International Journal of Applied Earth Observation and Geoinformation 122 (2023) 103420

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag

A lightweight building instance extraction method based on adaptive optimization of mask contour

Xiaoxue Liu [a,c], Yiping Chen [b,*], Cheng Wang [a], Kun Tan [d], Jonathan Li [e,*]

[a] Fujian Key Laboratory of Sensing and Computing for Smart Cities, School of Informatics, Xiamen University, Xiamen, China
[b] School of Geospatial Engineering and Science, Sun Yat-sen University, Zhuhai, China
[c] School of Big data and Computer Science, Guizhou Normal University, Guiyang, China
[d] East China Normal University, Key Laboratory of Spatial-Temporal Big Data Analysis and Application of Natural Resources in Megacities (Ministry of Natural Resources), Shanghai, China
[e] Department of Geography and Environmental Management and Department of Systems Design Engineering, University of Waterloo, Waterloo, Canada

ARTICLE INFO

Keywords:
Building instance extraction
Remote sensing imagery
Deep learning
Mask contour
Adaptive and lightweight

ABSTRACT

Automatic extraction of building instances from high spatial resolution optical remote sensing imagery is essential for urban infrastructure and smart management. In view of the severe challenges such as intricate building samples, high training overhead and inaccurate mask contours, this paper proposes a region-based, two-stage segmentation model assembled with components of adaptive feature extraction, guided anchoring and iterative subdivision mask, resulting in lightweight extraction of building instances and adaptive optimization of mask contours. We conducted comparison experiments with several classical, state-of-the-art instance segmentation methods on WHU aerial dataset, China city satellite dataset and Inria aerial dataset. The quantitative evaluation demonstrates that our method signally reduces the computational load by at least 34.5%, increases the mask AP on the three datasets by at least 0.8%, 4.6% and 4.1%, respectively, and increases the mean performance (mPC) and relative performance (rPC) under image corruption (12 corruptions and 2 severity levels) by at least 2.4% and 3.2%, respectively. The qualitative evaluation further verifies that our method is more discriminative for features such as color, texture, and shadow, and more adaptable to changes in shape, scale, and distribution and so on. Especially for large-sized buildings, various comparison methods either miss the partial interior area or excessively smooth the mask contour, whereas our method can output complete and accurate masks.

1. Introduction

Automatic extraction of buildings from remote sensing imagery can provide building geospatial data with wide coverage, precise spatial detail, and high update frequency for the fields of land planning, ecological protection, and disaster management. As the improvement of optical sensor performance, the decrease of remote sensing data acquisition charge and the development of instance segmentation technology, more and more researchers carry out theoretical research and experimental demonstration around the automatic extraction of building instances from high spatial resolution optical remote sensing imagery.

Automatic extraction of building instances from optical remote sensing imagery still faces severe challenges so far: (1) The complexity of building samples grows geometrically in different temporal and spatial dimensions. Smaller inter-class variance (e.g., courtyards, open parking lots, vehicles, containers, and many other man-made non-building objects may have similar spectral or spatial features as rooftops) and larger intra-class variance (involving the diversity and extreme of rooftops in terms of color/texture, geometry, size/aspect ratio, location/distribution, etc.) may lead to various extraction errors (Ghanea et al., 2016; Alshehhi et al., 2017). (2) High spatial resolution, massive remote sensing images, and high quality segmentation often imply high training overhead. (3) Existing region-based instance segmentation models such as Mask R-CNN (He et al., 2017), HTC (Chen et al., 2019), etc., usually predict masks based on low-resolution (28×28) regular grids. Such a compromise setting that balances oversampling versus undersampling may excessively smooth the mask contour of large-sized buildings.

To address the above technical bottlenecks, this paper proposes a

* Corresponding authors.
E-mail addresses: liuxiaoxue@stu.xmu.edu.cn, lxx@gznu.edu.cn (X. Liu), chenyp79@mail.sysu.edu.cn (Y. Chen), cwang@xmu.edu.cn (C. Wang), tankuncu@gmail.com (K. Tan), junli@uwaterloo.ca (J. Li).

https://doi.org/10.1016/j.jag.2023.103420
Received 14 March 2023; Received in revised form 19 June 2023; Accepted 8 July 2023
Available online 20 July 2023

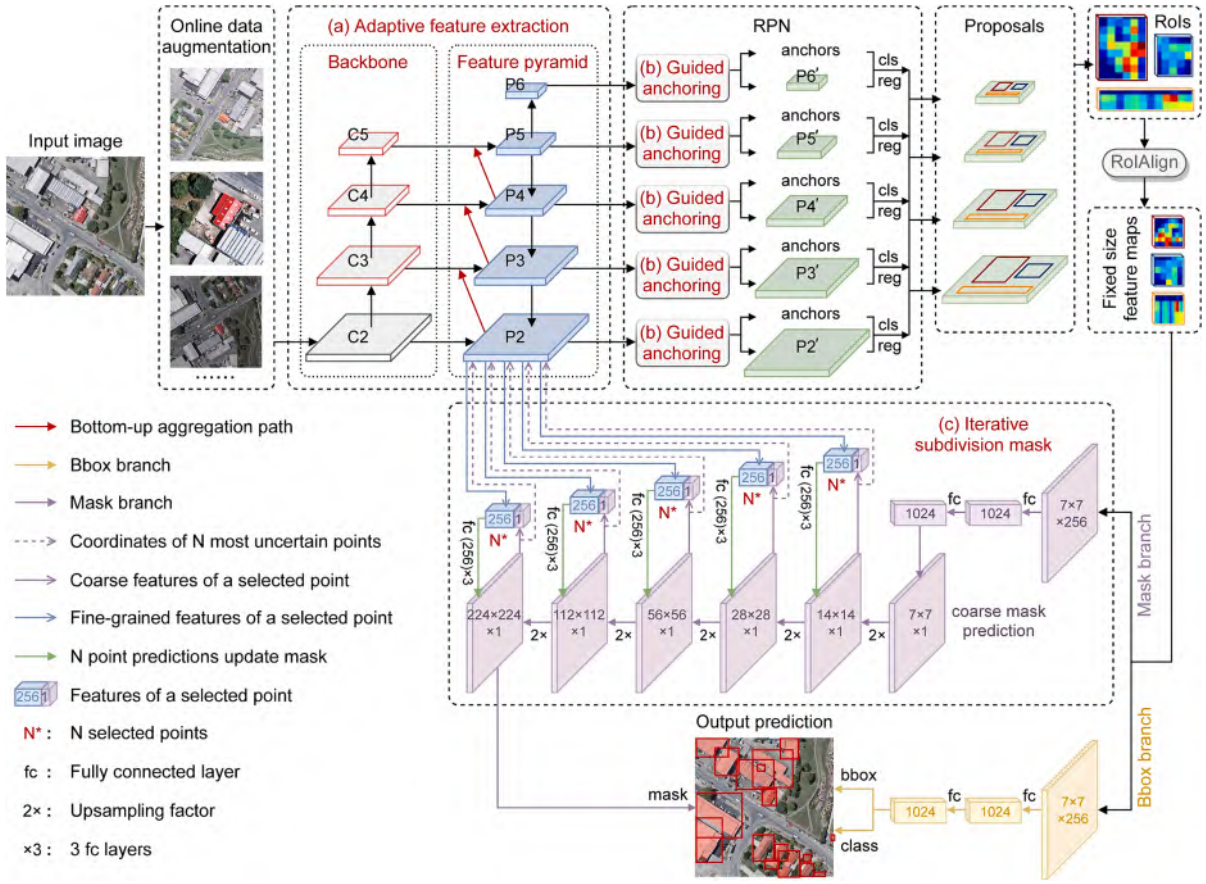

Fig. 1. Overall architecture of the proposed method. The main components include: (a) Adaptive feature extraction network; (b) Guided anchoring module; (c) Iterative subdivision mask module.

lightweight building instance extraction method based on adaptive optimization of mask contour. The contributions of this paper include: (1) Design an adaptive feature extraction network to cope with the diverse appearance and extreme scale of rooftops. (2) Construct a guided anchoring module to generate non-redundant, high-quality anchors to economize on computational resources and match rooftop shapes (width and height). (3) Assemble a lightweight iterative subdivision mask module to adaptively optimize the mask contour of large-sized buildings.

The rest of this paper is arranged as follows: Section 2 reviews the research status of building extraction. Section 3 describes the overall architecture and main components of the proposed method. Section 4 reports the experimental details and main results. Section 5 concludes our work.

2. Related work

Traditional methods need to design appropriate building features according to professional knowledge and cumulative experience, and then combine corresponding strategies and algorithms to semi-automatically extract pixel-level buildings from aerial or satellite-borne remote sensing imagery (Turker and Koc-San, 2015; Chaudhuri et al., 2016; Zhang et al., 2016; Li et al., 2017; Xu et al., 2018; Bi et al., 2019). Most of these methods tend to manually design low-rise features based on specific data (Liu et al., 2019), which is tough to maintain high extraction accuracy for discrepant imaging conditions, different geographical regions and abundant application requirements.

