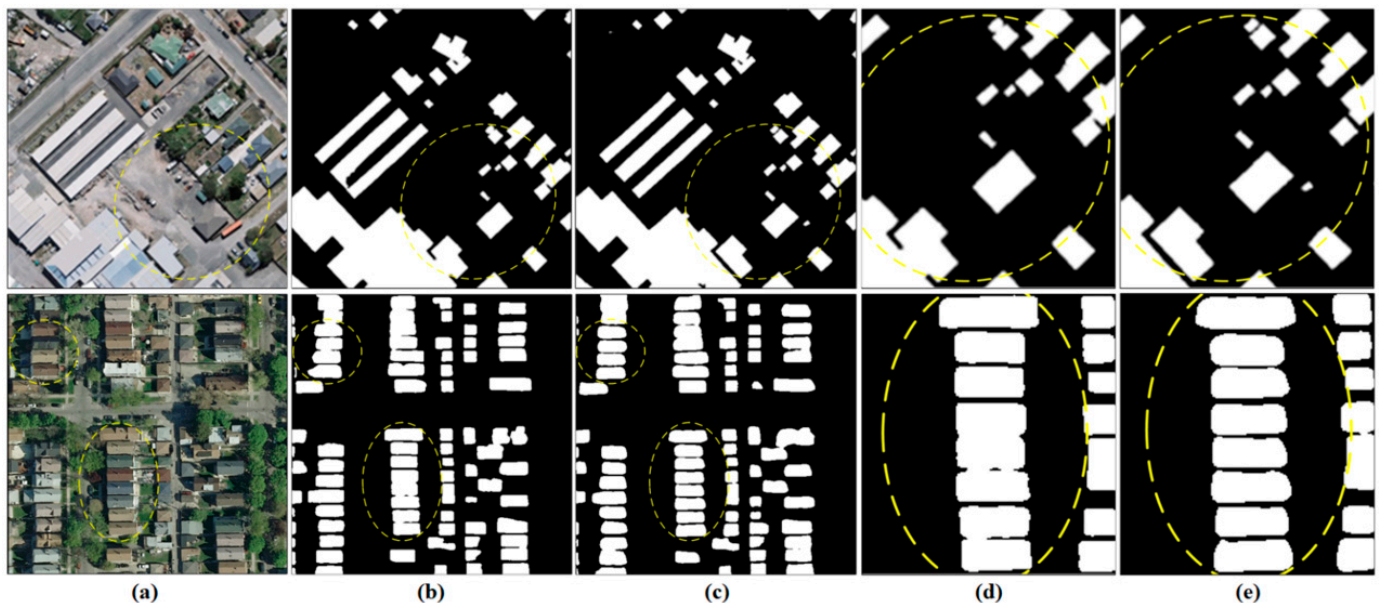

Figure 1. The building extracted by existing methods is zigzag and adhesive at the building boundary, specifically in adjacent areas. Columns (**a**–**c**) represent sample images, extracted results by MAP-Net, and ours, respectively. Columns (**d**,**e**) represent the details of the MAP-Net and ours.

To alleviate these problems, we proposed a boundary-preserved building extraction approach by jointly learning the contour and structure of buildings. We utilized the building contour information embedded in the labels to optimize the representation of building boundaries, then combined the contour information with multi-scale features extracted by the extractor of MAP-Net to alleviate the problems of edge blurring, zigzag, and boundary adhesion in existing methods. Specifically, the steps of this paper's method are as follows. First, we designed a building structure extraction module to learn the building geometric features, which ensures that the detailed boundary information is fully utilized. Second, we introduced dice loss in the structure feature extraction module to optimize the learned structural features and combine it with the cross-entropy loss for segmentation branching. Finally, we redesigned the multiscale feature extraction backbone, with more resolution to make it robust for remote sensing at different resolutions. Considering the large amounts of parameters introduced by the high-resolution feature, we replaced the convolution

layers with dilated convolutions [9] in an added feature extraction path to maintain a lower computational complexity.

In summary, our contributions are as follows: (1) We proposed a joint learning contour and structure information for a boundary-preserved building extraction method. (2) We proposed a structural feature constraint module that combines the structure information for refinement boundary extraction, especially for the edge of the dense, adjacent buildings. The rest of this paper is organized as follows. Section 2 summarizes the related work on building footprint extraction. The details of the proposed method for the refinement of building extraction are introduced in Section 3. Section 4 describes the experiments and analyses the results. The discussions and conclusions of this paper are presented in Section 5.

2. Related Works

Over the recent few decades, numerous building extraction algorithms have been proposed. They can be divided into traditional image processing-based and machine learning-based methods. Traditional building extraction methods utilize the threshold or design feature operators according to the characteristics of spectrum, texture, geometry, and shadow [10–16] to extract buildings from optical images. These methods can resolve only specific issues for specific data, since the feature operators vary with illumination conditions, sensor types, and building architecture. To relieve these problems, refs. [17–24] combined optical imagery with GIS data, digital surface models (DSMs) obtained from light detection and ranging (lidar), or synthetic aperture radar interferometry to distinguish non-buildings that are highly similar to buildings, increasing the robustness of building extraction, are used. However, obtaining a wide range of corresponding multisource data is always costly.

With the development of DCNNs in recent years, many algorithms have been proposed for processing remote sensing images [25–32]. The fully convolutional network [33] (FCN) replaces the fully connected layers with convolutional layers, making it possible for large-scale dense prediction. CNN-based methods [34–36] have been proposed to extract buildings from remote sensing data, but the details and boundaries of buildings are still inaccurate, because detailed spatial information is lost during repeated downsampling operations, which is difficult to recover.

To resolve this problem, encoder–decoder-based methods [5–7], represented by UNet [37], introduced skip connections to fuse high-resolution features extracted from shallow layers in the encoder stage during the decoder process to recover detailed information. These methods extract more details compared to the FCN-based methods, while noise can be introduced, since the high-resolution features are extracted from shallow layers. References [38–42] introduced many postprocesses for refining building boundaries, such as conditional random fields (CRF). Although the boundary could be more accurate, the complexity of the calculation was greatly increased. ResNet [43] improved the training stability and performance of a deeper CNN by introducing a residual connection module, which makes it possible to extract rich higher-level semantic features. It is often used as a backbone to achieve better performance in many tasks, such as in [44,45].

To improve the accuracy of multiscale building extraction, refs. [46–49] designed a module for multiscale input and fused multimodules to extract buildings with multiscales. It obviously improved the accuracy, but increased the complexity of the model, especially when processing a large number of remote sensing images. DeepLab [38,50] series methods introduced atrous convolution for segmentation tasks, which enlarged the receptive field without increasing the computational complexity. These methods could extract multiscale objects efficiently compared with [47,48]. Moreover, refs. [44,45,51,52] fused the multiscale features extracted from different stages to recover a detailed localization. MAP-Net extracts multiscale features through an independent parallel network that contains multiscale spatial details and rich high-level semantics, and achieved SOTA results on the WHU dataset.

Despite the most recent building extraction methods being able to obtain higher performance in many public benchmarks, the boundaries of buildings are rarely identified accurately, which is vital for many remote sensing applications. Many types of research have implied that edge information is potentially valuable for better segmentation boundaries. References [7,53,54] combined the superpixel segmentation result with deep features to enhance the segmentation boundary. Similarly, refs. [55–58] integrated the edge features extracted by an additional lightweight CNN from edge extraction operators, to improve the accuracy of the segmentation results. However, in research of CNN-based building extraction, inaccurate boundary identification, especially in the area of densely distributed buildings, is still unavoidable.

