remote sensing

MDPI

Article

Hybrid Task Cascade-Based Building Extraction Method in Remote Sensing Imagery

Runqin Deng [1], Meng Zhou [1,*], Yinni Huang [1] and Wei Tu [2]

[1] School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen 518107, China; dengrq5@mail2.sysu.edu.cn (R.D.); huangyn75@mail2.sysu.edu.cn (Y.H.)
[2] Department of Urban Informatics, School of Architecture and Urban Planning, Shenzhen University, Shenzhen 518060, China; tuwei@szu.edu.cn
* Correspondence: zhoum89@mail.sysu.edu.cn

Abstract: Instance segmentation has been widely applied in building extraction from remote sensing imagery in recent years, and accurate instance segmentation results are crucial for urban planning, construction and management. However, existing methods for building instance segmentation (BSI) still have room for improvement. To achieve better detection accuracy and superior performance, we introduce a Hybrid Task Cascade (HTC)-based building extraction method, which is more tailored to the characteristics of buildings. As opposed to a cascaded improvement that performs the bounding box and mask branch refinement separately, HTC intertwines them in a joint multilevel process. The experimental results also validate its effectiveness. Our approach achieves better detection accuracy compared to mainstream instance segmentation methods on three different building datasets, yielding outcomes that are more in line with the distinctive characteristics of buildings. Furthermore, we evaluate the effectiveness of each module of the HTC for building extraction and analyze the impact of the detection threshold on the model's detection accuracy. Finally, we investigate the generalization ability of the proposed model.

Keywords: deep learning; remote sensing; building extraction; instance segmentation; hybrid task cascade

Citation: Deng, R.; Zhou, M.; Huang, Y.; Tu, W. Hybrid Task Cascade-Based Building Extraction Method in Remote Sensing Imagery. *Remote Sens.* **2023**, *15*, 4907. https://doi.org/10.3390/rs15204907

Academic Editors: Qian Du, Yanni Dong and Xiaochen Yang

Received: 6 September 2023
Revised: 28 September 2023
Accepted: 7 October 2023
Published: 11 October 2023

1. Introduction

The ever-growing cities have witnessed drastic changes and exerted great influence on people's daily lives. Mapping the landscapes of cities is essential for better understanding the urban space and human activities, facilitating more informed decision making in urban policies. The high spatial resolution (HSR) remote sensing imagery contains abundant and detailed land cover information [1]. As one of the main artificial features in a city, the extraction and analysis of the building information provide valuable insights for a wide range of geographic and environmental applications. The automatic building extraction from high-resolution aerial images is currently an active research area and an issue of high importance to many urban scenarios, including disaster assessment, humanitarian aid, change detection in human settlements, urban planning and so on [2–6]. However, for building identification, it remains a challenge to fully leverage the inherent characteristics of building targets and different hierarchical features in building remote sensing images to obtain more accurate building locations.

The appearance of building rooftops in remote sensing images varies due to many factors, including lighting conditions, variety of reflections and diversity of image resolution [7–9]. Moreover, compared with natural images, the spatial scale of high-resolution remote sensing building objects in urban scenes varies greatly, especially for building rooftops in different shapes and sizes. These characteristics raise challenges to the accurate identification of building objects [10]. Previous traditional research usually relied on artificial design features, such as edge, spectral, shape, and texture features to extract buildings

from remote sensing imagery [11]. Some scholars also combined these handcrafted features with machine learning methods such as random forest (RF) [12] and support vector machines (SVM) [13] for building detection and classification. However, these approaches are labor-intensive with time-consuming tasks and their performance depends on the low-level hand-engineered local features.

Recent advancements in artificial intelligence (AI) have led to the growing popularity of deep learning algorithms for object detection tasks. In particular, deep Convolutional Neural Networks (CNNs) have been extensively utilized to extract objects from remote sensing images. The deep learning approach is capable of automatic learning and feature extraction from datasets. It also produces more stable results with stronger universality. In recent years, deep learning models have been increasingly adopted for building extraction from remote-sensed imagery. Relevant studies mainly made use of state-of-the-art models, e.g., Faster R-CNN [14], FCN [15], and Mask R-CNN [16], for a wide variety of purposes including object detection, semantic segmentation, and instance segmentation. Among them, Mask R-CNN-based methods have received growing attention in building recognition tasks. This is due to their ability to provide more detailed and accurate localization and segmentation results for the detected targets. More recently, cascade structure networks have been widely applied due to their ability to achieve higher detection accuracy and provide promising alternatives to the existing deep learning models. Moreover, the application of cascade structure networks in the remote sensing image analysis contributes to addressing specific challenges, such as detecting small objects and handling blurred object boundaries in target detection. However, cascade structure models have rarely been used in building object detections from remote sensing imagery. To address the challenges in automatic building extraction, we introduce the hybrid task cascade network (HTC) [17], which combines the advantages of Mask R-CNN and Cascade R-CNN [18] and propose a HTC-based automatic building detection method to extract building information from remote sensing imagery. By incorporating global semantic context (GSC), bounding box information (BBI) and mask information (MI), it can obtain more accurate, regular and smoother results of the building detection. The main contributions of this study are listed as follows:

1. This study introduces the hybrid task cascade network for building detection from remote sensing imagery based on the inherent characteristics of building targets. We also investigate the impact of global semantic features on the performance of building object detection.
2. We compare the influence of the integrated architecture in HTC, which makes the bounding box prediction and mask prediction intertwined, and the mask information flow architecture, which combines mask information at different stages of the cascade structure to improve the accuracy of mask detection.
3. Extensive experiments are conducted on three diverse remote sensing imagery building datasets, including CrowdAI mapping challenge dataset [19], WHU aerial image dataset [20]) and Chinese Typical City Buildings Dataset [21]. The results demonstrate the superiority of the HTC network over the swin-transformer-based method and other existing state-of-the-art (SOTA) segmentation methods.

The remainder of the paper is structured as follows. In Section 2, we provide a review of relevant literature on image segmentation and its applications in remote sensing imagery. Section 3 outlines the details of our HTC-based building extraction method. The experimental setup, including the datasets, evaluation metrics, and selected baseline models, is described in Section 4. In Section 5, we elaborate on the experimental results in detail. Section 6 summarizes the findings, provides discussions, and highlights our future work. Finally, Section 7 concludes the paper.

2. Related Works

This section provides an overview of the current trends in semantic segmentation and instance segmentation using deep learning models, specifically focusing on their

application in extracting buildings from aerial and satellite imagery. It also highlights the main contributions of promising CNN-based approaches in this field.

2.1. Encoder–Decoder Structure-Based Semantic Segmentation

In recent years, models based on encoder–decoder structures, such as FCN [15],U-Net [22] and FPN [23], have been widely used for their capability to perform dense predictions on images without using fully connected layers and generate segmentation maps of any size. The commonality among these models is that the encoder performs feature extraction with multi-stage down-sampling. Then, the decoder gradually recovers the size and structure of the image during up-sampling and generates semantic annotations. The fully convolutional network (FCN), proposed by Long et al., is a landmark pixel-based segmentation method. The decoder structure of FCN is the simplest and contains only one deconvolution operation, while U-Net and FPN adopt multiple up-sampling in the decoder structure. With such a U-shape structure, the U-Net and FPN models can extract features of different scales and combine high-level semantic information with low-level geometric details through the fusion of multiscale feature maps. Compared with U-Net which only make predictions in the last layer of the decoder, the feature pyramid network (FPN) makes independent predictions for feature maps at multiple scales and takes the pixel with the highest confidence as the result. All of these encoder–decoder methods aim to obtain richer information and enhance model performance by combining them in different ways.