In recent years, deep learning methods have significantly improved the accuracy and automation of interpretation by automatically learning the multi-level and multi-type (e.g. spectral features and spatial features) building features, which fundamentally promotes remote sensing image analysis and geoscience application research to get rid of the labor-intensive "tradition" gradually. Numerous semantic segmentation methods have emerged successively based on architectures such as convolutional neural networks (CNNs), generative adversarial networks (GANs) and graph convolution networks (GCNs) to automatically extract pixel-level buildings from aerial or satellite-borne remote sensing imagery (Alshehhi et al., 2017; Li et al., 2022; Abdollahi et al., 2020; Wei and Ji, 2022). The mainstream FCNs (Long et al., 2015) paradigm (Shrestha and Vanneschi, 2018; Zhu et al., 2021) includes U-Net (Liu et al., 2019; Girard et al., 2021; Guo et al., 2022), DeconvNet (Huang et al., 2016), SegNet (Sun et al., 2018), DeepLabv3+ (Li and Xin, 2021) and other variants of encoder-decoder based CNNs architecture (Kotaridis and Lazaridou, 2021). Emerging instance segmentation methods identify both semantic labels (background or building) and assign building numbers for each pixel in order to distinguish the image areas corresponding to different buildings (i.e. building instances) and thus accurately obtain object-level building information such as the location, contour, and area of each rooftop. Reviewing the related domestic and international literature from 2018 to the present, building instance extraction has made encouraging progress through research on attempting contour modeling, improving backbone networks, designing fusion mechanisms or parallel/cascade structures for multi-scale features, setting dynamic or rotatable anchors, optimizing or vectorizing mask contours, and exploring object-level evaluation (Zhao et al., 2018; Wen et al., 2019; Li et al., 2019; Wu et al., 2020; Li et al., 2020; Zhang et al., 2020; Fang et al., 2021a; Liu et al., 2021a, 2021b; Xu et al., 2021; Huang et al., 2022; Zorzi et al., 2022). Among them, solutions to bridge the mask contour discrepancy between deep neural network output and downstream application requirements each have shortcomings: On the

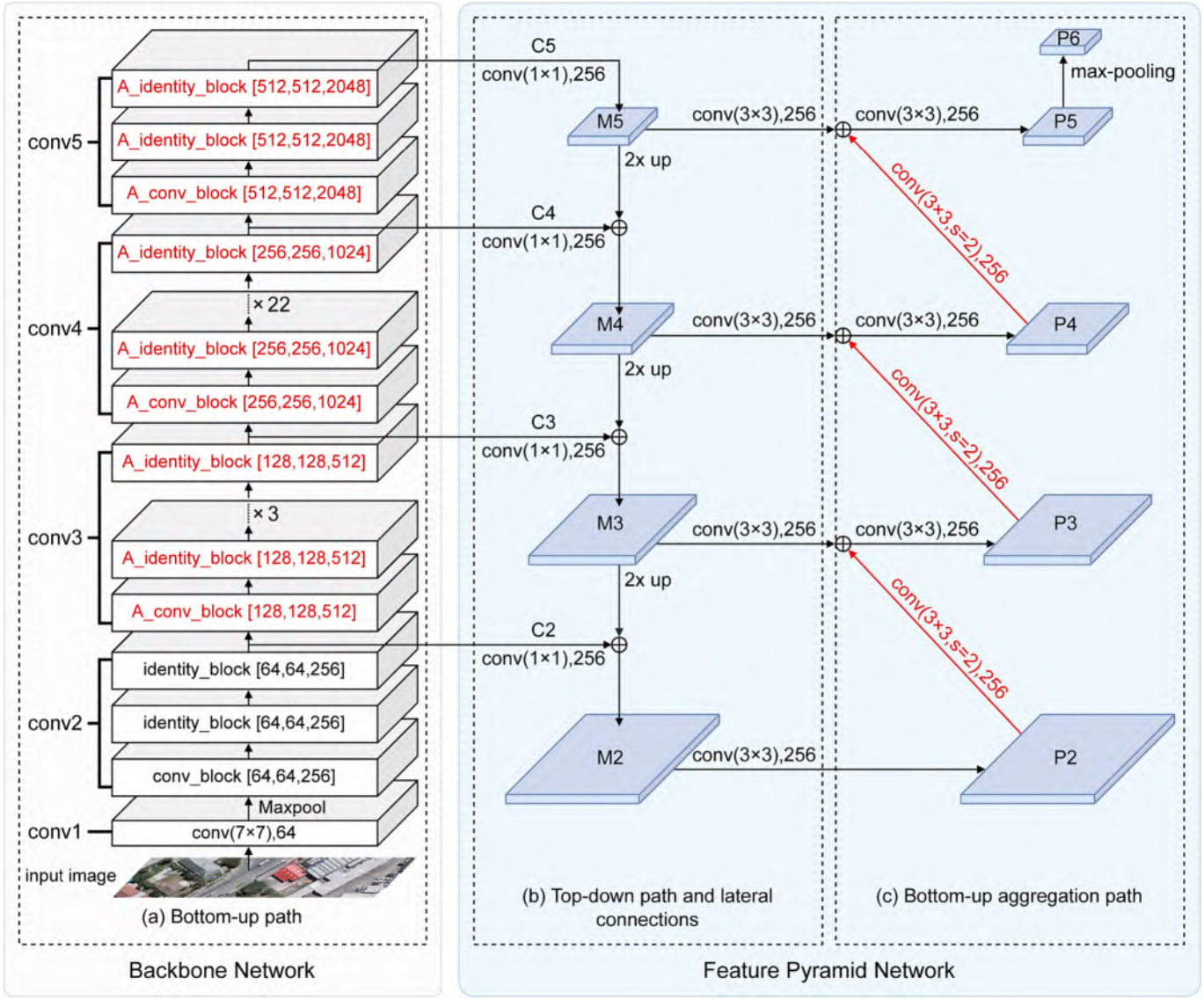

Fig. 2. Architecture of the adaptive feature extraction network.

one hand, no matter integrating traditional edge algorithms (Zhao et al., 2018; Zhang et al., 2020), or combining masks with keypoints (Li et al., 2020), or adjusting the edge constraint loss (Liu et al., 2021b), or sensing the geometric center of each rooftop (Xu et al., 2021), the mainstream Region-based CNN (R-CNN) paradigm cannot so far solve the challenge of excessively smoothing the mask contour of large-sized buildings due to multiple constraints such as model architecture and training overhead. On the other hand, although the contour modeling paradigm learns vector representation directly, the finite vertices encoded along the rooftop edges can only output simple polygonal masks without holes (Wu et al., 2020; Fang et al., 2021a). Furthermore, the training overhead (computation, memory, and training time) of integrated recurrent neural networks (RNNs) or graph neural networks (GNNs) is much higher than that of CNNs, and the repeated use of the same line segments to extract adjacent buildings is prone to more problems such as overlapping masks or excessive gaps between each other (Li et al., 2019; Zorzi et al., 2022). In general, building instance extraction has more practical value than semantic segmentation, but its segmentation accuracy and extraction efficiency currently cannot meet the demand for robustness and generalization in engineering applications, and more academic attention and cross-disciplinary communication are urgently needed.

3. Method

The overall architecture of the proposed method is illustrated in Fig. 1: (1) Data augmentation pipeline generates new random training samples online. (2) Adaptive feature extraction network (composed of backbone network and feature pyramid network) adaptively extracts and repeatedly fuses multi-scale features. (3) Guided anchoring module adaptively generates sparse anchors with shapes (width and length) that match the rooftop size and aspect ratio based on feature semantics, and adjusts the feature map receptive field to the anchor shapes. Region proposal network (RPN) further classifies (cls) and regresses (reg) according to the adjusted multi-level feature maps to generate regional proposals. (4) Regions of interest (RoIs) dynamically corresponding to these proposals are extracted from the corresponding level of feature maps. RoIs are pooled into fixed size feature maps by RoIAlign layer and then input into the bounding box (bbox) branch and mask branch, respectively. (5) Bbox branch is used for each RoI to correct the bounding box coordinates and predict the classification score. (6) Mask branch is used to segment the mask for each RoI. Iterative subdivision mask module optimizes its contour even further. In each iteration, the previously predicted mask is upsampled by 2 × and then the N most uncertain points that may located at the rooftop edge are adaptively

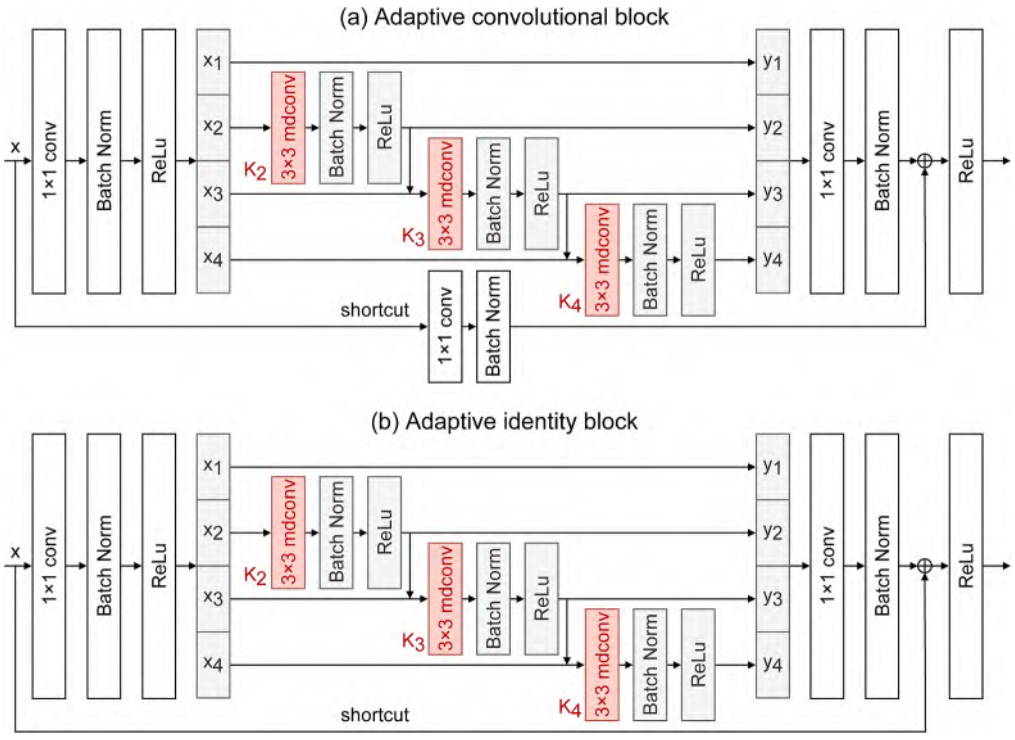

Fig. 3. Architecture of the improved residual block. (a) Adaptive convolutional block, (b) Adaptive identity block, corresponding to "A_conv_block" and "A_identity_block" in Fig. 2(a).

selected on this more dense regular grid. Construct feature representations for each of these N points individually, and then use a multi-layer perceptron (MLP, fully connected) to re-predict the label for each selected point, thereby optimizing the entire mask contour. In the following sections, several main components of the method are described in detail.