3. Methodology

3.1. Architecture Overview

To resolve the problem of the extracted results not being accurate at the building boundaries in existing methods, especially between building which are closely distributed, we proposed a boundary-preserved building extraction method by jointly learning the contour and structure of buildings. Our architecture is shown in Figure 2. It mainly contains two parts: a robust and efficient multiscale feature extraction backbone network, and a building structural constraint module. The backbone is inherited from the feature extractor of MAP-Net, due to its potential for preserving detailed multiscale features. First, we redesigned the multiscale feature extraction network for robustness to image resolution and a trade-off between accuracy and efficiency. Then, we designed a structural constrained module to learn contours from the gradient information for extracting refined building boundaries with a dice loss for structural feature optimization. The details of these parts are described in the following sections.
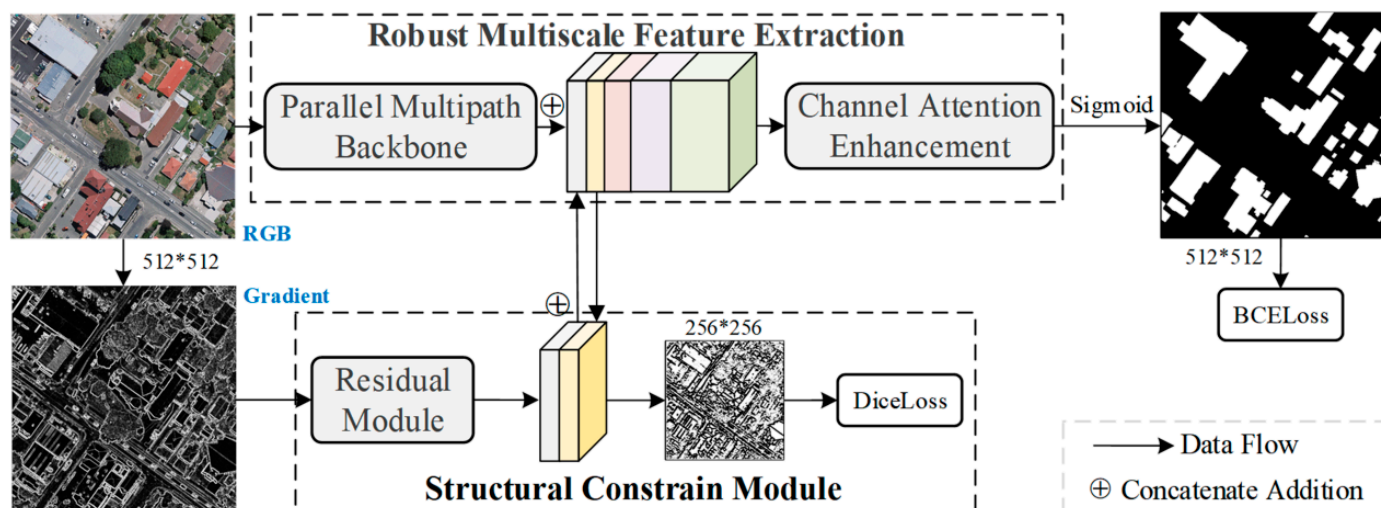

Figure 2. Structure of the proposed network. There are two main parts: a multiscale feature extraction backbone, and a building structural constraint module.

3.2. Structural Constraint Module

The existing building extraction methods cannot accurately identify boundaries, especially for buildings that are distributed closely adjacent. The CNN-based methods learn the weight of the convolution kernel from the features of each receptive field. The structural information at the edges of buildings can be underutilized.

Intuitively, this kind of boundary information is important for extracting accurate instances of boundaries between buildings. Based on the hypothesis that there is a significant difference between the building and the background at the junction, there is a high consistency in the area of the building instance or background. We introduced the

Sobel operator to capture the gradient information, since it has excellent characteristics for capturing the differences of the edges of buildings.

We designed a residual network branch to extract the structural features that capture the boundary information of buildings from the gradient, named the structural constraints module, as shown in Figure 3. The extracted features as a constraint were concatenated to the multiscale features extracted from the main segmentation branch. Similarly, the features with the highest resolution extracted from the main branch were concatenated to structural features as complementary information. In addition, we introduced the dice loss to evaluate the consistency between the learned structure features and the ground truth for optimizing the residual module, as described in Section 3.4. In this way, the extracted features retain the multiscale detailed information, and enhance the feature expression of the edge of the building.
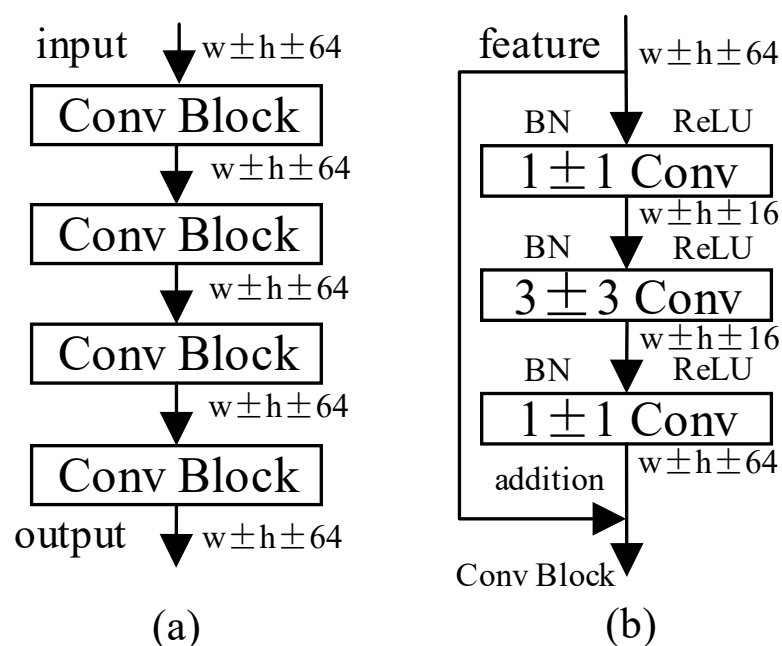
Figure 3. Details of the residual module as shown in (**a**). (**b**) represents the details of each convolutional block.

We visualized the learned features of the network without/with the structural constraint module for a clear understanding in Figure 4. The features have a higher response at the edge of the building, column (c) compared with column (b), which implies that the boundary was enhanced after the introduction of the structural features.