Many improved encoder–decoder methods have been employed for building extraction from remote sensing images in recent studies due to its ability to obtain more information and better address the scale-variance problems of building objects in remote sensing images. For instance, Shrestha et al. [24] proposed an improved FCN, which used the exponential linear unit (ELU) in place of the commonly used ReLU activation, and applied conditional random fields (CRFs) at the end of the network to reduce the noise and to sharpen the boundary of the buildings. Yang et al. [25] improved the accuracy of building extraction by paying more attention to the edge and using edge information to refine the segmentation results. Tang et al. [26] proposed Capsule–Encoder–Decoder and introduced a vector named capsule to store the characteristics of buildings and their components to improve generalization capabilities on different remote sensing datasets. In [27], three parallel pre-trained ResNet sub-CNNs followed by a fusion operation and a U-shaped deconvolution network were used to learn building features at different scales. It also used the Sobel filters in both the foreground and background as well as an edge detection loss function called the edge constraint loss (ECL) to obtain more precise masks. Huang et al. [9] proposed a novel approach for building extraction by introducing a gated feature labeling unit. This unit reduces unnecessary feature transmission and refines the coarse classification maps at each decoder stage of the fully convolutional network (FCN). Their method leveraged both HSR images and LiDAR point clouds, leading to significant improvements in building the extraction performance. In [28], a non-local block [29] was introduced at the top of the encoder to obtain global information by capturing the dependencies among pixels. Moreover, in the decoder part, they fused the multi-scale features by concatenating different deconvolution layers. In this way, they also achieved good performance. Yuan et al. [30] proposed a multi-scale adaptive semantic segmentation network (MSST-Net) and achieved good performance with the open access WHU building data set. They used a CNN to decode the features at different stages in Swin Transformer [31] and to concatenate the multi-level decoding outputs. Swin Transformer [31], inspired by the attention-based transformers [32] and attached with shifted windows, appeared promising to solve the scale-variance problem in remote sensing images for the powerful ability to catch spatial and global information. Girard et al. [33] trained a semantic segmentation network that aligned predicted frame fields to ground truth contours, in which they utilized the frame fields along with the raster segmentation to aid polygonization methods in resolving ambiguities caused by discrete probability maps.

2.2. Mask R-CNN Structure-Based Instance Segmentation

The instance segmentation approach combines the advantages of both object detection and semantic segmentation. It goes beyond the differentiation of individual objects and aims at also recovering more accurate locations and detailed shapes of objects. Recently, instance segmentation methods based on the Mask R-CNN structure, combined with other feature extraction techniques, have been widely applied to extract ground objects from remote sensing images [2,4,6]. For example, Zhao et al. [34] utilized the Douglas–Peucker (DP) algorithm [35] and Minimum Description Length (MDL) optimization [36] to reduce the number of points in the initial polygons generated by Mask R-CNN by 86%. This simplification process resulted in significantly simplified building polygons compared to the original irregular polygons generated by Mask R-CNN. The regularized polygons obtained through their approach were found to be suitable for various cartographic and engineering applications. By combining Sobel edge detection algorithm and Mask R-CNN, Zhang et al. [37] integrated the artificial edge features and features extracted by RestNet50 to improve the efficiency and accuracy of the building extraction. Fang et al. [38] proposed an attention-based FPN (AFPN), in which channel attention was introduced into each layer of the classical FPN to improve the identification of small buildings. Then, a two-stage coarse-to-fine contour sub-network was designed to refine building instance contours, which adjusted the deviation between the contours obtained by the Sobel operator and the ground truth through a loss function to further improve the contour accuracy. Based on an improved Mask R-CNN model, which used a modified ResNet101 [39] for feature extraction, Han et al. [5] achieved a high building detection accuracy in a manually annotated building dataset. Zhao et al. [40] introduced global context and boundary refinement blocks (BRB) to upgrade feature extraction, and added channel and spatial attention modules to boost the effectiveness of the detection block. They also employed a recurrent neural network (RNN) structure to sequentially decode the vertices of building polygons after the feature extraction stage of Mask R-CNN. In contrast to the traditional approach of converting pixel-level segmentation results into regularized building boundaries, this method directly obtains vector-formatted regularized building outlines. It provides convenience for engineering mapping and other related applications. Chen et al. [3] implemented Swin Transformer to replace the backbone of a current SOTA algorithm, the multiple attending path neural network (MAP-Net) [41], and obtained a more accurate result in the building extraction.

Instance segmentation methods have diversified the applications of building extraction from remote sensing imagery. Chen et al. [2] proposed a building area calculation method based on the number of building pixels, which enables the timely and accurate assessment of the losses caused by natural disasters such as earthquakes and floods. Amo-Boateng et al. [4] implemented Mask R-CNN to detect the rooftops of buildings in a typical rural settlement to estimate the solar generation potential of such areas.

To sum up, there has been a large body of research on the automatic building extraction from remote sensing images with deep learning models. However, few researchers have taken the detection threshold of the model into account to improve its accuracy. Moreover, the fact that building objects in remote sensing images have more regular contour shapes than other targets has been neglected. Based on this characteristic of buildings, we believe that high-quality bounding boxes can have a significant impact on improving the detection accuracy of building targets. In addition, existing models for building extraction from remote sensing images rarely take global semantic information into consideration. Moreover, the constant threshold commonly used for bounding box evaluation often leads to unsatisfactory results. Models with cascade structure, e.g., Cascade R-CNN [18], were proposed to tackle this issue with resampling by cascade regression, albeit with limited improvement. Hybrid Task Cascade (HTC) [17] proposed by Chen et al. provide a viable alternative for instance segmentation which could extract more comprehensive information and thus enable more precise recognition. It has rarely been used for building extraction tasks with an exception of Liu et al. [42]. However, despite the improvements made in the feature extraction backbone network and RPN network in this study, a detailed analysis

of the specific impacts of each module in the HTC network for building extraction from remote sensing images was not conducted, e.g., the influence of global semantic information. Furthermore, no relevant analysis was performed regarding the influence of detector thresholds on the model. Finally, they also failed to compare the HTC-based method with other SOTA cascade structure instance segmentation methods and more diverse classes of instance segmentation methods, e.g., the anchor-free instance segmentation method. Nonetheless, HTC could be leveraged for automatically extracting buildings from remote sensing imagery.

3. Methodology

In this study, the Hybrid Task Cascade network (HTC) [17] is introduced to enhance the performance of building extraction from high-resolution remote sensing imagery. The architecture of the proposed network is shown in Figure 1.

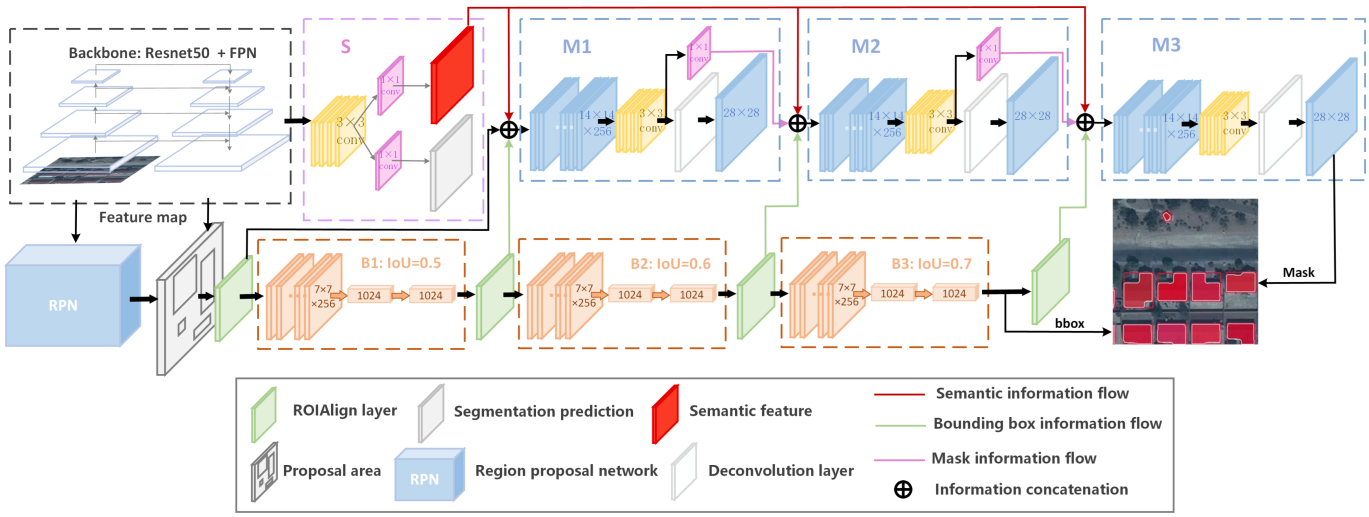

Figure 1. Structure of our HTC building detection model.