3.1. Adaptive feature extraction

At the red markers in Figs. 1(a), 2, and 3, the adaptive feature extraction network (denoted as AR101-BPFPN) improves the residual blocks and aggregates the bidirectional paths to accommodate the diverse appearances and extreme scales of rooftops in remote sensing scenes.

As depicted in Figs. 1(a) and 2(a), in the backbone network from conv3 to conv5 stages, we replace the three sets of 3 × 3 standard convolutions connected in a hierarchical residual style in Res2Net block (Gao et al., 2021) with modulated deformable convolutions (Zhu et al., 2019). Specifically, the architecture of the improved residual block is detailed in Fig. 3: The input feature maps x are equally divided into 4 input feature map subsets $x_i (i \in \{1,2,3,4\})$ by 1 × 1 convolution. Each x_i except x_1 has a corresponding 3 × 3 modulated deformable convolution (i.e. 3 × 3 mdconv in Fig. 3) K_i. The output feature map subset $y_i (i \in \{1,2,3,4\})$ corresponding to each x_i can be formulated as follows:

$$y_i(p) = \begin{cases} x_i(p) i = 1, \\ \sum_{k=1}^{9} x_i(p + p_k + \Delta p_k) \bullet w_k \bullet \Delta m_k \quad i = 2, \\ \sum_{k=1}^{9} z_i(p + p_k + \Delta p_k) \bullet w_k \bullet \Delta m_k \quad i \in \{3,4\}. \end{cases} \tag{1}$$

where $x_i(p)$, $z_i(p)$, and $y_i(p)$ are the features of the sampling position p in the input feature map subset x_i, the input feature map subset combination $z_i(z_i = x_i + y_{i-1})$, and the output feature map subset y_i, respectively; $p_k \in \{(-1,-1),(-1,0),\cdots,(1,1)\}$ and w_k are the pre-specified offset and corresponding weight for the k-th sampling position, respectively; Δp_k and Δm_k are the learnable offset and the corresponding modulation weight scalar for the k-th sampling position, respectively. Each K_i can receive more potential features than K_{i-1}, and each y_i can further adaptively expand the effective receptive field than y_{i-1}. Stacking more modulated deformable convolutions (Zhu et al., 2019) further expands the deformation range of geometric modeling, which makes the feature extraction network more adaptable to the diverse changes in rooftop appearance.

As depicted in Figs. 1(a) and 2(c), in the feature pyramid network, we add a bottom-up aggregation path with slight computational/memory load (Liu et al., 2018). As detailed in Fig. 2(c), shortening the path from the bottom to the top (less than 10 convolutional layers apart between *P2* and *P5*) preserves more edge and shape features in the output feature map that are conducive to identifying large-sized buildings, which makes the feature extraction network more adaptable to extreme changes in rooftop scale.

3.2. Guided anchor generation

Traditional RPN (Ren et al., 2017) generates anchors that are uniformly and densely distributed throughout the entire image region, and relies on prior knowledge to predefine a set of fixed anchor shapes (size and aspect ratio). On the basis of the semantic-guided design (Wang et al., 2019), we retune the shape projection and multi-task loss function, and ultimately generate sparsely distributed and shape-learnable (width and height) anchors only on the rooftop region to cope with the non-uniform rooftop distribution wasting a large number of anchoring computations and the diversiform rooftop sizes and aspect ratios.

As illustrated in Figs. 1(b) and 4, for a feature map P_i input from the feature pyramid, the guided anchoring module constructs the anchor location branch, anchor shape branch and feature adaptation branch to output a set of dynamic anchors represented in 4-tuple form (x, y, w, h) and a feature map P_i' with corrected receptive field, and the parameters for generating anchors are shared among all feature map levels. More

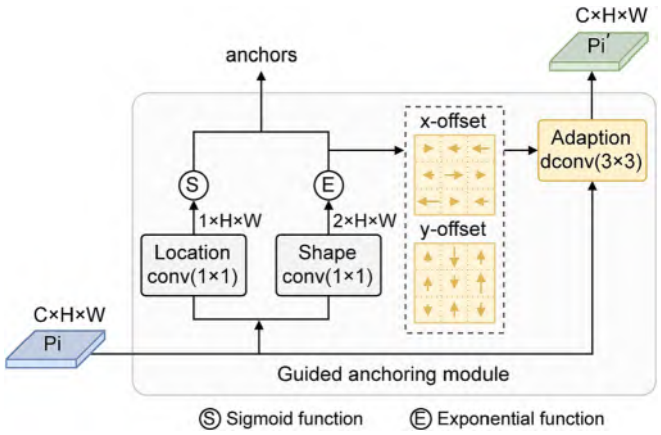

Fig. 4. Architecture of the guided anchoring module.

specifically, the architecture of the guided anchoring module is detailed in Fig. 4: (1) A 1 × 1 convolutional layer and an element-wise sigmoid function make the anchor location branch to only retain the location coordinates (x, y) in P_i that may be located in the central region of rooftops (that is, with conditional probability $p(\bullet|P_i)$ above the irrelevant region threshold λ). (2) A 1 × 1 convolutional layer and an element-wise exponential function enable the anchor shape branch to obtain a most likely shape (w, h) (that is, with the highest coverage compared to the nearest ground truth bounding box) corresponding to each retained location in P_i. The exponential function projects the predicted exponential (dw, dh), which is restricted to a small range, more stably onto the actual predicted shape (w, h):

$$w = W_i/2 \bullet e^{dw}, h = H_i/2 \bullet e^{dh} \tag{2}$$

where W_i and H_i are the width and height of corresponding level feature map P_i, respectively. (3) A 3 × 3 deformable convolutional (Dai et al., 2017) layer takes the predicted shape as offset field so that the feature adaptation branch corrects the receptive field of feature map P_i' to match the anchor shape. (4) Adjust $\mathscr{L}_{rpn}$ (multi-task loss function in RPN) as follows:

$$\mathscr{L}_{rpn} = \mathscr{L}_{loc} + \mathscr{L}_{shape} + \mathscr{L}_{cls} + \mathscr{L}_{reg} \tag{3}$$

where both $\mathscr{L}_{loc}$ (anchor location loss) and $\mathscr{L}_{cls}$ (classification loss) are Focal Loss (Lin et al., 2017), $\mathscr{L}_{shape}$ (anchor shape loss) and $\mathscr{L}_{reg}$ (regression loss) adopt Bounded IoU Loss (Tychsen-Smith and Petersson, 2018) and L1 Loss, respectively.(5) Anchors represented in 4-tuple form (x, y, w, h) are combined with the modified feature map P_i' for further classification and regression.

3.3. Iterative subdivision mask

The mainstream R-CNN (He et al., 2017) paradigm typically predicts segmentation masks based on low-resolution (28 × 28) regular grids to balance oversampling low-frequency regions (i.e. object internal regions) while undersampling high-frequency regions (i.e. object boundary regions). Given that the low-resolution setting may excessively smooth the mask contours of large-sized buildings while the high-resolution setting may substantially increase the direct and intensive computation, we assemble an iterative subdivision mask module (Kirillov et al., 2020) to adaptively sample a small quantity of non-uniform points (i.e. the most uncertain points) in the subdivided high-frequency region and then iteratively predict their segmentation labels, so as to output high-resolution (224 × 224) building masks with only a small amount of floating-point computations.

As illustrated in Figs. 1(c) and 5, the iterative subdivision mask module adopts a lightweight design of coarse mask head combined with point mask head (Kirillov et al., 2020): (1) Coarse mask head (i.e. an MLP with 2 hidden layers and 1024 output channels) similar to bbox head predicts the coarsest mask (7 × 7) for each RoI (7 × 7 feature map). (2) In the coarse mask subdivided by bilinear interpolation, N points with the most uncertain segmentation are adaptively selected. (3) Interpolated features obtained from both *P*2-level feature map and existing coarse mask prediction are concatenated as the point-wise features of each selected point. Fine-grained features and coarse

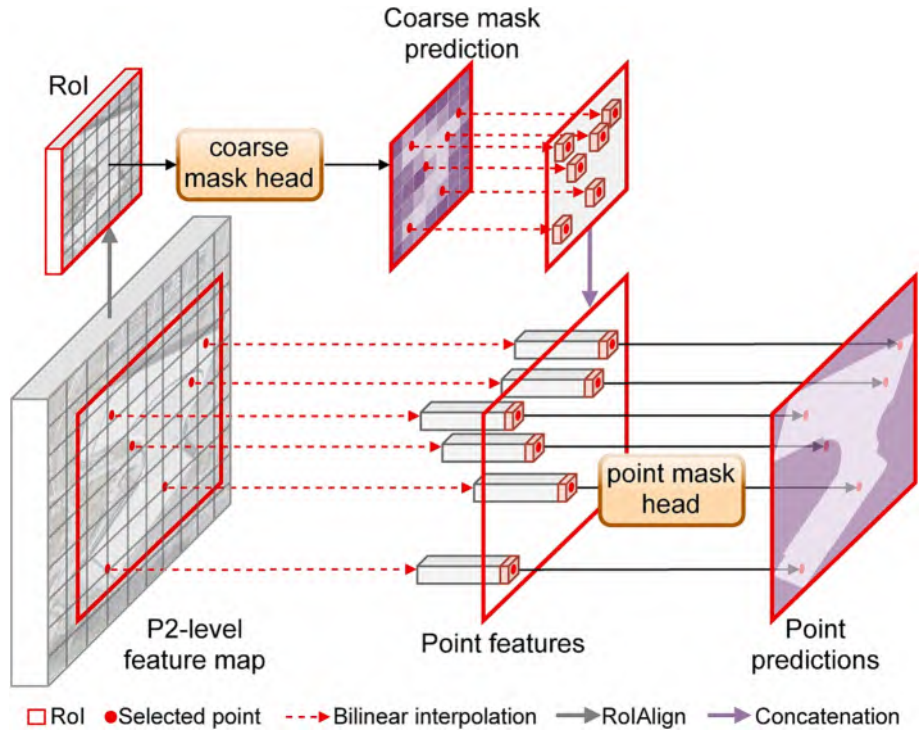

Fig. 5. Illustration of the iterative subdivision mask module.