3.3. Robust Feature Extraction

MAP-Net extracts multiscale features with a parallel multipath network. The resolution of features in each path remains fixed and independent during extraction to contain detailed context and rich semantic information. The optimal combination of scales is adaptively fused based on channel attention-based modules, which improves the accuracy of the building boundaries and small buildings, since detailed contexts are kept in the features with higher resolution. While the limited scale of features is not optimal for remote sensing images with different resolutions, as the paper mentioned, higher resolution feature extraction causes considerable computational complexity.
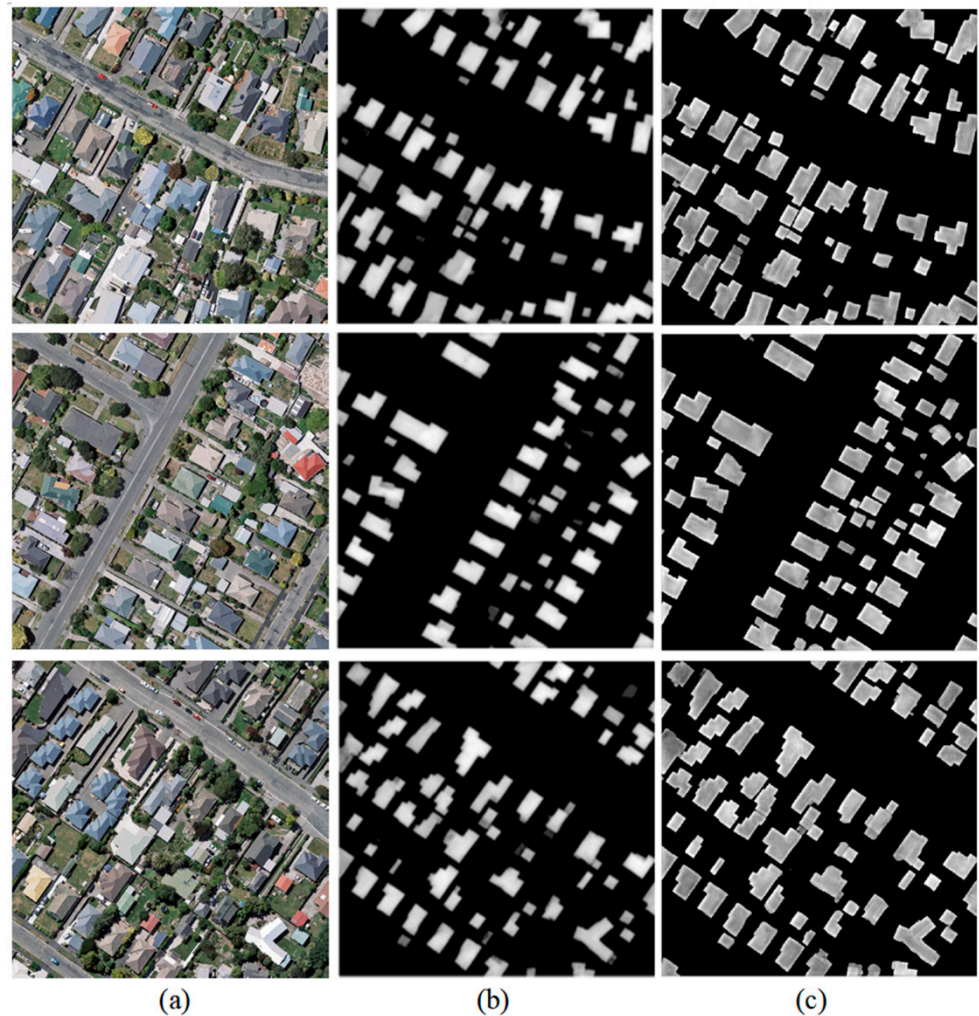
Figure 4. Comparison of the features extracted without/with the structural constraint (SC) module. (**a**): Original images; (**b**): features extracted without the SC module; (**c**): features extracted with the SC module.

In this paper, we inherited the structure of the multipath feature extraction network for buildings with multiple scales. Considering the different resolutions of remote sensing images, we added a path to extract features with half the resolution of the original image, which is vital for extracting buildings from low-resolution remote sensing images and identifying tiny buildings. Since the extracted features cover a greater scale space, the detailed contexts and semantics can be more abundant, which can enhance the robustness of the resolution. Additionally, we introduced dilated convolutions and omitted the spatial pyramid pooling module as a trade-off between the performance and complexity of the network.

3.4. Loss Function

For the loss function, we used binary entropy loss, which is presented as Formula (1) as follows, for the segmentation branch, since it has great performance for evaluating the distance between prediction and label, and has been widely used in building extraction networks. Where p_i represents predicted result for index i in the image, y_i represents the ground truth of pixel in the index i, and N represent the number of pixel in the image.

$$L_{bce} = \frac{1}{N}\sum_i L_i = -\frac{1}{N}\sum_i [y_i \cdot log(p_i) + (1 - y_i) \cdot log(1 - p_i)] \tag{1}$$

In addition, we introduced dice loss, which was first proposed in [59], so the structural features extracted from residual module have more consistency with the labels. As Formula (2) shows, p_i and y_i represent the predict result and ground truth of pixel i. The total loss is weighted by L_{bce} and L_{dice}, as shown in Formula (3). We conducted comparative experiments to compare different coefficients with a step of 0.1. The results showed little difference when α is between 0.3 and 0.5, and were superior to others, since we set α and β to 0.4 and 0.6 in our experiments.

$$L_{dice} = 1 - \frac{2\sum_i^N y_i \cdot p_i + 1}{\sum_i^N y_i^2 + \sum_i^N p_i^2 + 1} \tag{2}$$

$$Loss = \alpha \cdot L_{bce} + \beta \cdot L_{dice} \tag{3}$$

3.5. Evaluation Metrics

Semantic segmentation-based building footprint extraction from remote sensing images aims to label each pixel belong to the building or background for a specific input image. Generally, evaluation metric methodologies can be divided into pixel-level, and instance-level. To compare with most recent studies, we applied pixel-level metrics including precision, recall, F1-score, and IoU to evaluate the performance of our and other compared methods. Equations are given as follows:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{6}$$

$$IoU = \frac{Precision * Recall}{Precision + Recall - Precision * Recall} \tag{7}$$

4. Experiments and Analysis

4.1. Datasets

We evaluated the performance of the proposed method on three open datasets, including the WHU building dataset [60], the Inria aerial image labelling dataset (Aerial) [61], the and Massachusetts dataset [25]. The details of these datasets are described as follows.

The WHU building dataset includes both aerial and satellite subsets, with corresponding images and labels. We selected the aerial subset that has been widely used in existing works for comparison with the proposed algorithm. It covers a 450 km^2 area, with a 30 cm ground resolution. Each of the images has three bands, corresponding to red (R), green (G), and blue (B) wavelengths with a size of 512 $\times$ 512 pixels. There are 8188 tiles of images, including 4736, 2416, and 1036 tiles for the training, test, and validation datasets, respectively. We conducted our experiment with the original provided dataset partitioning.

The Aerial dataset contains 180 orthorectified RGB images for training with corresponding public labels, and 180 images for testing without public labels, covering 810 km^2 with a spatial resolution of 0.3 m. We only used the former in our experiments, which covers dissimilar urban settlements (Austin, Chicago, Kitsap, Tyrol, and Vienna). Similar to many existing works, we choose the first five images in each area for testing and the others for training, and we randomly clipped the images to a size of 512 $\times$ 512 pixels during the training stage.

The Massachusetts Building Dataset consists of 151 aerial images of the Boston area. The entire dataset covers approximately 340 square kilometers with a resolution of 1 m, and the size of each image is 1500 $\times$ 1500 pixels. There are 137 sets of aerial images and corresponding single-channel label images for training; 10 for the test, and four for

validation. We randomly clipped the training set to 512 × 512 pixels, while the validation and test sets were clipped to 512 × 512 for prediction, and merged to evaluate the accuracy.

4.2. Performance Comparison

We compared the performance of the proposed methods with some classical semantic segmentation methods, such as U-NetPlus [5], PSPNet [44], and DeepLabv3+ [50], as well as the most recent building extraction methods, such as MAP-Net, DE-Net, and MA-FCN, on the WHU, Aerial, and Massachusetts datasets.

U-NetPlus achieved a great improvement by redesigning the encoder according to VGG, and replacing the transposed convolution with the nearest interpolation based on UNet, which is widely used in remote sensing imagery segmentation. Since ResNet improved the training stability and performance of deeper CNNs by introducing residual connections, we implemented PSPNet with ResNet50 for feature extraction and replaced the upsampling module with ours for building extraction. In addition, DeepLabv3+ and MAP-Net were implemented as in the public source code.

Our method designed a structural constraint module to enhance the detailed information on the edge of the building. It outperformed the latest building extraction works and achieved SOTA results without pretraining and postprocessing on the three datasets with different spatial resolutions.