Compared to the classical Cascade-Mask-RCNN [43], it exhibits several distinctive characteristics: (1) Instead of performing bounding box regression and mask prediction in parallel, it integrates these tasks. (2) By using a direct path to feed the mask features from the previous stage to the current one, it reinforces the information flow between mask branches. (3) It adopts a fully convolutional branch to explore more global contextual information, fusing it with bbox and mask branches at the same time.

3.1. Network Stucture of HTC

The HTC network primarily consists of a feature extraction network and a three-stage prediction head, mainly focusing on the interaction and fusion of information between the backbone network and the various components. It performs joint processing by interleaving bounding box refinement and masking operations at each stage, as well as full convolutional branching to extract the global semantic context. By incorporating these tasks, the bounding box features, mask features, and global context form a tighter connection, which effectively contributes to the improvement of detection accuracy through the backpropagation.

The following provides a detailed description of the structure of the HTC network. The ResNet50 [39] with the Feature Pyramid Networks (FPN) is first used to extract building rooftop features from input imageries. The extracted feature maps are inputted into the Region Proposal Network (RPN) to extract candidate regions of buildings, and the proposed bounding boxes of the candidate regions are remapped onto the feature maps.

Simultaneously, a semantic segmentation branch is constructed based on the output of the feature pyramid. It performs up-sampling and down-sampling operations on the semantic features extracted at different levels of the feature pyramid, combining high-level

features with global information and low-level features with local information. The transformed feature maps from different levels are then fused through element-wise summation. Subsequently, further feature fusion is performed using four 3 × 3 convolutional layers. Then, two 1 × 1 convolutional layers are utilized to obtain global semantic features of buildings and the semantic segmentation results. The semantic segmentation results are compared with the semantic annotations to compute the loss of the branch, while the global semantic features are inputted into each bounding box and mask heads, allowing the model to be more discriminative on the cluttered background. By incorporating global semantic annotations for supervised learning, the model can pay more attention to the recognition of small and medium-sized building objects, leading to higher detection accuracy. Moreover, this approach aligns well with the focus of our building extraction task, for it optimizes the detection threshold progressively and effectively reduces redundant predictions and the overlapping of building targets.

Secondly, the proposal boxes selected by the region proposal network (RPN) are used to extract Regions of Interest (RoIs), and each RoI is pooled into a 7 × 7 feature map using the RoIAlign layer that was the bounding box feature x_1^{box}. Then, a fully connected branch is employed to regress and classify each RoI's bounding box. In the case of traditional Mask-RCNN, following the RPN and RoIAlign layer, the features within the proposal regions are fed into a segmentation head for mask prediction. However, HTC uses the optimized bounding box b_1, which is obtained from the first-stage bounding box head with an Intersection over Union (IoU) threshold of 0.5 for mask prediction. More accurate bounding boxes can provide more precise object positions and boundary information, thereby offering better initialization and localization for the subsequent fully convolutional mask branch. By using accurate bounding boxes as initial regions, the HTC model can achieve more accurate pixel-level segmentation of building objects. Subsequently, the features extracted by backbone network x, along with b_1 and the semantic segmentation features s, are fused and used as inputs to the first stage's mask prediction head M_1 to extract mask information x_1^{mask}. In the mask prediction head M_1, before using the deconvolution operation to obtain the mask prediction m_1, we also learn the intermediate feature m_1^- through four 3 × 3 and a 1 × 1 convolution layers. The mask information m_1^- from the first stage is further fused to the mask branch of the next stage, which allows for interconnection and mutual influence between adjacent stages' mask branches. All features are supervised through backpropagation, leading to more accurate mask segmentation.

At the same time, the optimized bounding boxes b_1 are fed into the bounding box head B_2, with an IoU threshold of 0.6, to further optimize their positions and sizes and extract the features x_2^{box} of the bounding boxes b_2. The mask feature m_1^- from mask head M_1, the bounding box feature x_2^{box} from bounding box head B_2 and the semantic segmentation feature s are fused again and then inputted into the mask prediction head M_2 to obtain the mask feature m_2^-.

Finally, in the third stage, the above operation is repeated. By using the bounding box head B_3 with an IoU threshold of 0.7, much more accurate bounding boxes b_2 can be obtained, together with the mask feature m_2^- and the semantic segmentation feature s to obtain mask information x_3^{mask}. The final mask prediction results m_3 are obtained in the mask head M_3 by a deconvolution operation. The formula of HTC is expressed as:

$$x_i^{bbox} = P(x, b_{i-1}) + P(S(x), b_{i-1}), \tag{1}$$

$$b_i = B_i(x_i^{box}), \tag{2}$$

$$x_i^{mask} = P(x, b_i) + P(S(x), b_i), \tag{3}$$

$$m_i = M_i(\mathcal{F}_i^{mask}, m_{i-1}^-)), \tag{4}$$

In the given expression, we denote x as the features extracted by the backbone network. The variables x_i^{box} and x_i^{mask} represent the box and mask features obtained from x and the input RoIs. The pooling operator $P(\cdot)$, such as RoI Align or ROI pooling, is applied. B_i and M_i refer to the box and mask heads at the i-th stage, while b_i and m_i denote the corresponding box predictions and mask predictions. m_i^- represents the intermediate feature of the i-th mask head. The semantic segmentation head is represented by S, and $\mathcal{F}$ denotes a function that combines the features at the current stage with the preceding one.

3.2. Loss Function

Firstly, the overall loss function *Loss* is formulated in the form of multi-task learning, which consists of the bounding box regression loss L_{bbox}, mask loss L_{mask} and semantic segmentation L_{seg}. The overall loss function is expressed as follows:

$$Loss = \sum_{i=1}^{3} \alpha_i (L_{bbox}^i + L_{mask}^i) + \beta L_{seg}, \tag{5}$$

Here, we set α_i = [1, 0.5, 0.25] in three stages of the HTC, respectively, and β = 1 by default, which follows the same definition as in HTC [17]. This multi-task learning framework enables the model to jointly optimize these three components, leading to improved performance in various aspects of the task.

Among the loss components, L_{bbox} combines two terms, L_{cls} and L_{reg}, which are used for classification and bounding box regression, respectively. L_{mask} is formulated using the binary cross-entropy form as described in Mask R-CNN [16]. Additionally, in our task, the semantic segmentation loss L_{seg} also appears in the form of binary cross-entropy, serving the purpose of distinguishing the foreground and background. They are given as follows:

$$L_{bbox}^i(c_i, b_i, \hat{c}_i, \hat{b}_i) = L_{cls}(c_i, \hat{c}_i) + L_{reg}(b_i, \hat{b}_i), \tag{6}$$

$$L_{mask}^i(m_i, \hat{m}_i) = BCE(m_i, \hat{m}_i), \tag{7}$$

$$L_{seg}(s, \hat{s}) = BCE(s, \hat{s}), \tag{8}$$

Here, the bbox regression loss L_{reg} is a smoothL1 loss and the classification loss also takes the form of a BCE loss. The bounding box regression loss is given as follows:

$$L_{reg} = \sum SmoothL1(y_i, \hat{y}_i), \tag{9}$$

where $\hat{y}_i$ is the predicted box and y_i is the box label. Finally, the smoothL1 loss and BCE loss is written as follows:

$$L_{SmoothL1}(y, \hat{y}) = \begin{cases} 0.5(\hat{y} - y)^2, & \text{if } |\hat{y} - y| < 1, \\ |\hat{y} - y| - 0.5, & \text{otherwise.} \end{cases} \tag{10}$$

$$L_{BCE}(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)], \tag{11}$$

where y_i represents the true label, $\hat{y}_i$ represents the predicted label probability and N is the number of samples.

4. Experiments

In this section, we present our experiment setups, datasets used, evaluation metrics and experimental results.

4.1. Implementation Details

To ensure the consistency of the experimental results, FPN was used in all backbones. The experiments were conducted using the PyTorch framework on a 3060 GPU. A weight decay rate of 0.0001 and a momentum value of 0.9 were applied. The detectors were trained

using a batch size of 16 images per GPU for 12 epochs. The initial learning rate was set to 0.02, and it was decreased by a factor of 0.1 after the 8th and 11th epochs, respectively.