Table 1
Description of several building datasets.

Dataset	WHU aerial dataset	China city satellite dataset	Inria aerial dataset
GSD (m)	0.3	0.29	0.3
Area (km^2)	450	120	405
Regions	1 New Zealand city	4 Chinese cities	3 U.S. cities, 2 Austrian cities
Data type	image/raster	image/label	image/raster
Bands	RGB	RGB	RGB
Data format	TIFF	TIFF/JSON	TIFF
Tiles	8188	7260	180
Tile size (pixels)	512 × 512	500 × 500	5000 × 5000

Table 2
Experimental configuration for training and testing.

Dataset	WHU aerial dataset	China city satellite dataset	Inria aerial dataset
Label format	JSON	JSON	JSON
Tile size (pixels)	512 × 512	500 × 500	512 × 512
Training tiles	5772	5985	9216
Testing tiles	2416	1275	2304

prediction features provide fine detail information and global context information for each selected point, respectively. (4) The trained point mask head (i.e. an MLP with 3 hidden layers and 256 channels, where the hidden and output layers are activated by ReLU and sigmoid functions, respectively) independently predicts the segmentation label for each selected point. The point mask head shares weights across all selected points and all RoIs. (5) Update these point predictions to the corresponding grid locations in the upsampled mask. Iterate the steps from (2) to (5) above until the mask is upsampled to the desired resolution (224 × 224).

The core task of this module is to adaptively select a small number of points (rather than all points) that are more densely located near high-frequency regions (i.e. rooftop edges). Different point selection strategies and number of points are used during inference and training. During inference, $N = 28^2$ points (at most) are selected by an iterative subdivision strategy: Except for the first subdivision step where only 14^2 points are available for selection, the most uncertain (that is, segmentation confidence close to about 0.5 or so) 28^2 points are selected from the subdivided mask in each iteration. This strategy only needs to predict 4.25×28^2 points, about 15 times less than the direct, intensive calculation of 224^2 points. During training, $N = 14^2$ points are selected by a mildly biased random sampling strategy (oversampling ratio $k = 3$, important sampling ratio $\beta = 0.75$) to train the point mask head, which balances performance and training overhead more than back-propagation through subdivision steps: $\beta \times 14^2$ points with the most uncertain segmentation (that is, the rooftop probability interpolated from the coarse mask prediction is close to about 0.5 or so) are selected from $k \times 14^2$ candidate points randomly sampled, and then another $(1-\beta) \times 14^2$ points are selected from the remaining candidate points uniformly distributed. Reasonable setting of oversampling ratio ($k > 1$) and important sampling ratio ($\beta \in [0, 1]$) can ensure that this strategy is more biased towards uncertain regions and simultaneously maintains a certain degree of uniform distribution.

4. Experiments

4.1. Datasets

The brief descriptions of several building datasets (e.g., ground sampling distance abbreviated as GSD) and the experimental configurations for training and testing are listed in Tables 1 and 2, respectively. WHU aerial dataset (Ji et al., 2019) was photographed in Christchurch, New Zealand, which contains various types of urban areas (e.g. commercial, industrial, residential, suburban, etc.) and buildings with diverse styles and extreme scales. China city satellite dataset (Fang et al., 2021a) provides 19-level satellite images (both orthophoto and non-orthophoto) of 4 typical cities, while Inria aerial dataset (Maggiori et al., 2017) covers 5 cities in the United States and Austria (e.g. densely populated Chicago, alpine town West Tyrol). Most of our experiments were conducted on WHU aerial dataset (Ji et al., 2019), and the other two datasets were used to further evaluate the generalization capabilities. Our data preprocessing includes: (1) Transform the label format from "TIFF" to COCO-style (Lin et al., 2014) "JSON" to fit deep learning methods. (2) Normalize the size and filenames of tiles in various datasets. Taking Inria aerial dataset (Maggiori et al., 2017) as an example: each tile is uniformly cropped into 8 rows × 8 columns, and then these small tiles are fine-tuned to 512 × 512 pixels. All tiles are renamed according to the rule of "region_id + tile_id + crop_row_id + crop_column_id". (3) During data loading, the samples without buildings are automatically cleared to reduce training overhead.

4.2. Implementation details

Configuration and hyperparameters for all methods: (1) Implementation is based on PyTorch deep learning framework and a single 16 GB GPU (Tesla V100) hardware environment. (2) Initialize from backbone models pre-trained on ImageNet dataset (Deng et al., 2009). (3) Both data augmentations (Aug.) are applied with 50% probability during training: random rotation by 90 degrees and cropping the random part of the input (scale range 384–480). (4) The default training schedule is 36 epochs, and Stochastic Gradient Descent (SGD) sets the initial learning rate to 0.0025 and the momentum factor to 0.9. The learning rate decreases at the 27th and 33nd epochs, the corresponding weight decay factor is 0.0001. QueryInst (Fang et al., 2021b) employs AdamW optimizer alone (designed by the method itself). (5) 4 images per GPU in each training batch. HTC (Chen et al., 2019) and QueryInst (Fang et al., 2021b) trained on Inria aerial dataset, DetectoRS (Qiao et al., 2021) trained on all datasets, with 2 images per GPU (limited by computational resources).

Data augmentation for our method. We design the transformation combination and application probability of data augmentation from the pixel and spatial dimensions (implemented on CPU). On the one hand, randomly select one transformation or combine multiple transformations from brightness, contrast, RGB offset, hue and saturation to modify the pixel value of the loaded image. On the other hand, randomly select one transformation from random rotation by 90 degrees, horizontal flip, vertical flip, diagonal transpose to apply the same spatial

Table 3
Comparison of complexity and memory overhead. Memory tested on WHU aerial dataset.

Method	Category	Input shape	Flops (G)	Params (M)	Mem (GB)
Mask R-CNN	Two-stage	512 × 512	134.20	62.74	4.4
SOLOv2	One-stage	512 × 512	422.81	**55.12**	**4.0**
HTC	Two-stage	512 × 512	278.85	98.67	9.9
DetectoRS	Two-stage	512 × 512	261.41	199.30	14.6
QueryInst (300 queries)	Two-stage	512 × 512	1021.65	191.26	12.6
Our method	Two-stage	512 × 512	**87.94**	81.66	4.7

Table 4
Comparison of accuracy and completeness on WHU aerial dataset.

Method	Backbone	Aug.	Epochs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR
Mask R-CNN	ResNet101-FPN	√	36	59.3	83.0	70.7	61.0	33.0	40.9	63.7
SOLOv2	ResNet101-FPN	√	36	60.1	84.1	71.1	61.5	7.5	39.4	64.4
HTC	ResNet101-FPN	√	36	60.6	84.8	71.3	62.2	29.0	39.5	64.7
DetectoRS	ResNet101-FPN	√	36	63.8	84.9	75.1	65.5	28.2	28.3	68.6
QueryInst (300 queries)	ResNet101-FPN	√	36	66.8	**90.2**	77.4	68.9	9.0	30.4	**74.8**
Our method	AR101-BPFPN	√	36	**67.6**	90.0	**78.8**	**69.0**	**44.6**	**47.6**	72.1

transformation to the loaded image and label. Extending the data pipeline online enables our method to effectively avoid overfitting and simultaneously improve model performance at minimal cost.

4.3. Evaluation metrics

Mask evaluation metrics. AP (i.e. average precision at IoU thresholds from 0.5 to 0.95 with an interval of 0.05) and AR (i.e. average recall given 300 detections per image) based on COCO evaluation benchmark (Lin et al., 2014) quantitatively evaluate the accuracy and completeness of object-level masks. In addition to mask AP (%) as the primary challenge metric, there are mask AP50 and AP75 each corresponding to a single IoU threshold (0.50 and 0.75). Subscripts S, M and L represent the appraised scales of buildings based on COCO evaluation benchmark (Lin et al., 2014). S: small-sized buildings (rooftop area less than 32 × 32). M: medium-sized buildings (rooftop area between 32 × 32 to 96 × 96). L: large-sized buildings (rooftop area greater than 96 × 96).

Robustness evaluation metrics. Performance on clean data (P_{clean}), mean performance under corruption (mPC) and relative performance under corruption (rPC) (Hendrycks and Dietterich, 2019; Michaelis et al., 2019) quantitatively evaluate the robustness to common image corruption. We simulate 12 types of image quality damages from 4 dimensions of blur, noise, weather and digital, and subdivide each type into 2 severity levels (i.e. less extreme but perceptible image distortion) (Hendrycks and Dietterich, 2019). mPC and rPC evaluate the average performance and relative performance degradation under corruption respectively (Michaelis et al., 2019):

$$\mathrm{mPC} = \frac{1}{N_c}\sum_{c=1}^{N_c}\frac{1}{N_s}\sum_{s=1}^{N_s}P_{c,s}, \tag{4}$$

$$\mathrm{rPC} = \frac{\mathrm{mPC}}{\mathrm{P_{clean}}} \tag{5}$$

where $P_{c,s}$ is a specific performance measure to evaluate the test data of corruption c under severity-level s, while $N_c = 12$ and $N_s = 2$ represent the number and severity-level of corruption respectively, $\mathrm{P_{clean}}$ is mask AP (i.e. the COCO "average precision" metric).

Visualization rules. We customize a set of notation and color rules to highlight the results: (1) The best results are shown in bold all tables; (2) Yellow boxes mark buildings that are entirely omitted; (3) Green boxes or arrows mark semantic confusion areas (that is, pixel areas that confuse buildings with the background, or are repeatedly identified as different buildings); (4) Blue boxes or arrows mark semantic lack areas (that is, pixel areas that are not completely identified inside buildings).

Table 5
Comparison of robustness on WHU aerial dataset. The higher the values of mPC and rPC, the better the performance. Square brackets denote evaluation metric.

Method	P_{clean} [AP]	mPC [AP]	rPC [%]
Mask R-CNN	59.3	17.4	29.3
SOLOv2	60.1	18.8	31.3
HTC	60.6	18.8	31.1
DetectoRS	63.8	19.9	31.1
QueryInst (300 queries)	66.8	21.6	32.4
Our method	**67.6**	**24.0**	**35.6**

4.4. Results

4.4.1. Quantitative results

Tables 3 to 7 and Fig. 6 present the quantitative evaluation results of our method compared to several classical, state-of-the-art instance segmentation methods.