Our research was implemented in TensorFlow using a single 2080Ti GPU with 11 Gigabytes of memory. The Adam optimizer was chosen with an initial learning rate of 0.001, and beta1 and beta2 were set to default, as recommended. All compared methods in the experiment were trained from scratch for approximately 150 epochs to ensure that the training model converged, with random rotation and flipping for data augmentation on each dataset described in Section 4.1. The batch size was set to four, restricted by the GPU memory size, to ensure that the hyperparameters remained the same among the compared methods for equal performance evaluation.
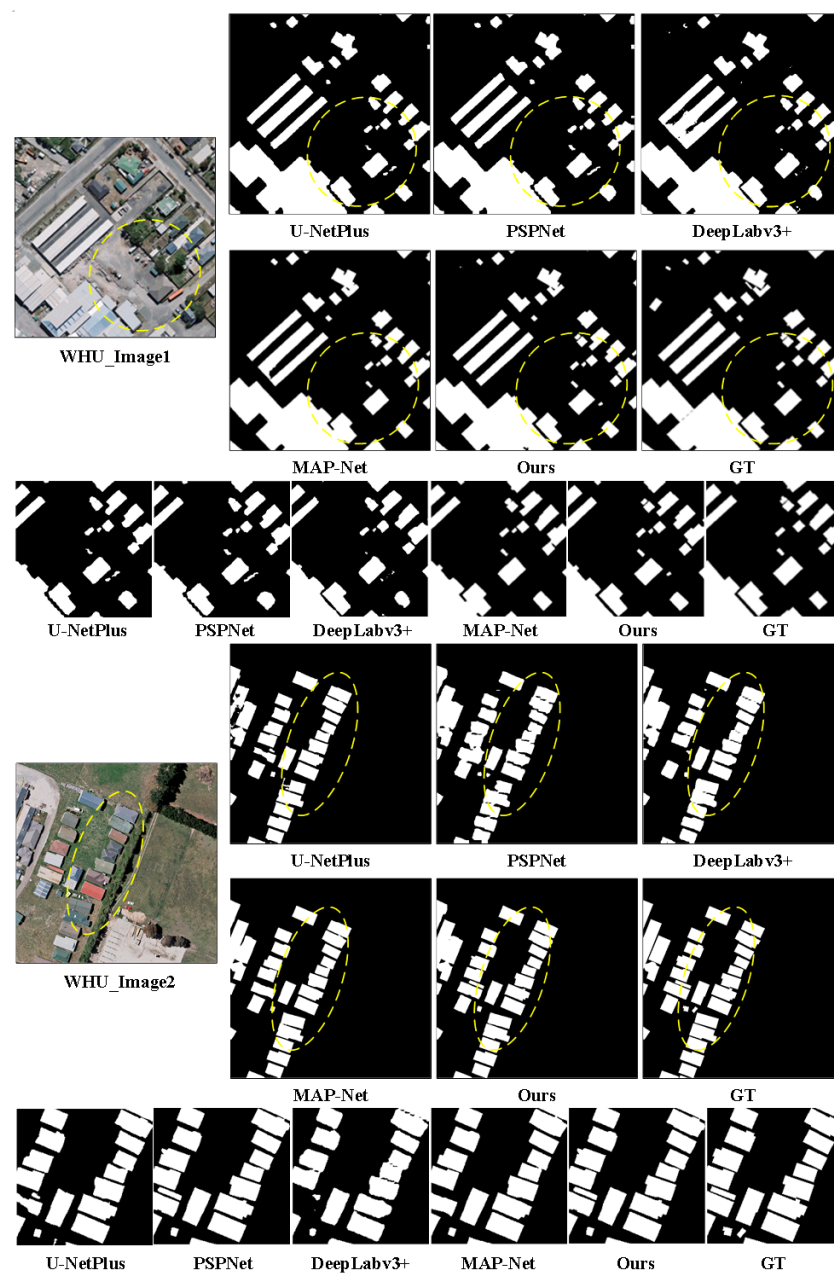
4.2.1. Experiments on the WHU Dataset

The WHU dataset has a higher spatial resolution, which means the intersection of the buildings and the backgrounds contains more accurate and valuable contour information. We conducted comparative experiments on the WHU dataset to compare the similarity of the building boundaries extracted by our, and related, methods.

SRI-Net [45] generated multiscale features by a spatial residual inception module based on the ResNet101 encoder. DE-Net [51] aggregated many lasting segmentation techniques, such as linear activation units, residual blocks, and densely upsampling convolutions to recover better spatial information. EU-Net [52] used deep spatial pyramid pooling (DSPP) and introduced focal loss to reduce the effect of incorrect labels. MA-FCN [40] introduced a feature pyramid network (FPN) for multiscale feature extraction, and a polygon regularization strategy for boundary optimization. The experimental results were as shown in Table 1. The results of the related methods marked as quoted are referenced from the corresponding papers, since the source code is unavailable.

Our method outperformed the compared methods, and achieved a 0.78% IoU improvement in pixel-level metrics compared with our previous work, MAP-Net. To compare the results extracted by our, and other compared, methods, we visualized examples of results on WHU datasets in Figure 5. There are two sample images and the corresponding extracted results through the related method. The partial details are shown in the last row for convenient comparison. The proposed method can extract buildings more accurately, as shown in the area marked by yellow dashed circles. In particular, the edges between buildings with dense, adjacent areas were identified more clearly.

Table 1. Performance evaluation of the compared methods and ours on the WHU test datasets.

Methods	IoU (%)	Precision (%)	Recall (%)	F1-Score (%)
UNet	89.51	95.11	93.83	94.47
PSPNet	88.87	94.28	93.93	94.10
DeepLabv3+	88.16	94.64	92.79	93.71
MAPNet	90.86	95.62	94.81	95.21
SRI-Net [45]	89.23	95.67	93.69	94.51
DE-Net [51]	90.12 + 0.24	95.00 + 0.16	94.60 + 0.19	94.80 + 0.18
EU-Net [52]	90.56	94.98	95.10	95.04
MA-FCN [40]	90.70	95.20	95.10	95.15
Ours	91.64	95.83	95.44	95.64

Figure 5. Examples of the extracted results on the WHU test dataset. The bottom column represents the corresponding partial details for convenience comparison.

4.2.2. Experiments on the Aerial Dataset

The buildings are distributed very densely in the Aerial dataset. Existing methods have difficultly in distinguishing the boundaries of building instances accurately, resulting in multiple adjacent buildings being adhering together in the extracted results. We conducted comparative experiments on the Aerial dataset to compare the segmentation results extracted by our model and the comparison methods at adjacent building boundaries. As shown in Table 2, the results of related methods marked as quoted are referenced from the corresponding papers, since the source code is unavailable.

Table 2. Performance evaluation of the compared methods and ours on the Aerial test dataset.

Methods	IoU (%)	Precision (%)	Recall (%)	F1-Score (%)
UNet	77.51	88.83	85.88	87.33
PSPNet	78.18	88.48	87.40	87.75
DeepLabv3+	73.83	87.27	82.75	84.95
MAPNet	80.33	89.61	88.58	89.09
SRI-Net [45]	71.76	85.77	81.46	83.56
EU-Net [52]	80.50	90.28	88.14	89.20
Ours	81.15	91.78	87.51	89.59

Our method outperformed the classical semantic methods and the most recent works. Especially, our method achieved a 0.82% IoU improvement in pixel-level metrics compared with MAP-Net, and 0.65% compared with the EU-Net, which, as far as we know, is the most recent published result on this dataset.