4.2. Datasets

The proposed method was evaluated on three challenging building instance segmentation datasets: CrowdAI Mapping Challenge Dataset [19], the WHU aerial image dataset [20] and the self-annotated dataset proposed in [21]. Traditional instance annotations are used to supervise the bbox and mask branches, while the semantic branch is supervised by COCO-stuff [44] annotations. By employing the Python COCOAPI, we convert traditional COCO annotations into the COCO-stuff format and perform grayscale transformation to obtain foreground–background semantic annotations.

4.2.1. CrowdAI Mapping Challenge Dataset

The CrowdAI Mapping Challenge dataset is a large-scale remote sensing imagery building dataset published by [19] for the mapping challenge. It is comprised of a total of 280,741 satellite images for training, along with an additional 60,317 images for testing. In academic research, the experiment is commonly conducted with small-scale CrowdAI datasets, which usually comprise 8366 images for training and 1820 images for testing. Each image is presented in JPEG format with a resolution of 300 × 300 pixels. Furthermore, their annotations are provided in the MS-COCO format [45], offering useful information about the objects and their characteristics within the images. The samples displayed in Figure 2 showcase a diverse range of buildings in the CrowdAI dataset, in varying sizes and shapes.

Figure 2. Samples of CrowdAI Mapping Challenge Dataset and corresponding building labels.

4.2.2. WHU Aerial Image Dataset

The WHU aerial image dataset, proposed by Ji et al. [20], consists of JPEG images with a resolution of 512 × 512 pixels. The dataset contains over 187,000 buildings with varying sizes and appearances, captured in Christchurch, New Zealand. It is divided into a training set of 4736 images and a test set of 2416 images. Some samples are illustrated in Figure 3.

Figure 3. Samples of WHU Aerial Image Dataset and building labels.

4.2.3. Chinese Typical City Buildings Dataset

The dataset utilized in this study was provided by Fang et al. [21]. It comprises a total of 7260 tiles, with 5985 images allocated for training purposes and 1275 for testing. The dataset encompasses a diverse range of urban areas in four major Chinese cities: Beijing, Shanghai, Shenzhen and Wuhan. In total, it contains 63,886 instances of buildings. Each image in the dataset is presented in TIF format and possesses dimensions of 500 × 500 pixels, with a spatial resolution of 0.29 m per pixel. The dataset exhibits a wide coverage, encompassing diverse architectural styles with significant variations in terms of shape, size, color and other distinctive features of buildings. Some samples from this building instance segmentation dataset are shown in Figure 4.

Figure 4. Samples of Self-Annotated Dataset and building labels.

4.3. Evaluation Metrics

In this study, the proposed method is evaluated using the standard MS COCO metrics [45], which include mean average precision (AP) and mean average recall (AR) at multiple IoU values, e.g., AP, AP_{50}, AP_{75}. Additionally, the mean average precision is computed for objects of different scales, denoted as AP_s, AP_m and AP_l. The IoU is calculated using the equation below:

$$IoU = \frac{intersection}{union} = \frac{S_{pred} \cap S_{gt}}{S_{pred} \cup S_{gt}}, \tag{12}$$

where S_{pred} and S_{gt} represent the prediction results and the ground truth, respectively, and the AP and AR are calculated at 10 IoU overlap thresholds ranging from 0.50 to 0.95 in increments of 0.05, following the evaluation metrics [45]. The formulas are given as follows:

$$AP = \frac{AP_{0.50} + AP_{0.55} + \ldots + AP_{0.90} + AP_{0.95}}{10}, \tag{13}$$

$$AR = 2\int_{0.5}^{1} recall(o)do = \frac{2}{n}\sum_{i=1}^{n} max(IoU(gt_i) - 0.5, 0), \tag{14}$$

In addition, AP_{50} and AP_{75} are AP calculated at IoU thresholds of 0.5 and 0.75, respectively. Between them, the AP_{75} is a much more strict evaluation metric, which can better reflect the position accuracy of the algorithm. Furthermore, the metrics AP_s, AP_m, and AP_l are used to evaluate the performance of the methods on detecting buildings of different sizes. Specifically, AP_s measures the performance in detecting small buildings with an area $S < 32 \times 32$ pixels, AP_m measures the performance in detecting buildings with an area $32 \times 32 < S < 96 \times 96$ pixels, and AP_l measures the performance in detecting large buildings with an area $S > 96 \times 96$ pixels.

4.4. Baseline Methods

In this study, the HTC builing instance segmentation method is compared with five baselines, Mask R-CNN [16], Cascade Mask R-CNN [43], SCNet [46], SOLOv2 [47] and Swin Mask R-CNN [48]. Mask R-CNN is a classic instance segmentation method; recent works showed its popularity for building the instance extraction. Cascade Mask R-CNN effectively integrates cascade structures into instance segmentation, which can obtain better detection accuracy than traditional Mask R-CNN. SCNet, proposed by Vu et al. [46], utilizes stacking and skip connect operations to connect mask branches of the Cascade Mask R-CNN into a sequence of successive convolutional layers, ensuring sample consistency and improving the model training speed and accuracy. It also introduces a feature relay operation to establish a feature prior for bounding box feature to mask prediction. In addition, they introduced a global context branch similar to that in the HTC model for multi-label prediction and fused a global feature with a box and mask branch. SOLOv2 is a powerful, novel and efficient anchor-free instance segmentation model that dynamically segments each instance in an image by decomposing the mask branches into mask kernel prediction and mask feature learning. Swin Mask R-CNN, which introduces Swin Transformer into traditional Mask R-CNN, employs a patch-based approach and the "shift window" conception to effectively capture pixel relationships within each patch while promoting an information exchange between patches. Consequently, this architecture significantly enhances the performance of object detection and instance segmentation tasks and has been frequently used in SOTA methods. All of the baseline models adopt the same settings and use ResNet50 and FPN as the backbone, except for Swin Mask R-CNN, which uses Swin Transformer as the backbone with the window size set to 7 × 7 and transformer blocks with sizes of 2, 2, 6 and 2 in the 4 stages, respectively.

5. Experimental Results

This section presents a comparative analysis of the results obtained from five baseline models and our proposed HTC building detection method, followed by a comparison of their actual performance on the three test sets.

5.1. Results on the CrowdAI Mapping Challenge Dataset

Table 1 presents a comparison of results obtained using different instance segmentation algorithms on the CrowdAI Mapping Challenge Dataset. In general, the HTC-based method demonstrates superior performance across various evaluation metrics, including bounding box and mask accuracy. Our approach exhibits significant improvements in AP and AR for both the bounding box branch and mask branch, surpassing the baseline methods. Specifically, AP is increased by 0.6–3% and 0.3–2.2% in the bounding box and mask branches, respectively, while AR is enhanced by 1.0–3.8% and 0.8–2.9%, respectively. Moreover, for small-sized building instances, the bounding box branch and mask branch

achieve an improvement of 0.3–2.6% and 0.4–2.9% in AP_s. For medium-sized buildings, AP_m is enhanced by 0.3–3.1% and 0.1–1.2% in the two branches. Additionally, for large-scale building instances, AP_l shows improvements of 1.7–4.2% and 0.3–4.0% in the bounding box branch and mask branch, respectively.

Table 1. Experimental results on the CrowdAI Mapping Challenge Dataset.