Complexity **vs**. memory overhead. Our method does not rely on intricate designs to obtain performance gains. Table 3 demonstrates that lightweight components enable our method to reasonably balance and efficiently reduce the computational load (Flops), parameter scale (Params) and memory overhead (Mem), especially the computational

Table 6
Comparison of generalizability (similar style buildings) on China city satellite dataset.

Method	Backbone	Aug.	Epochs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR
Mask R-CNN	ResNet101-FPN	√	36	40.3	64.8	44.8	48.2	3.0	12.8	50.7
SOLOv2	ResNet101-FPN	√	36	41.0	67.1	44.4	49.6	2.3	11.4	54.7
HTC	ResNet101-FPN	√	36	42.1	67.0	46.4	50.6	5.2	9.8	54.5
DetectoRS	ResNet101-FPN	√	36	41.9	65.9	46.4	50.6	6.5	6.8	54.7
QueryInst (300 queries)	ResNet101-FPN	√	36	42.7	66.8	47.8	52.2	**17.1**	**23.4**	**62.5**
Our method	AR101-BPFPN	√	36	**47.3**	**71.9**	**52.9**	**56.0**	5.5	17.0	60.4

Table 7
Comparison of generalizability (dissimilar style buildings) on Inria aerial dataset.

Method	Backbone	Aug.	Epochs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR
Mask R-CNN	ResNet101-FPN	√	36	35.2	63.4	36.7	39.0	10.8	10.1	42.8
SOLOv2	ResNet101-FPN	√	36	36.0	65.8	36.8	40.7	2.6	5.6	45.2
HTC	ResNet101-FPN	√	36	39.3	69.9	41.5	43.9	5.1	39.6	48.2
DetectoRS	ResNet101-FPN	√	36	40.6	70.5	43.8	45.4	10.1	9.3	49.9
QueryInst (300 queries)	ResNet101-FPN	√	36	34.6	63.8	35.1	38.8	8.4	10.3	48.9
Our method	AR101-BPFPN	√	36	**44.7**	**74.4**	**50.2**	**48.7**	**12.5**	**43.4**	**54.5**

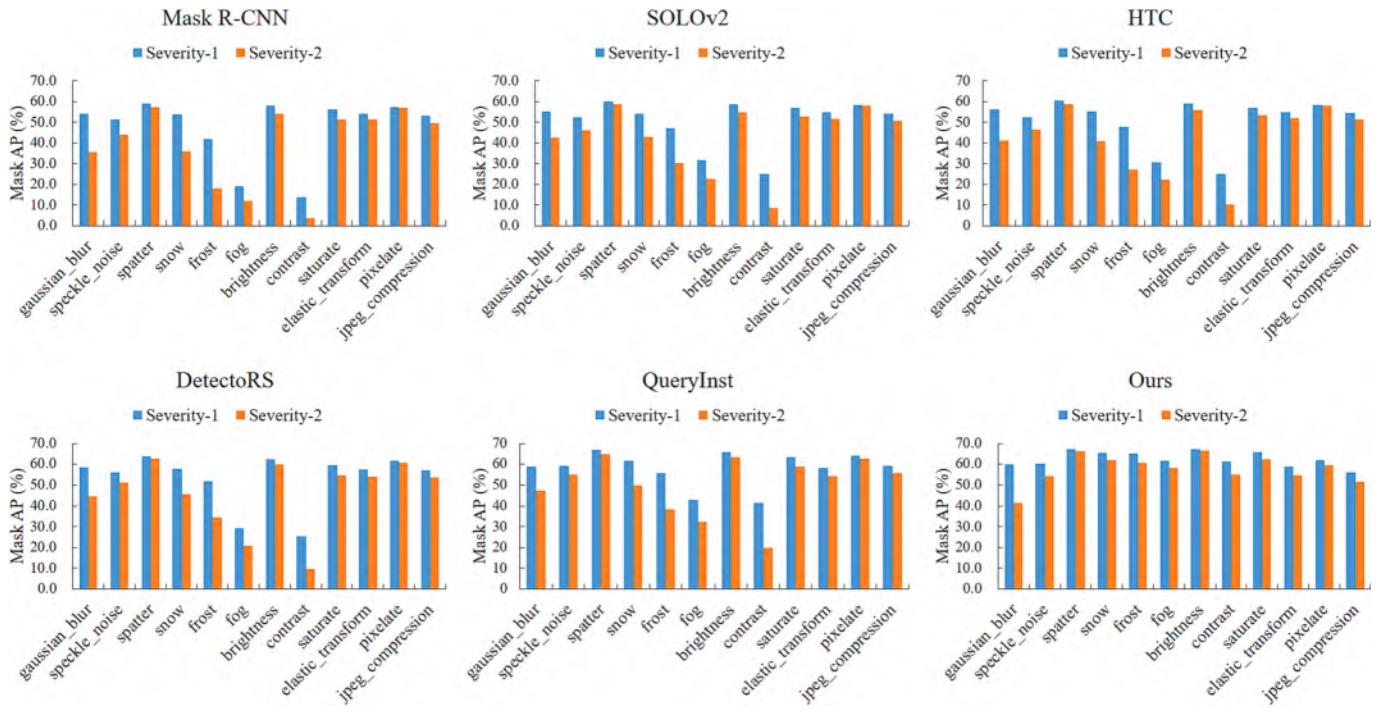

Fig. 6. Comparison of mask AP (%) for various methods on 12 corruptions and 2 severity levels.

load (Flops) by at least 34.5%.

Accuracy **vs**. completeness. On WHU aerial dataset, Table 4 shows that our method surpasses other methods by at least 0.8% of the mask AP, while the mask AR is 2.7% lower only than QueryInst (Fang et al., 2021b). For medium-sized buildings with scarce samples, the mask APM corresponding to SOLOv2 (Wang et al., 2020) and QueryInst (Fang et al., 2021b) severally plummet to just 7.5% and 9.0%. On the contrary, our method simultaneously improves the accuracy at different scale ranges, especially the mask AP_M and AP_L with fewer samples are conspicuously increased by at least 11.6% and 6.7%, respectively. These comparisons effectively validate that our method is more adaptable to the scale change of buildings and the sample distribution of remote sensing scenes.

Robustness**.** Investigate the vulnerability of CNNs to familiar corruptions. Existing computer vision systems are not stable robustness of human vision system (Azulay and Weiss, 2018). Deep learning classifiers may be confused by many forms of corruption (such as blur, noise, contrast, saturation, pixelation, data compression, snow, fog, rain, etc.) or be deceived by small changes in the target object (such as partial areas being occluded) (Hendrycks and Dietterich, 2019). Table 5 shows that all instance segmentation models suffer severe performance impairments under common image corruption. Our method outperforms other comparison methods in all robustness evaluation metrics: mPC increased by at least 2.4% and rPC increased by at least 3.2%. Fine-tuning one corruption type does not improve the performance of other corruption types (Geirhos et al., 2018). As detailed in Fig. 6, our method is most vulnerable to "blur" and "image compression". Furthermore, as the severity level increases, the performance degradation caused by "blur" is also much more drastic than the fine-tuned digital categories (e.g., contrast, saturation, etc.).

Generalizability**.** Test the generalizability of similar or dissimilar style buildings across multiple data sources. On China city satellite dataset, Table 6 shows that our method surpasses other methods by at least 4.6% mask AP, while mask AR is lower only than QueryInst (Fang et al., 2021b) by 2.1%. Multiple factors such as numerous non-orthophoto images, non-uniform distribution of sample scales, and intricate interference details on medium and large-sized rooftops combine to result in very low mask AP_M and AP_L for all methods. Especially, some low definition images (Wuhan urban area) trigger strong performance oscillations. As shown in Fig. 6, our method is most susceptible to "blur" damage; on the contrary, QueryInst (Fang et al., 2021b) is the most robust to "blur" damage among all methods and thus outperforms our method on AP_M and AP_L. Compared with China city satellite dataset, instead of splitting the same urban areas into training and test subsets simultaneously, Inria aerial dataset divides different cities into training and test subsets respectively, which further challenges the generalizability of deep learning methods for buildings with dissimilar styles. On Inria aerial dataset, Table 7 shows that our method outperforms other methods on all object-level evaluation metrics (Lin et al., 2014). The mask AP and AR increased by at least 4.1% and 4.6%, respectively. AP at different scales (AP_S, AP_M and AP_L) increased by at least 3.3%, 1.7% and 3.8%, respectively. These performances on the above two datasets forcefully validates that our method is more adaptable to the abundant changes in building appearance and surrounding