We also visualized examples of results on Aerial datasets in Figure 6 to compare the results extracted by our and other compared methods. As can be seen from the experimental results, our method could extract a more accurate building boundary than the compared methods, since we introduced learnable contour information to joint-optimize the building outline. Our proposed method distinguished the boundaries between adjacent building instances more accurately than the compared methods, especially for the densely distributed building, such as the results between our method and MAP-Net in the partial detail results.

4.2.3. Experiments on the Massachusetts Dataset

To validate the generalization of the proposed method on imagery with a different spatial resolution, we conducted experiments on the Massachusetts dataset to compare the performance of our and related works. BRRNet [62] designed a prediction module based on the encoder–decoder structure with a residual refinement module for accurate boundary extraction.

The results showed that our method outperformed the classical semantic segmentation methods and the most recent building extraction methods, as shown in Table 3. This implies that combining the contour information embedded in the building and the background with the semantic features could enhance the robustness of the spatial resolution of images.

We illustrate the example results on the Massachusetts datasets in Figure 7 for intuitive comparison. The proposed methods could identify more detailed instance boundaries, especially for buildings that were densely distributed, as shown in the first sample image. In addition, the outlines of the extracted building were also closer to the ground-truth than those without combining the contour information, as shown in the partial detail results.
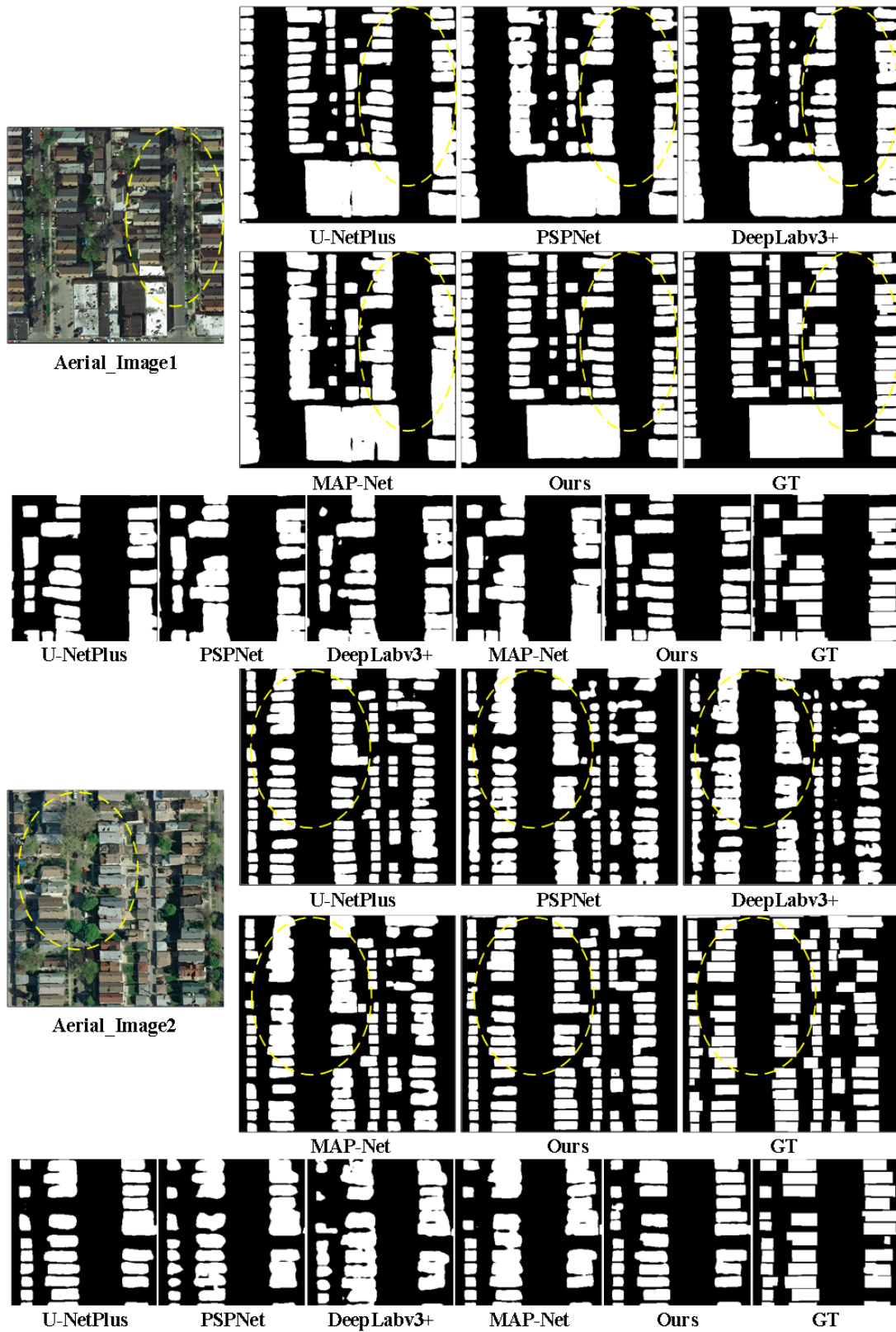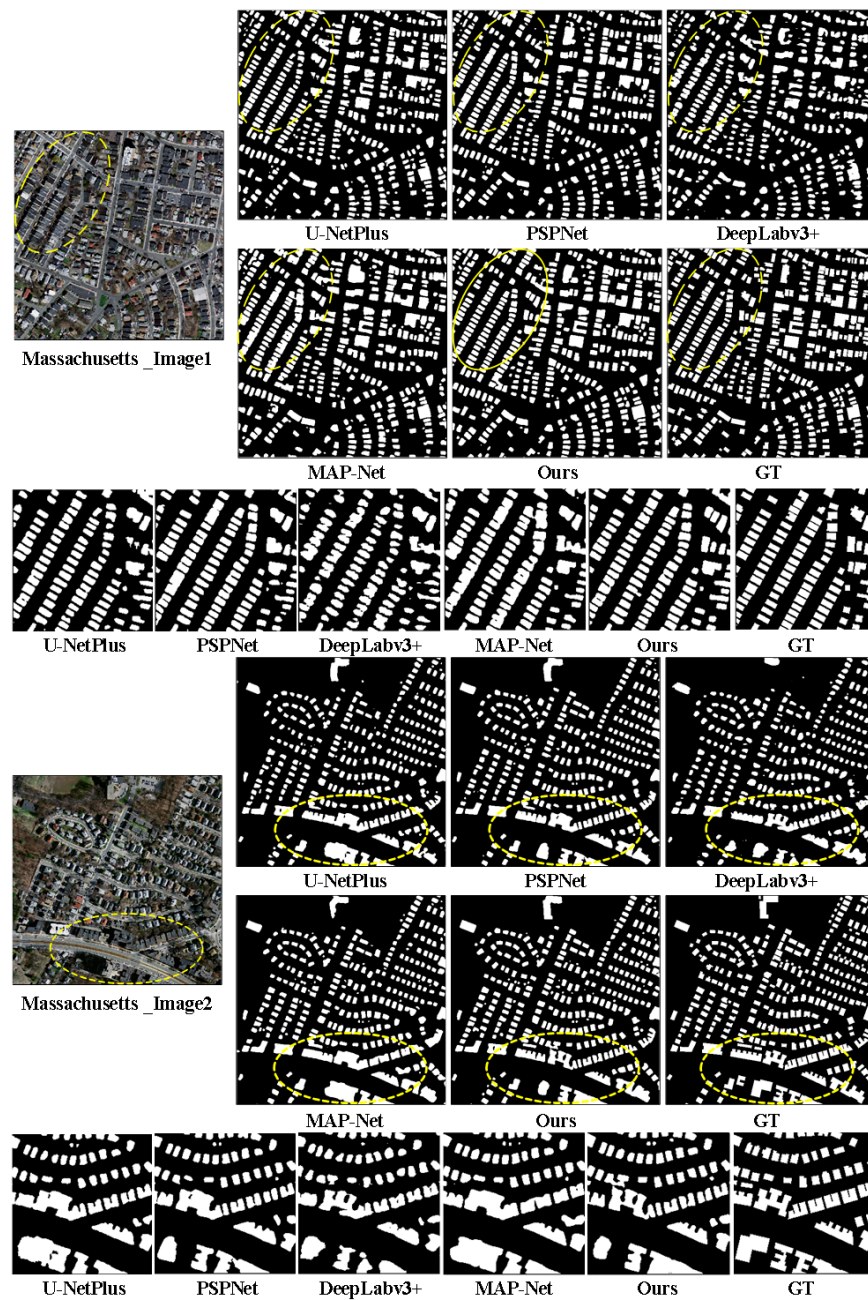
Figure 6. Example of extracted results on the Aerial test dataset. The bottom column represents the corresponding partial details for convenience of comparison.