Methods	Bounding Box (%)							Mask (%)						
	AP	***AP*$_{50}$**	***AP*$_{75}$**	***AP*$_s$**	***AP*$_m$**	***AP*$_l$**	***AR***	***AP***	***AP*$_{50}$**	***AP*$_{75}$**	***AP*$_s$**	***AP*$_m$**	***AP*$_l$**	***AR***
Mask R-CNN	55.7	82.4	64.6	27.1	71.4	69.8	61.1	54.3	82.6	64.4	26.4	69.2	68.5	59.6
Swin Mask R-CNN	56.0	82.5	65.5	28.4	71.4	72.0	61.0	54.9	83.4	64.9	27.7	69.4	71.6	59.8
SOLOv2	—	—	—	—	—	—	—	53.7	82.4	63.0	25.3	69.4	72.0	60.2
Cascade Mask R-CNN	58.1	83.5	66.1	29.4	73.9	71.0	63.5	55.2	83.5	65.0	27.8	69.7	70.1	60.5
SCNet	58.1	83.8	66.1	29.2	74.2	72.3	63.9	55.6	83.7	65.3	27.5	70.3	72.2	61.7
Proposed Method	58.7	84.1	66.6	29.7	74.5	74.0	64.9	55.9	84.2	66.1	28.2	70.4	72.5	62.5

To further illustrate and analyze the results, selected samples are shown in Figure 5, with a focus on the areas outlined by blue circles. Firstly, as illustrated in the first row, Swin Mask R-CNN exhibits an imperfect segmentation of certain building contours. Secondly, in the second row, baseline methods overlook some small-scale buildings and edge targets, whereas our method successfully detects them. Furthermore, in the third rows, baseline methods erroneously detect non-building objects, e.g., vehicles and shadows. Moreover, in the fourth row, compared to the baseline method, our approach extracts building masks with more regular contours and closer approximation to real building outlines. Finally, from the fifth line, we can observe that our method performs well in identifying buildings with interior holes, whereas the baseline methods either fails to recognize the interior holes or produces blurred boundaries. These findings indicate that the proposed method outperforms the baseline methods.

Figure 5. Example of results on the CrowdAI Mapping Challenge Dataset: (**a**) Original image, (**b**) Label, (**c**) Mask R-CNN, (**d**) Swin Mask R-CNN, (**e**) SOLOv2, (**f**) Cascade Mask R-CNN, (**g**) SCNet and (**h**) HTC.

5.2. Results on the WHU Aerial Image Dataset

The experimental results on WHU Aerial Image Dataset are listed in Table 2. Our method achieves a state-of-the-art performance in all evaluation metrics related to bounding

box accuracy. However, it slightly falls behind Swin Mask R-CNN in certain evaluation metrics for mask detection accuracy. The proposed method obtains more precise bounding box results than the five baselines, with AP, AP_{50}, AP_{75} and AR values of 66.5%, 85.5%, 75.8% and 70.7%, respectively. Specifically, AP is increased by 0.6–2.8%, while AR is enhanced by 0.8–2.7% compared to the baseline models.

Table 2. Experimental results on the WHU Aerial Image Dataset.

Methods	Bounding Box (%)							Mask (%)						
	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	AR	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	AR
Mask R-CNN	63.7	83.9	73.3	49.3	79.8	73.6	68.0	60.4	84.8	71.4	44.5	77.3	75.2	64.2
Swin Mask R-CNN	63.9	84.8	74.2	50.0	79.5	74.2	68.2	62.7	84.9	73.5	47.2	79.1	78.1	66.5
SOLOv2	—	—	—	—	—	—	—	55.1	83.3	64.3	37.2	74.3	76.8	59.4
Cascade Mask R-CNN	65.9	84.7	74.7	51.2	82.3	76.6	69.9	60.5	84.8	71.5	44.5	77.5	76.7	64.1
SCNet	64.9	85.2	74.3	50.2	81.5	74.7	69.2	60.3	84.8	71.1	44.2	77.4	76.8	64.1
Proposed Method	66.5	85.5	75.8	52.1	82.7	76.5	70.7	61.7	85.5	73.1	46.1	78.3	77.9	65.4

Although our method slightly lags behind Swin Mask R-CNN in several mask accuracy metrics, we can derive valuable insights from the specific experimental results shown in Figure 6 (with a focus on the areas outlined by blue circles). Firstly, compared to the baseline methods, our approach achieves more regular and smoother contour shapes. Secondly, in the third and fourth lines, we can observe that our method demonstrates higher detection accuracy for very small building instances. Furthermore, the precise identification of small interior holes in buildings in the fourth line also demonstrates the superior accuracy of our method in recognizing building outlines. Lastly, in the last row, for different regions of large-scale building instances, the baseline methods exhibit varying degrees of redundant detections. In contrast, our detection method, by gradually increasing the IoU threshold, produces better mask predictions for such buildings, significantly reducing instances of overlapping predicted masks. This characteristic makes our method more suitable for remote sensing image building detection tasks, especially for scenarios with large and unconventionally shaped buildings.

Figure 6. Example of results on the WHU Aerial Image Dataset: (**a**) Original image, (**b**) Label, (**c**) Mask R-CNN, (**d**) Swin Mask R-CNN, (**e**) SOLOv2, (**f**) Cascade Mask R-CNN, (**g**) SCNet and (**h**) HTC.

5.3. Results on the Chinese Typical City Buildings Dataset

The results obtained using different instance segmentation algorithms in this dataset are presented in Table 3. This dataset consists of non-orthophoto images with complex backgrounds, posing challenges in extracting building instances. Moreover, these challenges might have contributed to the relatively lower detection performance of Swin Mask R-CNN. However, the HTC method performs the best in all other evaluation metrics, except for a slight inferiority to SCNet in the metric of mask precision AP_l. Firstly, the bounding box branch exhibits an improvement of 0.8–3.9% in AP, while the mask branch shows an enhancement of 0.3–4.0%. Secondly, the bounding box branch and mask branch demonstrate an increase of 1.9–7.5% and 1.0–6.2% in AR, respectively. Additionally, for small-sized building instances, the bounding box branch and mask branch achieve an improvement of 0.5–2.5% and 0.2–5.1% in AP_s, respectively. Here, we find that SOLOv2 is quite ineffective at detecting small target building objects. For medium-sized buildings, the AP_m in both branches is enhanced by 0.7–3.8% and 0.5–3.6%. Furthermore, when considering large-scale building instances, the bounding box branch demonstrates notable enhancements in AP_l, with improvements ranging from 0.6% to 4.7%. The SCNET achieves nearly the same accuracy in large target building detection. These data highlight the superiority and robustness of the HTC method for building extraction from non-orthophoto remote sensing imagery.

Table 3. Experimental results on the Chinese Typical City Buildings Dataset.

Methods	Bounding Box (%)							Mask (%)						
	AP	***AP*$_{50}$**	***AP*$_{75}$**	***AP*$_s$**	***AP*$_m$**	***AP*$_l$**	***AR***	***AP***	***AP*$_{50}$**	***AP*$_{75}$**	***AP*$_s$**	***AP*$_m$**	***AP*$_l$**	***AR***
Mask R-CNN	47.8	72.1	52.7	23.2	54.7	58.9	61.7	45.3	71.5	49.7	17.7	51.0	58.3	58.8
Swin Mask R-CNN	46.9	72.1	52.4	22.4	53.9	57.3	59.4	45.1	71.4	49.8	17.3	51.1	57.5	57.3
SOLOv2	—	—	—	—	—	—	—	43.0	69.9	46.9	13.6	49.4	58.7	56.3
Cascade Mask R-CNN	50.0	71.7	55.4	24.0	57.0	61.4	64.3	46.1	71.6	50.9	18.1	51.9	59.2	59.4
SCNet	49.9	72.9	54.6	24.4	56.8	60.7	65.0	46.7	72.7	51.5	18.5	52.5	60.7	61.5
Proposed Method	50.8	73.2	56.0	24.9	57.7	62.0	66.9	47.0	72.8	52.1	18.7	53.0	60.6	62.5

Examples of detailed experimental results, as shown in Figure 7 (with a focus on the areas outlined by blue circles), demonstrate where the HTC method outperforms the baseline methods in building extraction. The HTC method achieves more accurate predictions with fewer errors and clearer building contour extraction. It also reduces the occurrence of repetitive outputs and overlapping objects. From the first, fourth, and fifth rows, we can observe that our method can more accurately identify small target buildings and buildings located at the edges compared to the baseline methods. Meanwhile, from the second and fifth rows, we can observe that our method extracts building masks with more regular and smoother contours. Finally, from the third row, we find that our method can effectively filter out interferences, such as trees, shadows of buildings, and other disturbances, and thus accurately identify building targets and their contours.