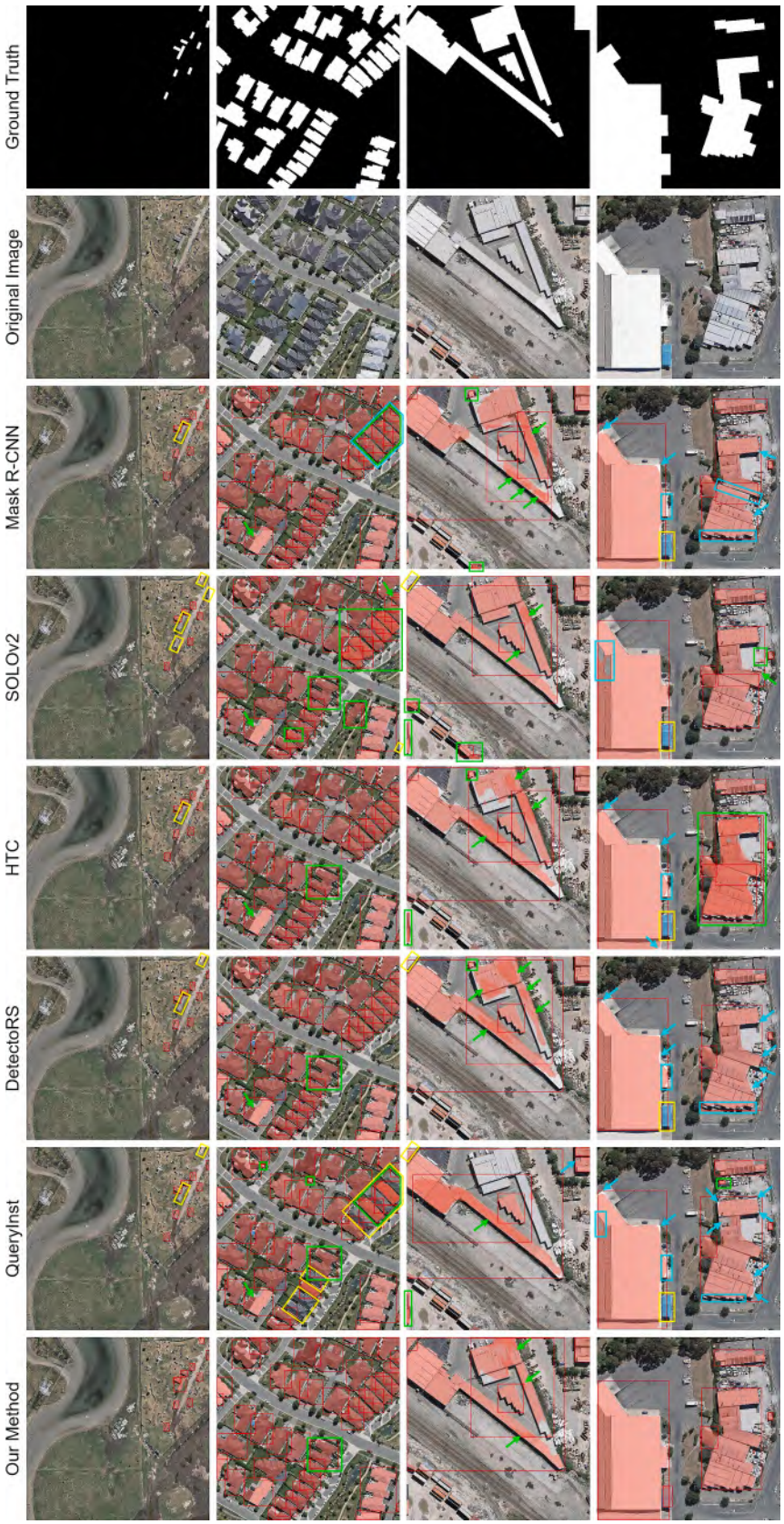

Fig. 7. Visual comparison of buildings extracted on WHU aerial dataset.

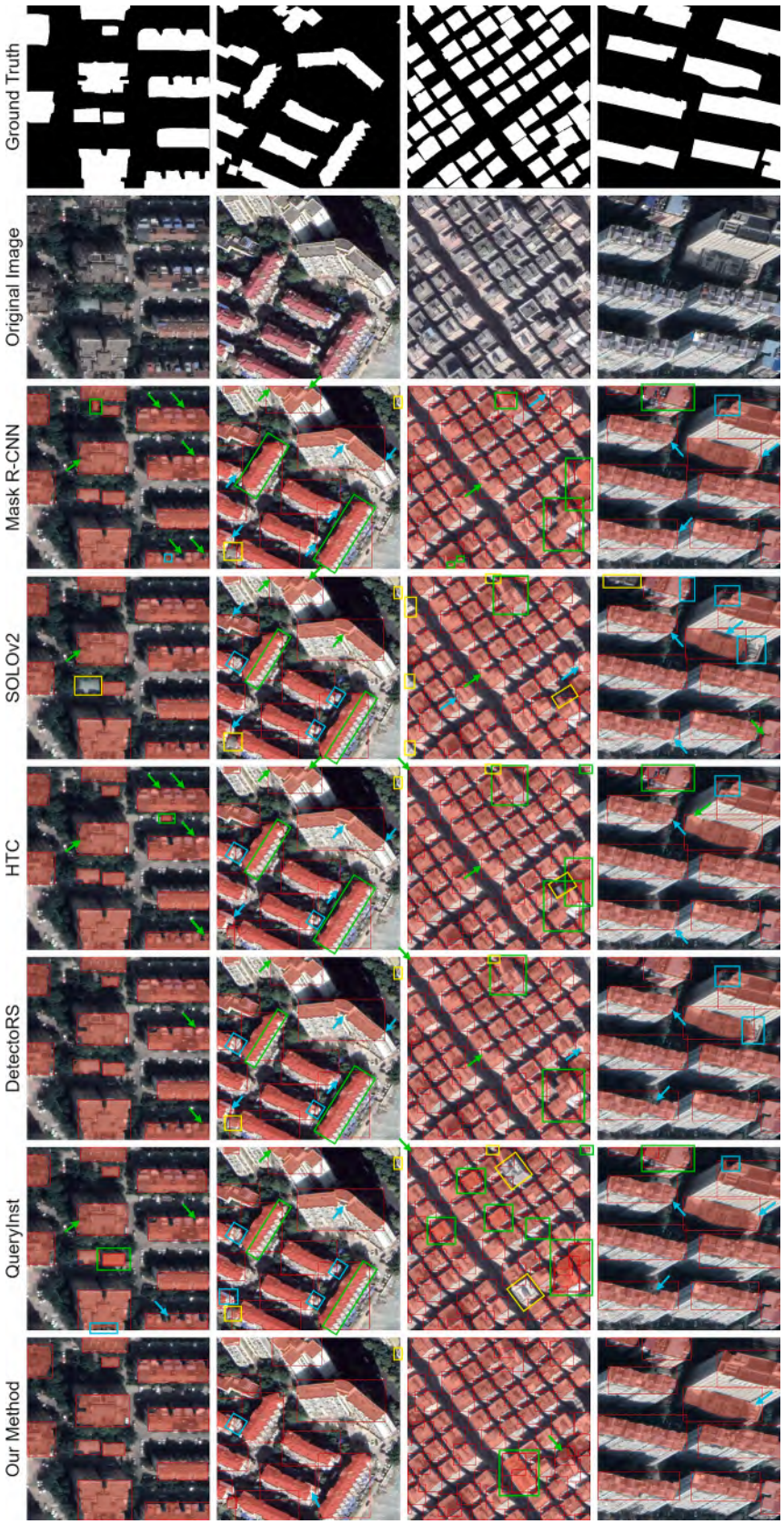

Fig. 8. Visual comparison of buildings with different data sources but similar styles extracted on China city satellite dataset.

Fig. 9. Visual comparison of buildings with different data sources and dissimilar styles extracted on Inria aerial dataset.

Table 8
Effects of components on complexity and memory overhead.

DA	FE	GA	IM	Input shape	Flops (G)	Params (M)	Mem (GB)
				512 × 512	134.20	**62.74**	**4.4**
√				512 × 512	134.20	**62.74**	**4.4**
√	√			512 × 512	136.80	69.29	5.3
√	√	√		512 × 512	136.74	69.88	4.6
√	√	√	√	512 × 512	**87.94**	81.66	4.7

environment.

4.4.2. Qualitative results

Figs. 7–9 visualize the qualitative evaluation results of our method compared with several classical, state-of-the-art instance segmentation methods.

Columns 1 and 2 in Fig. 7 contain some sparsely or densely distributed small and medium-sized buildings whose colors and textures are confused with their surrounding environment. As expected in Section 3.1, our method outperforms by capturing multi-scale features more granularly and adaptively expanding richer receptive field combinations. Other methods more or less miss some small-sized buildings, while SOLOv2 (Wang et al., 2020) and QueryInst (Fang et al., 2021b) even confuse or miss many medium-sized buildings. Columns 3 and 4 in Fig. 7 contain some large-sized buildings or building groups with large aspect ratios and unique appearances. In the face of large aspect ratios, our method effectively improves omission, semantic confusion and semantic lack by expanding deformation modeling, aggregating more location information from the bottom feature map as described in Sections 3.1, and learning aspect ratio as described in Sections 3.2. In the face of large scale rooftops with atypical and irregular geometric shapes, various comparison methods either miss the partial interior region or excessively smooth the mask contour as estimated in Section 3.3, whereas our method benefits from the iterative subdivision algorithm to completely and accurately output the masks of large-sized buildings or buildings groups. Figs. 8 and 9 contain some buildings with similar or dissimilar styles from several data sources. In contrast, our method is more discriminative for color, texture, shadow and other features, and more adaptable to larger changes of buildings in shape, scale, distribution and so on under the condition of few-sample training. Notably: as shown in Fig. 8, column 2 and Fig. 9, column 3, except for our method performs well, the other methods can neither accurately extract the abundant (jagged) contour details nor perfectly fit the sharp corners and regular edges of large rooftops; as shown in Fig. 9, column 1, all methods tend to confuse buildings with their surroundings when confronted with buildings occluded by lush vegetation, thus suffering from large-area semantic confusion or lack. In response to the above technical defect, we plan to model each region of interest (RoI) separately in the future by

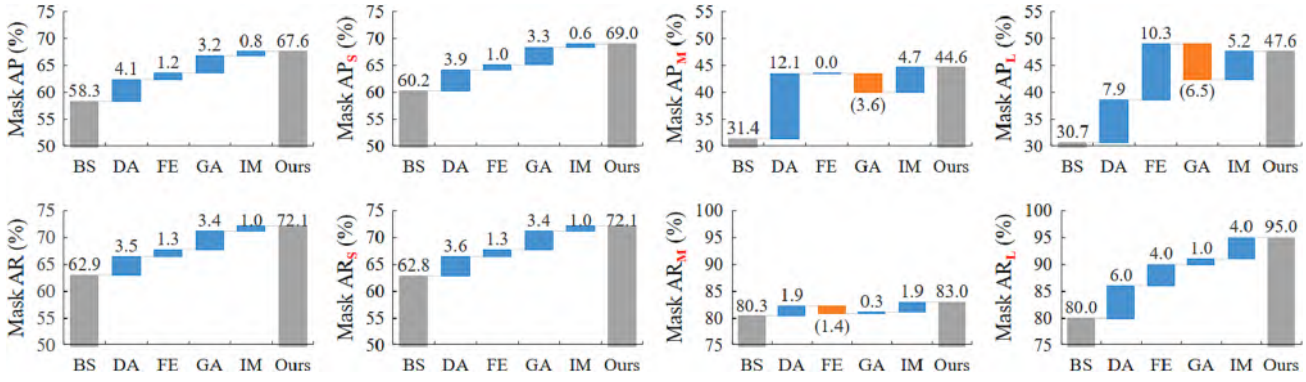

Fig. 10. Waterfall plots analyze the effects of components on accuracy and completeness. ■: Increase. ■: Decrease. ■: Summary.