Table 3. Performance evaluation of the compared methods and ours on the Massachusetts test dataset.

Methods	IoU (%)	Precision (%)	Recall (%)	F1-Score (%)
UNet	70.99	85.53	80.68	83.33
PSPNet	71.57	88.24	79.12	83.43
DeepLabv3+	66.18	82.84	76.69	79.65
MAPNet	73.34	85.49	83.76	84.62
EU-Net [52]	73.93	86.70	83.40	85.01
BRRNet [62]	74.46	-	-	85.36
Ours	74.51	85.44	85.34	85.39

Figure 7. Example of extracted results on the Massachusetts test dataset. The bottom column represents the corresponding partial details for convenience of comparison.

4.3. Analysis on Instance-Level Metrics

To verify whether the proposed method significantly improved the edge of extracted buildings, especially the boundary discrimination between adjacent buildings in dense areas, we also used instance-level metrics to verify the effectiveness of the proposed method in optimizing building instance boundary discrimination, compared with the pixel-level metrics, on the WHU, Aerial, and Massachusetts datasets.

The instance-level evaluation method is similar to the pixel-level method, except that the basic unit of calculation is the building instance instead of pixels. The building is identified correctly when the IoU between every building instance in the prediction results and the corresponding instance in the labels exceeds a certain threshold value. Usually, the higher the threshold is, the more stringent the requirement for accuracy in the assessment of the prediction result. The results are presented in Table 4, which shows the improved performance on instance-level F1 and IoU scores under different instance thresholds, from 0.4 to 0.8 on three datasets.

Table 4. The instance-level performance improvement between our method and MAP-Net, with different thresholds on the WHU, Aerial, and Massachusetts datasets.

Data Sets	TH Methods	0.4		0.5		0.6		0.7		0.8	
		F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU
WHU	MAP-Net	90.70	82.99	89.48	80.97	87.68	78.06	84.84	73.67	78.80	65.02
	Ours	91.66	84.59	90.66	82.92	89.22	80.55	86.80	76.68	81.46	68.73
	Improved	+0.96	+1.60	+1.18	+1.95	+1.54	+2.49	+1.96	+3.01	+2.66	+3.71
Aerial	MAP-Net	79.56	66.06	75.53	60.68	70.98	55.02	62.81	45.78	45.57	29.51
	Ours	81.87	69.30	77.95	63.87	73.75	58.41	66.40	49.70	49.31	32.72
	Improved	+2.31	+3.24	+2.42	+3.19	+2.77	+3.39	+3.59	+3.92	+3.74	+3.21
Massa	MAP-Net	87.29	77.44	82.65	70.44	73.67	58.31	53.01	36.06	20.49	11.41
	Ours	89.64	81.23	85.48	74.65	77.49	63.26	57.94	40.79	24.11	13.70
	Improved	+2.35	+3.79	+2.83	+4.21	+3.82	+4.95	+4.93	+4.73	+3.62	+2.29

The results implied that our proposed method achieved greater accuracy improvement in the instance-level metric than in the pixel-level metric, which is reflected in Table 3. In addition, the improved accuracy grew with the threshold, especially in higher resolution remote sensing imagery, such as the WHU dataset. This can be explained by, the higher the resolution of the images, the more separable the boundaries between buildings, as well as the more accurate identification of tiny buildings. Since our research introduced gradient information to enhance building boundary response, this information becomes weaker in low-resolution images for unclear building contours. This probably explains why when the threshold becomes extremely large, for 0.8 example, the improved instance-level accuracy becomes reduced on the Massachusetts dataset. We display the improved instance-level performance in F1 and IoU metrics on the WHU, Aerial, and Massachusetts datasets in Figure 8, for a more obvious representation.

4.4. Ablation Experiments

To evaluate the effectiveness of the different modules involved in the proposed method, we designed an ablation experiment to compare the contributions with precision, recall, F1-Score, and IoU metrics on the WHU dataset. The results are shown in Table 5.
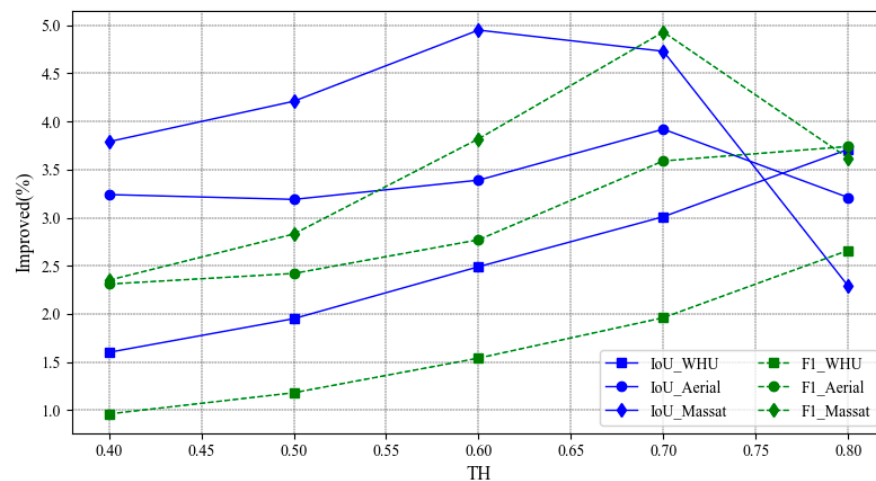
Figure 8. Visualization of the instance-level performance improvement with different thresholds on the three datasets.

Table 5. Comparison the performance of the modules in the proposed methods on the WHU dataset.

Methods	Structural	Multi-Loss	IoU	Precision	Recall	F1-Score
MAP-Net [8]	×	×	90.86	95.62	94.81	95.21
Baseline	×	×	91.00	95.21	95.36	95.29
Baseline-S	√	×	91.43	95.95	95.10	95.52
BaseLine+S+M	√	√	91.64	95.83	95.44	95.64

First, our baseline was based on MAP-Net, which was composed of four paths that extract features with different resolutions and introduce dilated convolution, which captures more global context during multiscale feature extraction, and reduces computational complexity. It slightly surpassed MAP-Net, with a lightweight network, since the extracted features cover more scales, and make it more robust to image resolution. Second, a parallel prior structural constraints module (S) was combined with the baseline, which was represented as Baseline+S, and achieved 0.43% IoU improvement on the baseline. This means that the introduced low-level geometric prior has a greater impact on building extraction since it has stronger generalizability than the RGB values. Finally, the proposed method introduced multiloss, which combines CE loss for segmentation and dice loss for the structural constraint module, based on Baseline+S. It enhanced the IoU metric since the dice loss promotes the optimization of structural features extracted by the prior structural constraint module.