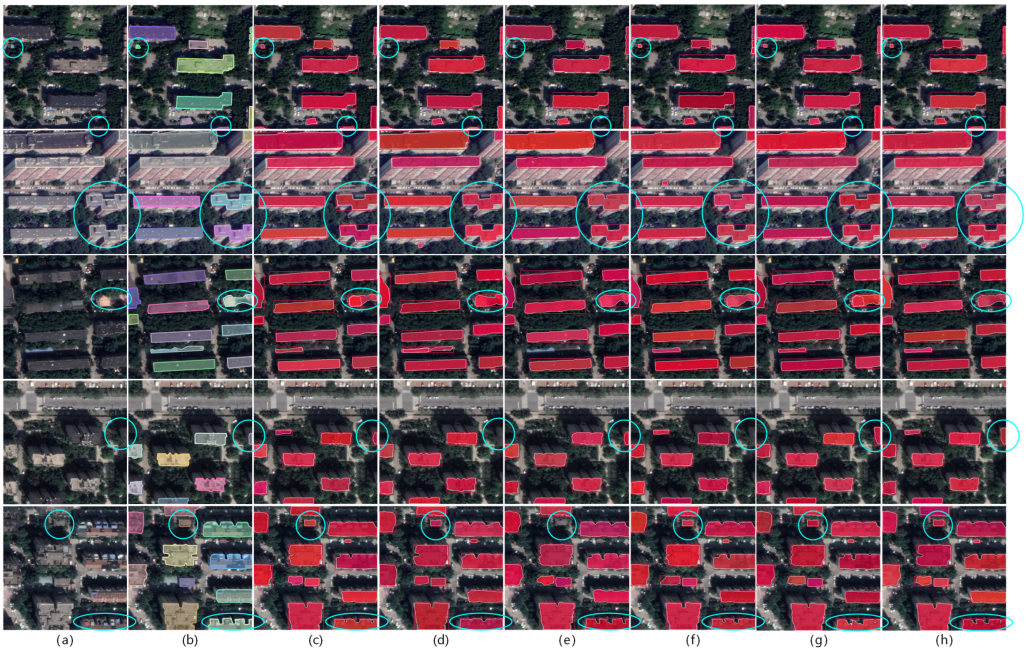

Figure 7. Example of results on Chinese Typical City Buildings Dataset: (**a**) Original image, (**b**) Label, (**c**) Mask R-CNN, (**d**) Swin Mask R-CNN, (**e**) SOLOv2, (**f**) Cascade Mask R-CNN, (**g**) SCNet and (**h**) HTC.

5.4. Ablation Experiments

In order to understand the specific effects of each component of HTC for remote sensing building extraction, we conducted experiments on the CrowdAI Mapping Challenge Dataset and discovered some interesting experimental results.

As shown in Tables 4 and 5, the integrated structure, mask information flow and semantic segmentation branch module all contribute to the improvements in model accuracy compared to the vanilla Cascade Mask R-CNN. However, when each of the three modules is individually excluded, the impact on the final model's detection accuracy is not evident, except for the detection of large-scale building objects. The absence of any single module leads to a decrease in the model's detection accuracy for large-scale building objects AP_l by 1.3–2.3% in the bounding box branch and 0.5–1.5% in the mask branch.

Additionally, it is noteworthy that when only the semantic segmentation branch is used without the integrated structure and mask information flow, the model achieves optimal detection performance on the CrowdAI validation dataset. Compared to HTC, the detection accuracy for small-scale building objects AP_s is improved by 4.9% and 4.7% in the bounding box branch and mask branch, respectively. The detection accuracy for medium-sized buildings AP_m is also improved by 1.1% and 0.9% in the bounding box branch and mask branch, respectively. Although there is a slight decrease of 0.9% and 1.1% in the detection accuracy of large-scale building objects AP_l, overall improvements in AP and AR are observed with only the semantic segmentation module added to Cascade Mask R-CNN, largely due to the more frequent presence of small and medium-sized building objects in the images. While this result may be influenced by the characteristics of the dataset itself and may lack generalizability, it still highlights the importance of global semantic information for the detection of small-scale building objects in remote sensing imagery. Small-size buildings have limited available features, and their semantic information appears in low-level feature maps. As the network deepens, their detailed information may be completely lost. However, HTC addresses this issue by incorporating an additional branch for contextual information utilization in semantic segmentation and utilizing global semantic annotation for supervised learning, thus better preserving the information of small target buildings.

Table 4. Ablation experiments for the bounding box branch on the CrowdAI Mapping Challenge Dataset.

Cascade	Integrated	Mask Info	Semantic	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	AR
✓				58.1	83.5	66.1	29.4	73.9	71.0	63.5
✓	✓	✓		58.6	84.0	66.9	29.8	74.3	72.0	64.9
✓		✓	✓	58.7	84.1	66.6	29.8	74.4	71.7	64.9
✓	✓		✓	58.9	84.1	66.7	30.0	74.7	72.7	65.0
✓			✓	61.4	87.4	69.3	34.6	75.6	73.1	67.9
✓	✓	✓	✓	58.7	84.1	66.6	29.7	74.5	74.0	64.9

Table 5. Ablation experiments for the Mask branch on the CrowdAI Mapping Challenge Dataset.

Cascade	Integrated	Mask Info	Semantic	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	AR
✓				55.2	83.5	65.0	27.8	69.7	70.1	60.5
✓	✓	✓		55.9	84.1	65.9	28.3	70.4	71.1	62.5
✓		✓	✓	55.8	84.2	65.7	28.3	70.3	71.0	62.5
✓	✓		✓	55.8	84.1	65.7	28.3	70.3	72.0	62.3
✓			✓	58.3	87.5	68.4	32.9	71.3	71.4	65.2
✓	✓	✓	✓	55.9	84.2	66.1	28.2	70.4	72.5	62.5

We conducted similar experiments on another two datasets, and the results were largely consistent, except for a slight decrease in model accuracy compared to HTC when using only the semantic segmentation branch module.

5.5. Effects of different IoU threshold

The IoU threshold has an impact on the selection of positive and negative samples during the model training process, thereby affecting the training effectiveness of the model. A suitable IoU threshold improves training sample quality and model performance. While 0.5 is commonly used for optimal results in general visual tasks, for building extraction in remote sensing imagery, buildings often have regular and well-defined shapes, leading to higher IoU values for bounding boxes. To validate this hypothesis, we conducted experiments on different datasets. Table 6 shows the results of experiments on the CrowdAI dataset using Mask R-CNN and HTC models, respectively. Increasing the IoU threshold by approximately 0.1 yields slight improvements in key performance metrics, and results in higher AP and AR.

Table 6. Experimental results on the CrowdAI Dataset with different IoU thresholds.

Methods	IoU	Bounding Box (%)							Mask (%)						
		AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	AR	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	AR
Mask R-CNN	0.5	55.7	82.4	64.6	27.1	71.4	69.8	61.1	54.3	82.6	64.4	26.4	69.2	68.5	59.6
Mask R-CNN	0.6	56.1	82.1	64.8	27.7	72.0	68.3	61.5	54.4	82.9	64.3	26.7	69.3	68.9	59.9
HTC	(0.5, 0.6, 0.7)	58.7	84.1	66.6	29.7	74.5	74.0	64.9	55.9	84.2	66.1	28.2	70.4	72.5	62.5
HTC	(0.6, 0.7, 0.8)	59.1	83.9	66.9	29.8	75.1	72.7	65.1	56.2	84.1	66.1	28.3	70.9	72.5	62.6

Table 7 shows the results of experiments on the WHU dataset using Mask R-CNN and HTC models, respectively. By appropriately increasing the IoU threshold, we obtained improved results. Specifically, for the HTC model, increasing the detector threshold by 0.1 led to improved precision in terms of AP, AR and detection accuracy for buildings of different scales. Additionally, we conducted an set of experiments with an increased IoU threshold of 0.05 to further validate our hypothesis.

Table 7. Experimental results on the WHU Dataset with different IoU thresholds.