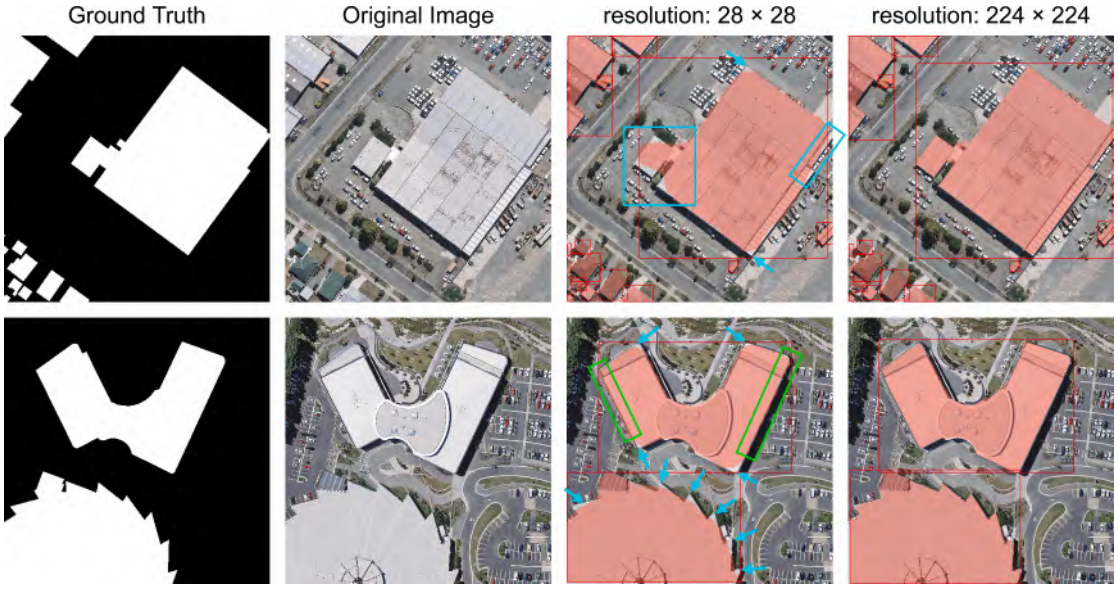

Fig. 11. Visual comparison of mask contours with different output resolutions (28 × 28 *vs.* 224 × 224).

Table 9
Parameter analysis of IM module. k: oversampling ratio. β: important sampling ratio.

(a) Point selection for training.				(b) Iterative subdivision steps.			(c) Point selection for inference.		
Selection strategy	k	β	AP	Out resolution	Selected points	AP	Out resolution	Selected points	AP
uniform	1	0.0	67.0	56 × 56	28^2	67.3	224 × 224	14^2	67.3
unbiased	3	0.5	67.3	112 × 112	28^2	67.5	224 × 224	28^2	**67.6**
mildly biased	3	0.75	**67.6**	224 × 224	28^2	**67.6**	224 × 224	56^2	67.4
heavily biased	9	1.0	67.5	448 × 448	28^2	67.4	224 × 224	112^2	**67.6**

two GCNs explicitly (Ke et al., 2021) to naturally decouple the occluding objects (or background regions) and the occluded buildings into two disjoint graph spaces, and to consider the influence of mutual occlusion in the mask regression process.

4.5. Ablation studies

With Mask R-CNN (ResNet-101-FPN) as the baseline (denoted as BS), we conducted ablation studies on WHU aerial dataset. Table 8 and Fig. 10 visually illustrate the respective effects of data augmentation (DA), feature extraction (FE), guided anchoring (GA), and iterative mask (IM) on our method.

DA increases mask AP and AR by 4.1% and 3.5%, respectively. Pixel-level augmentations, as exhibited in Fig. 6, make the model more robust against corresponding digital corruptions such as noise, contrast. Spatial-level augmentations improve the scale distribution imbalance while increasing training samples. Among them, the mask AP_M and AP_L of medium and large-sized buildings with fewer original samples increased conspicuously by 12.1% and 7.9%, respectively.

FE increases mask AP and AR by 1.2% and 1.3%, respectively, mainly due to the fact that adaptive residual blocks divide the feature maps and receptive fields equally and then recombine them hierarchically, which is beneficial to increase the mask AP_S and AR_S by 1.0% and 1.3%, respectively. Furthermore, feature pyramid significantly increases mask AP_L and AR_L by 10.3% and 4.0% respectively by aggregating more location information from the bottom feature map.

GA module increases mask AP and AR separately by 3.2% and 3.4% with 13.2% memory saving, which illustrates that matching the location and shape of anchors to ground truth rooftops is conducive to generate more accurate proposals, and controlling the number of anchors can affect training overhead. The training samples of WHU aerial dataset are mostly small-sized buildings, which makes the dynamic anchors generated by GA module are closer to the scale distribution of small-sized buildings, and indirectly leads to a significant decrease of mask AP_M and AP_L by 3.6% and 6.5%, respectively.

IM module increases mask AP and AR by 0.8% and 1.0% respectively with 35.7% less computation (Flops) and only 2.2% more memory. AP_S, AP_M and AP_L increased by 0.6%, 4.7% and 5.2% respectively, AR_S, AR_M and AR_L increased by 1.0%, 1.9% and 4.0% respectively, which indicates that the larger the building scale, the more obvious the performance of IM module. Intuitively, Fig. 11 demonstrates that high-resolution output is more effective for large-sized buildings, aligning the mask contour as closely as possible to the sharp corners and regular boundaries of their rooftops. Parameter analysis of IM module in Table 9: (a) Appropriately oversampling ($2 < k < 5$) and mildly biased ($0.5 < \beta < 1$) towards uncertain regions (rooftop edges) can obtain the best performance; (b) Higher output resolution obtained with iterative subdivision steps can continuously improve prediction results until 224 × 224; (c) Selecting 28^2 points in each iterative subdivision step saturates the mask AP.

5. Conclusion

In this paper, we propose a lightweight method capable of adaptively optimizing the mask contour to automatically extract building instances. The design of each component aims to be adaptive to the extraction task while minimizing the computational and memory overhead. Comprehensive experiments on WHU aerial, China city satellite and Inria aerial building datasets demonstrate that compared to classical, state-of-the-art instance segmentation methods, our method not only improves accuracy, robustness and generalizability, economizes significant computational and memory overheads, but also the extracted masks better fit the sharp corners and regular edges of rooftop contours. In future work, we will focus on decoupling the occluding objects and the occluded buildings, describing the interior and exterior contours of rooftops separately, as well as investigating gain-robust technology suitable for remote sensing scenes.

CRediT authorship contribution statement

Xiaoxue Liu: Conceptualization, Methodology, Software, Data curation, Writing – original draft, Writing – review & editing. **Yiping Chen:** Conceptualization, Methodology, Writing – review & editing, Project administration, Funding acquisition. **Cheng Wang:** Resources, Writing – review & editing. **Kun Tan:** Resources, Writing – review & editing. **Jonathan Li:** Conceptualization, Supervision, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The experimental data were obtained from three open source datasets, and the download sources have been indicated in the manuscript.

References

Abdollahi, A., Pradhan, B., Gite, S., Alamri, A., 2020. Building footprint extraction from high resolution aerial images using generative adversarial network (GAN) architecture. IEEE Access 8, 209517–209527. https://doi.org/10.1109/ACCESS.2020.3038225.

Alshehhi, R., Marpu, P.R., Woon, W.L., Mura, M.D., 2017. Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks. ISPRS J. Photogramm. Remote Sens. 130, 139–149. https://doi.org/10.1016/j.isprsjprs.2017.05.002.

Azulay, A., Weiss, Y., 2018. Why do deep convolutional networks generalize so poorly to small image transformations?. arXiv preprint arXiv:1805.12177v1.

Bi, Q., Qin, K., Zhang, H., Zhang, Y., Li, Z., Xu, K., 2019. A multi-scale filtering building index for building extraction in very high-resolution satellite imagery. Remote Sens. 11 (5), 482. https://doi.org/10.3390/rs11050482.

Chaudhuri, D., Kushwaha, N.K., Samal, A., Agarwal, R.C., 2016. Automatic building detection from high-resolution satellite images based on morphology and internal gray variance. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. 9 (5), 1767–1779. https://doi.org/10.1109/JSTARS.2015.2425655.

Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., Loy, C., Lin, D., 2019. Hybrid task cascade for instance segmentation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog. 4969–4978. https://doi.org/10.1109/CVPR.2019.00511.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y., 2017. Deformable convolutional networks. In: Proc. IEEE Int. Conf. Comput. Vis. 764–773. https://doi.org/10.1109/ICCV.2017.89.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Li, F.-F., 2009. ImageNet: a large-scale hierarchical image database. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog. 248–255. https://doi.org/10.1109/CVPR.2009.5206848.

Fang, F., Wu, K., Liu, Y., Li, S., Wan, B., Chen, Y., Zheng, D., 2021a. A coarse-to-fine contour optimization network for extracting building instances from high-resolution remote sensing imagery. Remote Sens. 13 (19), 3814. https://doi.org/10.3390/rs13193814.

Fang, Y., Yang, S., Wang, X., Li, Y., Fang, C., Shan, Y., Feng, B., Liu, W., 2021b. Instances as Queries. In: Proc. IEEE Int. Conf. Comput. Vis. 6890–6899. https://doi.org/10.1109/ICCV48922.2021.00683.

Gao, S., Cheng, M., Zhao, K., Zhang, X., Yang, M., Torr, P., 2021. Res2Net: a new multi-scale backbone architecture. IEEE Trans. Pattern Anal. Mach. Intell. 43 (2), 652–662. https://doi.org/10.1109/TPAMI.2019.2938758.