In summary, the proposed method achieved a great improvement compared with MAP-Net, since robust feature extraction, generalized structural information, and edges contributed to the optimal function. The prior structural constraints had the greatest contribution to the extraction of the generalized context.

4.5. Efficiency Comparison

According to the above experiments, the proposed method obtained a better accuracy than the most recent related building extraction methods on the three datasets. To verify the timing performance, we calculated the trainable parameters and the number of floating-point operations (FLOPs) of the compared methods. The experimental results are shown in Table 6.

Table 6. Comparison of the efficiency of the related methods on the WHU test dataset.

Methods	IoU (%)	FLOPs (M)	Parameters (M)
U-NetPlus	89.51	17.28	8.64
PSPNet	88.87	93.48	46.72
DeepLabv3+	88.16	51.40	25.63
MAP-Net	90.86	48.09	24.00
Ours	91.64	49.42	24.55

Although the U-NetPlus was the most efficient, its accuracy was much lower than ours. Comparing our method with the MAP-Net, the number of trainable network parameters had no significant increase. However, the proposed method significantly improved the IoU accuracy of building recognition on the WHU test dataset.

In summary, although the backbone of the proposed method extracts multi-scale features to improve the robustness of varied image resolutions, it does this without introducing higher computational complexity due to the introduction of dilated convolution. The proposed method introduced building structural information and achieved better performance without significantly increasing computational complexity, since the structural constraint module is lightweight.

5. Discussion and Conclusions

This paper inherited the structure of MAP-Net and improved the trade-off between accuracy and computational complexity to extract building footprints accurately, especially for the boundaries of densely distributed building instances. To relieve the problem of not being able to exactly detect the building boundary in recent methods, we first designed the structural prior constraints module to enhance the feature representation of the building edges, under the hypothesis that the buildings have great differences from the background, especially between the boundaries. Furthermore, we introduced dice loss to optimize the representation of geometric features in the S module combined with the CE loss in the segmentation branch.

We explored the effectiveness of the proposed method through sufficient experimental demonstration and proved that the proposed method outperforms the most recent building extraction methods on the WHU, Aerial, and Massachusetts datasets, and achieved SOTA results. In addition, we evaluated the improved performance of our method with instance-level metrics to prove its ability to extract fine building boundaries. This research performed sufficient experiments on building extraction and implied that some low-level features could be considered for better generalization. We will perform further studies on land-cover classification on a large scale in the future. The source code is available at https://github.com/liaochengcsu/jlcs-building-extracion (accessed on 8 February 2021).

Author Contributions: Conceptualization, C.L. (Cheng Liao) and H.H.; methodology, C.L. (Cheng Liao); software, C.L. (Cheng Liao); validation, C.L. (Cheng Liao); formal analysis, C.L. (Cheng Liao); investigation, X.G., M.C., and C.L. (Chuangnong Li); resources, C.L. (Cheng Liao); data curation, C.L. (Cheng Liao); writing—original draft preparation, C.L. (Cheng Liao); writing—review and editing, H.H.; visualization, C.L. (Cheng Liao); supervision, H.H, H.L., and Q.Z.; project administration, H.H.; funding acquisition, H.H and Q.Z. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in reference number [25,60,61].

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

References

1. Jun, W.; Qiming, Q.; Xin, Y.; Jianhua, W.; Xuebin, Q.; Xiucheng, Y. A Survey of Building Extraction Methods from Optical High Resolution Remote Sensing Imagery. *Remote Sens. Technol. Appl.* **2016**, *31*, 653–662.
2. Mayer, H. Automatic Object Extraction from Aerial Imagery—A Survey Focusing on Buildings. *Comput. Vis. Image Underst.* **1999**, *74*, 138–149. [CrossRef]
3. 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]
4. 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]
5. Hasan, S.; Linte, C.A. U-NetPlus: A modified encoder-decoder U-Net architecture for semantic and instance segmentation of surgical instrument. *arXiv* **2019**, arXiv:1902.08994.
6. Diakogiannis, F.I.; Waldner, F.; Caccetta, P.; Wu, C. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 94–114. [CrossRef]
7. Gharibbafghi, Z.; Tian, J.; Reinartz, P. Modified superpixel segmentation for digital surface model refinement and building ex-traction from satellite stereo imagery. *Remote Sens.* **2018**, *10*, 1824. [CrossRef]
8. Zhu, Q.; Liao, C.; Hu, H.; Mei, X.; Li, H. MAP-Net: Multiple Attending Path Neural Network for Building Footprint Extraction From Remote Sensed Imagery. *IEEE Trans. Geosci. Remote Sens.* **2020**, 1–13. [CrossRef]
9. Xie, S.; Tu, Z. Holistically-Nested Edge Detection. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 11–18 December 2015; pp. 1395–1403.
10. Cote, M.; Saeedi, P. Automatic Rooftop Extraction in Nadir Aerial Imagery of Suburban Regions Using Corners and Variational Level Set Evolution. *IEEE Trans. Geosci. Remote Sens.* **2012**, *51*, 313–328. [CrossRef]
11. Li, Z.; Shi, W.; Wang, Q.; Miao, Z. Extracting Man-Made Objects From High Spatial Resolution Remote Sensing Images via Fast Level Set Evolutions. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 883–899. [CrossRef]
12. Liasis, G.; Stavrou, S. Building extraction in satellite images using active contours and colour features. *Int. J. Remote Sens.* **2016**, *37*, 1127–1153. [CrossRef]
13. Li, Q.; Wang, Y.; Liu, Q.; Wang, W. Hough Transform Guided Deep Feature Extraction for Dense Building Detection in Remote Sensing Images. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018.
14. 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]
15. Wang, X.; Li, P. Extraction of urban building damage using spectral, height and corner information from VHR satellite images and airborne LiDAR data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 322–336. [CrossRef]
16. Zhang, K.; Chen, H.; Xiao, W.; Sheng, Y.; Su, D.; Wang, P. Building Extraction from High-Resolution Remote Sensing Images Based on GrabCut with Automatic Selection of Foreground and Background Samples. *Photogramm. Eng. Remote Sens.* **2020**, *86*, 235–245. [CrossRef]
17. Huang, Z.; Cheng, G.; Wang, H.; Li, H.; Shi, L.; Pan, C. Building extraction from multi-source remote sensing images via deep deconvolution neural networks. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016.
18. Kemker, R.; Salvaggio, C.; Kanan, C. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 60–77. [CrossRef]
19. Sun, Y.; Zhang, X.; Xin, Q.; Huang, J. Developing a multi-filter convolutional neural network for semantic segmentation using high-resolution aerial imagery and LiDAR data. *ISPRS J. Photogramm. Remote Sens.* **2018**, *143*, 3–14. [CrossRef]
20. Bittner, K.; Adam, F.; Cui, S.; Korner, M.; Reinartz, P. Building Footprint Extraction From VHR Remote Sensing Images Combined With Normalized DSMs Using Fused Fully Convolutional Networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 2615–2629. [CrossRef]
21. Feng, M.; Zhang, T.; Li, S.; Jin, G.; Xia, Y. An improved minimum bounding rectangle algorithm for regularized building boundary extraction from aerial LiDAR point clouds with partial occlusions. *Int. J. Remote Sens.* **2019**, *41*, 300–319. [CrossRef]
22. Zhang, S.; Han, F.; Bogus, S.M. Building Footprint and Height Information Extraction from Airborne LiDAR and Aerial Imagery. In *Construction Research Congress 2020: Computer Applications*; American Society of Civil Engineers: Reston, VA, USA, 2020; pp. 326–335.
23. Dey, E.K.; Awrangjeb, M.; Stantic, B. Outlier detection and robust plane fitting for building roof extraction from LiDAR data. *Int. J. Remote Sens.* **2020**, *41*, 6325–6354. [CrossRef]
24. Zhang, P.; Du, P.; Lin, C.; Wang, X.; Li, E.; Xue, Z.; Bai, X. A Hybrid Attention-Aware Fusion Network (HAFNet) for Building Extraction from High-Resolution Imagery and LiDAR Data. *Remote Sens.* **2020**, *12*, 3764. [CrossRef]
25. Mnih, V. *Machine Learning for Aerial Image Labeling*; University of Toronto: Toronto, ON, Canada, 2013.