Methods	IoU	Bounding Box (%)							Mask (%)						
		AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	*AR*	*AP*	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	*AR*
Mask R-CNN	0.5	63.7	83.9	73.3	49.3	79.8	73.6	68.0	60.4	84.8	71.4	44.5	77.3	75.2	64.2
Mask R-CNN	0.6	64.6	83.8	74.6	50.0	81.4	72.9	68.6	60.9	83.9	72.4	44.7	78.1	75.1	64.3
HTC	(0.5, 0.6, 0.7)	66.5	85.5	75.8	52.1	82.7	76.5	70.7	61.7	85.5	73.1	46.1	78.3	77.9	65.4
HTC	(0.55, 0.65, 0.75)	66.9	85.3	75.6	52.3	83.4	77.0	71.1	62.0	85.5	73.1	46.0	78.9	78.9	65.7
HTC	(0.6, 0.7, 0.8)	67.2	85.1	76.2	52.6	83.9	77.0	71.2	62.4	85.2	73.6	46.5	79.2	78.8	66.0

The findings indicate that appropriately increasing the model's IoU threshold significantly enhances training effectiveness and improves the detection accuracy of the model for automatic building extraction using remote sensing imagery.

5.6. Efficiency and Generalizability Tests

While our method outperforms baseline models in terms of model accuracy, the HTC model is also comparatively more complex with more parameters, which could affect the computational costs. To better evaluate our model, we compared the parameter sizes as well as the detection speed of the proposed HTC-based model with other baseline models on the WHU dataset. As shown in Table 8, while obtaining higher accuracy, our model also sacrifices efficiency to a certain extent. This undoubtedly limits applications of the HTC-based model to scenarios that require fast processing of remote sensing images such as disaster monitoring. To further analyze the model efficiency of HTC-based models, we carried out additional experiments with a modified model structure. We observe that removing the semantic segmentation branch substantially improves the efficiency of HTC by 38.9%. Although from the previous experimental results in Tables 4 and 5 we found that missing global semantic information affects the detection accuracy for large target buildings, the effect on the overall accuracy is modest. Therefore, for application scenarios that require more efficient detection, a modified HTC model without a semantic segmentation branch could be utilized.

Table 8. Parameter sizes and detection speed on the WHU Dataset.

Methods	Parameters (*M*)	fps (*Frames per Second*)
Mask R-CNN	335	20.64
Cascade Mask R-CNN	587	14.12
Swin Mask R-CNN	542	18.58
SOLOv2	352	15.19
SCNet	720	8.42
HTC	609	8.92
HTC (without semantic branch)	588	12.39

To further validate our model backbone selection, we used four different backbones other than ResNet50 [39] for the experimental comparison on the WHU dataset. The experimental results are shown in Table 9. We found that although ResNet18 [39], with a lower number of network layers, reduces the size and the training time of the model, it sacrifices more in terms of detection accuracy. In addition, deeper networks, ResNet101 [39] and ResNext101 [49], produce slower training without any gain in accuracy. Finally, when we use HRNetV2p [50] as the backbone, which introduce more cross-connections among branches, most of the accuracy metrics are improved, especially for small and medium-sized building targets, although it lags behind in terms of efficiency. It suggests that the method is worth being considered in scenarios where higher detection accuracy is required.

Table 9. Experimental results on HTC with different backbone.

Backbone	Mask (%)							Parameters (*M*)	fps (*Frames per Second*)
	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	*AR*		
Resnet-18	59.9	84.5	70.6	43.5	77.0	76.4	64.0	509	9.47
Resnet-50	61.7	85.5	73.1	46.1	78.3	77.9	65.4	609	8.92
Resnet-101	60.3	84.7	71.2	44.2	76.8	77.9	64.2	754	8.16
Resnext-101	61.3	85.0	72.4	45.5	78.2	77.7	64.8	751	8.10
HRNetV2p	62.3	85.9	73.6	46.6	79.3	77.7	65.9	980	7.60

6. Discussion

In this section, we discuss the findings based on our experiment results and discuss our future work.

6.1. Summary of Results

Firstly, our experiments demonstrate the promising capabilities of the HTC model based on the cascaded structure for building extraction from remote sensing imagery. Compared to the selected baseline algorithms, the HTC method achieves higher detection accuracy and generates more regular building contours. In addition, the HTC method integrates the execution of tasks such as the bounding box refinement, masking, and global semantic information extraction. This integration strengthens the information interaction and connection between different branches, ultimately leading to accuracy improvement through backpropagation. This approach has been shown to be effective in enhancing the performance compared to traditional non-hybrid models that handle these tasks separately. The method's accuracy and robustness are validated on three challenging datasets.

Secondly, we show that through cascaded optimization algorithms, our method not only achieves higher recognition accuracy but also reduces the overlap of predicted targets in practical applications. This aligns well with the inherent characteristics of building targets in remote sensing imagery, making the obtained images more suitable for real-world engineering applications. Therefore, the HTC-based method is promising as a suitable alternative for applications in building extraction from remote sensing imagery tasks.

Thirdly, we conduct experiments to evaluate the effectiveness of different modules in the HTC model, revealing their contributions to improving model accuracy and their impacts on recognizing buildings of different scales. Additionally, we find that the utilization of global semantic information plays a vital role in accurately identifying small-sized building targets.

Furthermore, we propose a hypothesis regarding optimal IoU thresholds based on the relatively regular shape of building targets in remote sensing imagery. Our experiments confirm that appropriately increasing the detector's IoU threshold positively affects model accuracy. We believe that this can provide valuable insights for the future application of instance segmentation methods for automatic building detection.

Finally, some of the current limitations of the HTC method are addressed. We have conducted generalization experiments that enable the HTC-based building instance segmentation method to be effectively applied to more scenarios, such as drone disaster assessment (DDS) [51]. In summary, we believe that our proposed method is feasible and has sufficient application prospects.

6.2. Future Work

Our future work will primarily focus on several key directions. Firstly, we aim to investigate the effectiveness of employing additional stages to further enhance the model's accuracy. In a recent study by Wu et al. [52], they proposed an enhanced cascade structure network called Cascade R-CNN++ for high-quality object detection in multi-resolution remote sensing imagery. They found that adding a fourth stage to Cascade R-CNN led to a performance degradation. They attributed this to a potential mismatch between the RoI

features and classifier. Additionally, we plan to investigate alternative backbone networks to assess their potential impact on improving HTC's performance.

Secondly, we intend to incorporate building regularization methods to obtain more regular and realistic building contours. Girard et al. [33] proposed a frame-field-based approach for building contour regularization. They trained a deep neural network that aligned a predicted frame field to ground truth contours and utilized the frame field to facilitate polygonization.

Moreover, we aim to explore the applicability of global semantic information in other instance segmentation models. We also extend our efforts to explore how to reduce parameters and improve the speed of model training while preserving globle semantic information.

Finally, we aim to apply our model to real-world scenarios. Experimental results on real-world datasets demonstrate great potential for practical applications, including mapping, urban planning, change detection, and the integration of multi-source geospatial data for urban functional studies [53].

7. Conclusions

This study introduces the HTC model for building extraction from remote sensing imagery and evaluates its effectiveness and robustness on three challenging building extraction datasets. Experimental results demonstrate that the proposed model outperforms existing techniques with higher accuracy and more precise contours. Moreover, the research highlights the positive impact of incorporating global semantic information on efficient and accurate extraction of building targets in remote sensing imagery. Additionally, by considering the inherent characteristics of building targets in remote sensing imagery, the HTC model produces building maps that are more readily applicable.

Finally, through an experimental analysis, we conclude that appropriately increasing the detector's IoU threshold can improve the detection accuracy of building extraction in remote sensing imagery to a certain extent. In summary, our proposed method holds great potential for applications in map-making, urban planning, intelligent urban transportation, geological exploration, urban life quality assessment and beyond. However, alongside these advantages, it is important to address the limitations of large parameter counts and the relatively slow training speed in HTC models, which warrants further investigation in future studies. Our future work will focus on continuing to improve detection accuracy, while reducing time costs and obtaining regular and smooth contours.

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

Funding: This research was funded by the National Natural Science Foundation of China (42201495), the Guangdong Basic and Applied Basic Research Foundation (2021A1515110049), and Shenzhen Science and Technology Program (JCYJ20220818100200001).