26. Chen, L.; Zhu, Q.; Xie, X.; Hu, H.; Zeng, H. Road extraction from VHR remote-sensing imagery via object segmentation con-strained by Gabor features. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 362. [CrossRef]
27. Guo, M.; Liu, H.; Xu, Y.; Huang, Y. Building Extraction Based on U-Net with an Attention Block and Multiple Losses. *Remote Sens.* **2020**, *12*, 1400. [CrossRef]
28. Zhu, Q.; Zhang, J.; Ding, Y.; Liu, M.; Li, Y.; Feng, B.; Miao, S.; Yang, W.; He, H.; Zhu, J. Semantics-Constrained Advantageous Information Selection of Multimodal Spatiotemporal Data for Landslide Disaster Assessment. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 68. [CrossRef]
29. Zhou, D.; Wang, G.; He, G.; Long, T.; Yin, R.; Zhang, Z.; Chen, S.; Luo, B. Robust Building Extraction for High Spatial Resolution Remote Sensing Images with Self-Attention Network. *Sensors* **2020**, *20*, 7241. [CrossRef]
30. Wagner, F.H.; Dalagnol, R.; Tarabalka, Y.; Segantine, T.Y.; Thomé, R.; Hirye, M. U-Net-Id, an Instance Segmentation Model for Building Extraction from Satellite Images—Case Study in the Joanópolis City, Brazil. *Remote Sens.* **2020**, *12*, 1544. [CrossRef]
31. Zhang, Y.; Zhu, J.; Zhu, Q.; Xie, Y.; Li, W.; Fu, L.; Zhang, J.; Tan, J. The construction of personalized virtual landslide disaster environments based on knowledge graphs and deep neural networks. *Int. J. Digit. Earth* **2020**, *13*, 1637. [CrossRef]
32. Zhu, Q.; Chen, L.; Hu, H.; Pirasteh, S.; Li, H.; Xie, X. Unsupervised Feature Learning to Improve Transferability of Landslide Susceptibility Representations. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 3917–3930. [CrossRef]
33. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
34. 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**, *XLII-1/W1*, 481–486. [CrossRef]
35. Wu, G.; Shao, X.; Guo, Z.; Chen, Q.; Yuan, W.; Shi, X.; Xu, Y.; Shibasaki, R. Automatic Building Segmentation of Aerial Imagery Using Multi-Constraint Fully Convolutional Networks. *Remote Sens.* **2018**, *10*, 407. [CrossRef]
36. 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]
37. 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.
38. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
39. Pan, X.; Zhao, J.; Xu, J. An End-to-End and Localized Post-Processing Method for Correcting High-Resolution Remote Sensing Classification Result Images. *Remote Sens.* **2020**, *12*, 852. [CrossRef]
40. 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]
41. Li, E.; Femiani, J.; Xu, S.; Zhang, X.; Wonka, P. Robust Rooftop Extraction From Visible Band Images Using Higher Order CRF. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 4483–4495. [CrossRef]
42. Huang, X.; Yuan, W.; Li, J.; Zhang, L. A new building extraction postprocessing framework for high-spatial-resolution re-mote-sensing imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *10*, 654–668. [CrossRef]
43. He, K.; Zhang, X.; Ren, S.; Sun, J. *Identity Mappings in Deep Residual Networks*; European Conference on Computer Vision; Springer: Berlin/Heidelberg, Germany, 2016; pp. 630–645.
44. 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; pp. 2881–2890.
45. 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]
46. Sun, G.; Huang, H.; Zhang, A.; Li, F.; Zhao, H.; Fu, H. Fusion of Multiscale Convolutional Neural Networks for Building Extraction in Very High-Resolution Images. *Remote Sens.* **2019**, *11*, 227. [CrossRef]
47. Li, L.; Liang, J.; Weng, M.; Zhu, H. A Multiple-Feature Reuse Network to Extract Buildings from Remote Sensing Imagery. *Remote Sens.* **2018**, *10*, 1350. [CrossRef]
48. Ji, S.; Wei, S.; Lu, M. A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery. *Int. J. Remote Sens.* **2019**, *40*, 3308–3322. [CrossRef]
49. Yu, Y.; Ren, Y.; Guan, H.; Li, D.; Yu, C.; Jin, S.; Wang, L. Capsule Feature Pyramid Network for Building Footprint Extraction From High-Resolution Aerial Imagery. *IEEE Geosci. Remote Sens. Lett.* **2020**, 1–5. [CrossRef]
50. 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.
51. Liu, H.; Luo, J.; Huang, B.; Hu, X.; Sun, Y.; Yang, Y.; Xu, N.; Zhou, N. DE-Net: Deep Encoding Network for Building Extraction from High-Resolution Remote Sensing Imagery. *Remote Sens.* **2019**, *11*, 2380. [CrossRef]
52. 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]

53. Huang, H.; Sun, G.; Zhang, A.; Hao, Y.; Rong, J.; Zhang, L. Combined Multiscale Convolutional Neural Networks and Super-pixels for Building Extraction in Very High-Resolution Images. In Proceedings of the IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 7415–7418.

54. Derksen, D.; Inglada, J.; Michel, J. Geometry Aware Evaluation of Handcrafted Superpixel-Based Features and Convolutional Neural Networks for Land Cover Mapping Using Satellite Imagery. *Remote Sens.* **2020**, *12*, 513. [CrossRef]

55. Marmanis, D.; Schindler, K.; Wegner, J.D.; Galliani, S.; Datcu, M.; Stilla, U. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS J. Photogramm. Remote Sens.* **2018**, *135*, 158–172. [CrossRef]

56. Cheng, D.; Meng, G.; Xiang, S.; Pan, C. FusionNet: Edge Aware Deep Convolutional Networks for Semantic Segmentation of Remote Sensing Harbor Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 5769–5783. [CrossRef]

57. Chen, L.-C.; Barron, J.T.; Papandreou, G.; Murphy, K.; Yuille, A.L. Semantic Image Segmentation with Task-Specific Edge Detection Using CNNs and a Discriminatively Trained Domain Transform. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.

58. Takikawa, T.; Acuna, D.; Jampani, V.; Fidler, S. Gated-SCNN: Gated Shape CNNs for Semantic Segmentation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.

59. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016.

60. 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]

61. 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 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; IEEE: New York, NY, USA, 2017; Volume 2017, pp. 3226–3229.

62. Shao, Z.; Tang, P.; Wang, Z.; Saleem, N.; Yam, S.; Sommai, C. BRRNet: A Fully Convolutional Neural Network for Automatic Building Extraction From High-Resolution Remote Sensing Images. *Remote Sens.* **2020**, *12*, 1050. [CrossRef]