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

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

References

1. 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; pp. 3226–3229.
2. Chen, J.; Wang, G.; Luo, L.; Gong, W.; Cheng, Z. Building Area Estimation in Drone Aerial Images Based on Mask R-CNN. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 891–894. [CrossRef]
3. Chen, D.; Tu, W.; Cao, R.; Zhang, Y.; He, B.; Wang, C.; Shi, T.; Li, Q. A hierarchical approach for fine-grained urban villages recognition fusing remote and social sensing data. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *106*, 102661. [CrossRef]

4. Amo-Boateng, M.; Sey, N.E.N.; Amproche, A.A.; Domfeh, M.K. Instance segmentation scheme for roofs in rural areas based on Mask R-CNN. *Egypt. J. Remote Sens. Space Sci.* **2022**, *25*, 569–577. [CrossRef]
5. Han, Q.; Yin, Q.; Zheng, X.; Chen, Z. Remote sensing image building detection method based on Mask R-CNN. *Complex Intell. Syst.* **2021**, *8*, 1847–1855. [CrossRef]
6. Wang, Y.; Li, S.; Teng, F.; Lin, Y.; Wang, M.; Cai, H. Improved mask R-CNN for rural building roof type recognition from uav high-resolution images: A case study in hunan province, China. *Remote Sens.* **2022**, *14*, 265. [CrossRef]
7. Powers, R.P.; Hay, G.J.; Chen, G. How wetland type and area differ through scale: A GEOBIA case study in Alberta's Boreal Plains. *Remote Sens. Environ.* **2012**, *117*, 135–145. [CrossRef]
8. Hu, L.; Zheng, J.; Gao, F. A building extraction method using shadow in high resolution multispectral images. In Proceedings of the 2011 IEEE International Geoscience and Remote Sensing Symposium, Vancouver, BC, Canada, 24–29 July 2011; pp. 1862–1865. [CrossRef]
9. Huang, J.; Zhang, X.; Xin, Q.; Sun, Y.; Zhang, P. Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network. *ISPRS J. Photogramm. Remote Sens.* **2019**, *151*, 91–105. [CrossRef]
10. Yuan, J.; Cheriyadat, A.M. Learning to count buildings in diverse aerial scenes. In Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Fort Worth, TX, USA, 4–7 November 2014; Association for Computing Machinery: New York, NY, USA, 2014; pp. 271–280. [CrossRef]
11. Blaschke, T.; Hay, G.J.; Kelly, M.; Lang, S.; Hofmann, P.; Addink, E.; Feitosa, R.Q.; van der Meer, F.; van der Werff, H.; van Coillie, F.; et al. Geographic Object-Based Image Analysis—Towards a new paradigm. *ISPRS J. Photogramm. Remote Sens.* **2014**, *87*, 180–191. [CrossRef]
12. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
13. Noble, W.S. What is a support vector machine? *Nat. Biotechnol.* **2006**, *24*, 1565–1567. [CrossRef]
14. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Volume 28.
15. 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.
16. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
17. Chen, K.; Pang, J.; Wang, J.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Shi, J.; Ouyang, W.; et al. Hybrid task cascade for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 4974–4983.
18. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.
19. Mohanty, S.P. Crowdai Mapping Challenge 2018: Baseline with Mask RCNN. GitHub Repository. 2018. Available online: https://github.com/crowdai/crowdai-mapping-challenge-mask-rcnn (accessed on 6 October 2023).
20. Ji, S.; Wei, S.; Lu, M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 574–586. [CrossRef]
21. Fang, F.; Wu, K.; Zheng, D. A dataset of building instances of typical cities in China [DB/OL]. *Sci. Data Bank* **2021**. [CrossRef]
22. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2015.
23. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
24. Shrestha, S.; Vanneschi, L. Improved fully convolutional network with conditional random fields for building extraction. *Remote Sens.* **2018**, *10*, 1135. [CrossRef]
25. Yang, G.; Zhang, Q.; Zhang, G. EANet: Edge-aware network for the extraction of buildings from aerial images. *Remote Sens.* **2020**, *12*, 2161. [CrossRef]
26. Tang, Z.; Chen, C.Y.C.; Jiang, C.; Zhang, D.; Luo, W.; Hong, Z.; Sun, H. Capsule–Encoder–Decoder: A Method for Generalizable Building Extraction from Remote Sensing Images. *Remote Sens.* **2022**, *14*, 1235. [CrossRef]
27. Liu, Y.; Chen, D.; Ma, A.; Zhong, Y.; Fang, F.; Xu, K. Multiscale U-Shaped CNN Building Instance Extraction Framework with Edge Constraint for High-Spatial-Resolution Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 6106–6120. [CrossRef]
28. 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]
29. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
30. Yuan, W.; Xu, W. Msst-net: A multi-scale adaptive network for building extraction from remote sensing images based on swin transformer. *Remote Sens.* **2021**, *13*, 4743. [CrossRef]
31. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–16 October 2021; pp. 10012–10022.

32. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
33. Girard, N.; Smirnov, D.; Solomon, J.; Tarabalka, Y. Polygonal building extraction by frame field learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 5891–5900.
34. Zhao, K.; Kang, J.; Jung, J.; Sohn, G. Building extraction from satellite images using mask R-CNN with building boundary regularization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 247–251.
35. Douglas, D.H.; Peucker, T.K. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartogr. Int. J. Geogr. Inf. Geovis.* **1973**, *10*, 112–122. [CrossRef]
36. Sohn, G.; Jwa, Y.; Jung, J.; Kim, H. An implicit regularization for 3D building rooftop modeling using airborne lidar data. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2012**, *1*, 305–310. [CrossRef]
37. 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]
38. Fang, F.; Wu, K.; Liu, Y.; Li, S.; Wan, B.; Chen, Y.; Zheng, D. A coarse-to-fine contour optimization network for extracting building instances from high-resolution remote sensing imagery. *Remote Sens.* **2021**, *13*, 3814. [CrossRef]
39. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
40. Zhao, W.; Persello, C.; Stein, A. Building outline delineation: From aerial images to polygons with an improved end-to-end learning framework. *ISPRS J. Photogramm. Remote Sens.* **2021**, *175*, 119–131. [CrossRef]
41. Zhu, Q.; Liao, C.; Hu, H.; Mei, X.; Li, H. MAP-Net: Multi Attending Path Neural Network for Building Footprint Extraction from Remote Sensed Imagery. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 6169–6181. [CrossRef]
42. Liu, X.; Chen, Y.; Wei, M.; Wang, C.; Goncalves, W.N.; Marcato, J.; Li, J. Building Instance Extraction Method Based on Improved Hybrid Task Cascade. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 3002005. [CrossRef]
43. Cai, Z.; Vasconcelos, N. Cascade R-CNN: High quality object detection and instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 1483–1498. [CrossRef]
44. Caesar, H.; Uijlings, J.; Ferrari, V. Coco-stuff: Thing and stuff classes in context. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1209–1218.
45. Lin, T.Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollár, P. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014.
46. Vu, T.; Kang, H.; Yoo, C.D. Scnet: Training inference sample consistency for instance segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 2701–2709.
47. Wang, X.; Zhang, R.; Kong, T.; Li, L.; Shen, C. Solov2: Dynamic and fast instance segmentation. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 17721–17732.
48. Fu, R.; He, J.; Liu, G.; Li, W.; Mao, J.; He, M.; Lin, Y. Fast seismic landslide detection based on improved mask R-CNN. *Remote Sens.* **2022**, *14*, 3928. [CrossRef]
49. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
50. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3349–3364. [CrossRef]
51. Xu, C.; Wang, G.; Yan, S.; Yu, J.; Zhang, B.; Dai, S.; Li, Y.; Xu, L. Fast vehicle and pedestrian detection using improved Mask R-CNN. *Math. Probl. Eng.* **2020**, *2020*, 5761414. [CrossRef]
52. Wu, B.; Shen, Y.; Guo, S.; Chen, J.; Sun, L.; Li, H.; Ao, Y. High Quality Object Detection for Multiresolution Remote Sensing Imagery Using Cascaded Multi-Stage Detectors. *Remote Sens.* **2022**, *14*, 2091. [CrossRef]
53. Zhang, Y.; Li, Q.; Tu, W.; Mai, K.; Yao, Y.; Chen, Y. Functional urban land use recognition integrating multi-source geospatial data and cross-correlations. *Comput. Environ. Urban Syst.* **2019**, *78*, 101374. [CrossRef]