Geirhos, R., Temme, C.R.M., Rauber, J., Schütt, H.H., Bethge, M., Wichmann, F.A., 2018. Generalisation in humans and deep neural networks. In: in: Advances in Neural Information Processing Systems 31, pp. 7549–7561.

Ghanea, M., Moallem, P., Momeni, M., 2016. Building extraction from high-resolution satellite images in urban areas: recent methods and strategies against significant challenges. Remote Sens. 37 (21), 5234–5248. https://doi.org/10.1080/01431161.2016.1230287.

Girard, N., Smirnov, D., Solomon, J., Tarabalka, Y., 2021. Polygonal building extraction by frame field learning. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog. 5887–5896. https://doi.org/10.1109/CVPR46437.2021.00583.

Guo, H., Du, B., Zhang, L., Su, X., 2022. A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery. ISPRS J. Photogramm. Remote Sens. 183, 240–252. https://doi.org/10.1016/j.isprsjprs.2021.11.005.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN. In: Proc. IEEE Int. Conf. Comput. Vis., pp. 2980-2988, doi: 10.1109/ICCV.2017.322.

Hendrycks, D., Dietterich, T., 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. arXiv preprint arXiv:1903.12261.

Huang, Z., Cheng, G., Wang, H., Li, H., Shi, L., Pan, C., 2016. Building extraction from multi-source remote sensing images via deep deconvolution neural networks. In: Proc. IEEE Int. Geosci. Remote Sens. Sympos. 1835–1838. https://doi.org/10.1109/IGARSS.2016.7729471.

Huang, W., Tang, H., Xu, P., 2022. OEC-RNN: object-oriented delineation of rooftops with edges and corners using the recurrent neural network from the aerial images. IEEE Trans. Geosci. Remote Sens. 60, 1–12. https://doi.org/10.1109/TGRS.2021.3076098.

Ji, S., Wei, S., Lu, M., 2019. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Trans. Geosci. Remote Sens. 57 (1), 574–586. https://doi.org/10.1109/TGRS.2018.2858817.

Ke, L., Tai, Y., Tang, C., 2021. Deep occlusion-aware instance segmentation with overlapping BiLayers. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog. 4018–4027. https://doi.org/10.1109/CVPR46437.2021.0040.

Kirillov, A., Wu, Y., He, K., Girshick, R., 2020. PointRend: image segmentation as rendering. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog. 9796–9805. https://doi.org/10.1109/CVPR42600.2020.00982.

Kotaridis, I., Lazaridou, M., 2021. Remote sensing image segmentation advances: a meta-analysis. ISPRS J. Photogramm. Remote Sens. 173, 309–322. https://doi.org/10.1016/j.isprsjprs.2021.01.020.

Li, Q., Mou, L., Hua, Y., Sun, Y., Jin, P., Shi, Y., Zhu, X., 2020. Instance segmentation of buildings using keypoints. In: Proc. IEEE Int. Geosci. Remote Sens. Sympos. 1452–1455. https://doi.org/10.1109/IGARSS39084.2020.9324457.

Li, W., Sun, K., Zhao, H., Li, W., Wei, J., Gao, S., 2022. Extracting buildings from high-resolution remote sensing images by deep ConvNets equipped with structural-cue-guided feature alignment. Int. J. Appl. Earth Observation and Geoinformation. 113, 102970 https://doi.org/10.1016/j.jag.2022.102970.

Li, Z., Wegner, J.D., Lucchi, A., 2019. Topological map extraction from overhead images. In: Proc. IEEE Int. Conf. Comput. Vis. 1715–1724. https://doi.org/10.1109/ICCV.2019.00180.

Li, Z., Xin, Q., 2021. Corner-guided building polygon construction from aerial images using deep multitask learning. In: Proc. IEEE Int. Geosci. Remote Sens. Sympos. 4043–4046. https://doi.org/10.1109/IGARSS47720.2021.9554624.

Li, E., Xu, S., Meng, W., Zhang, X., 2017. Building extraction from remotely sensed images by integrating saliency cue. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. 10 (3), 906–919. https://doi.org/10.1109/JSTARS.2016.2603184.

Lin, T., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal Loss for dense object detection. In: Proc. IEEE Int. Conf. Comput. Vis. 2999–3007. https://doi.org/10.1109/ICCV.2017.324.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., Zitnick, C.L., 2014. Microsoft COCO: common objects in context. Proc. Eur. Conf. on Comput. Vis. 8693, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48.

Liu, X., Chen, Y., Wei, M., Wang, C., Goncalves, W.N., Marcato, J., Li, J., 2021a. Building instance extraction method based on improved hybrid task cascade. IEEE Geosci. Remote Sens. Lett. 19, 1–5. https://doi.org/10.1109/LGRS.2021.3060960.

Liu, Y., Chen, D., Ma, A., Zhong, Y., Fang, F., Xu, K., 2021b. Multiscale U-Shaped CNN building instance extraction framework with edge constraint for high-spatial-resolution remote sensing imagery. IEEE Trans. Geosci. Remote Sens. 59 (7), 6106–6120. https://doi.org/10.1109/TGRS.2020.3022410.

Liu, S., Qi, L., Qin, H., Shi, J., Jia, J., 2018. Path Aggregation Network for Instance Segmentation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog., pp. 8759-8768, doi: 10.1109/CVPR.2018.00913.

Liu, W., Yang, M., Xie, M., Guo, Z., Li, E., Zhang, L., Pei, T., Wang, D., 2019. Accurate building extraction from fused DSM and UAV images using a chain fully convolutional neural network. Remote Sens. 11 (24), 2912. https://doi.org/10.3390/rs11242912.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog., pp. 3431-3440, doi: 10.1109/CVPR.2015.7298965.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In: Proc. IEEE Int. Geosci. Remote Sens. Sympos., pp. 3226-3229, doi: 10.1109/IGARSS.2017.8127684.

Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A. S., Bethge, M., Brendel, W., 2019. Benchmarking Robustness in Object Detection: Autonomous Driving when Winter is Coming. arXiv preprint arXiv:1907.07484.

Qiao, S., Chen, L.C., Yuille, A., 2021. DetectoRS: Detecting Objects with Recursive Feature Pyramid and Switchable Atrous Convolution. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog., pp. 10208-10219, doi: 10.1109/CVPR46437.2021.01008.

Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39 (6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031.

Shrestha, S., Vanneschi, L., 2018. Improved fully convolutional network with conditional random fields for building extraction. Remote Sens. 10 (7), 1135. https://doi.org/10.3390/rs10071135.

Sun, Y., Zhang, X., Zhao, X., Xin, Q., 2018. Extracting building boundaries from high resolution optical images and LiDAR data by integrating the convolutional neural network and the active contour model. Remote Sens. 10 (9), 1459. https://doi.org/10.3390/rs10091459.

Turker, M., Koc-San, D., 2015. Building extraction from high-resolution optical spaceborne images using the integration of support vector machine (SVM) classification, Hough transformation and perceptual grouping. Int. J. Appl. Earth Observation and Geoinformation. 34, 58–69. https://doi.org/10.1016/j.jag.2014.06.016.

Tychsen-Smith, L. and Petersson, L., 2018. Improving Object Localization with Fitness NMS and Bounded IoU Loss. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog., pp. 6877-6885, doi: 10.1109/CVPR.2018.00719.

Wang, J., Chen, K., Yang, S., Loy, C.C., Lin, D., 2019. Region Proposal by Guided Anchoring. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog., pp. 2960-2969, doi: 10.1109/CVPR.2019.00308.

Wang, X., Zhang, R., Kong, T., Li, L., Shen, C., 2020. SOLOv2: Dynamic and Fast Instance Segmentation. In: Proc. Adv. Neural Inf. Process. Syst., pp. 17721-17732.

Wei, S., Ji, S., 2022. Graph Convolutional Networks for the Automated Production of Building Vector Maps From Aerial Images. IEEE Trans. Geosci. Remote Sens. 60, 1–11. https://doi.org/10.1109/TGRS.2021.3060770.

Wen, Q., Jiang, K., Wang, W., Liu, Q., Guo, Q., Li, L., Wang, P., 2019. Automatic building extraction from Google Earth images under complex backgrounds based on deep instance segmentation network. Sensors. 19 (2), 333. https://doi.org/10.3390/s19020333.

Wu, T., Hu, Y., Peng, L., Chen, R., 2020. Improved anchor-free instance segmentation for building extraction from high-resolution remote sensing images. Remote Sens. 12 (18), 2910. https://doi.org/10.3390/rs12182910.

Xu, L., Li, Y., Xu, J., Guo, L., 2021. Gated Spatial Memory and Centroid-Aware Network for Building Instance Extraction. IEEE Trans. Geosci. Remote Sens. 60, 1–14. https://doi.org/10.1109/TGRS.2021.3073164.

Xu, S., Pan, X., Li, E., Wu, B., Bu, S., Dong, W., Xiang, S., Zhang, X., 2018. Automatic Building Rooftop Extraction From Aerial Images via Hierarchical RGB-D Priors. IEEE Trans. Geosci. Remote Sens. 56 (12), 7369–7387. https://doi.org/10.1109/TGRS.2018.2850972.

Zhang, Q., Huang, X., Zhang, G., 2016. A Morphological Building Detection Framework for High-Resolution Optical Imagery Over Urban Areas. IEEE Geosci. Remote Sens. Lett. 13 (9), 1388–1392. https://doi.org/10.1109/LGRS.2016.2590481.

Zhang, L., Wu, J., Fan, Y., Gao, H., Shao, Y., 2020. An efficient building extraction method from high spatial resolution remote sensing images based on improved Mask R-CNN. Sensors. 20 (5), 1465. https://doi.org/10.3390/s20051465.

Zhao, K., Kang, J., Jung, J., Sohn, G., 2018. Building Extraction from Satellite Images Using Mask R-CNN with Building Boundary Regularization. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops, pp. 242-2424, doi: 10.1109/CVPRW.2018.00045.

Zhu, X., Hu, H., Lin, S., Dai, J., 2019. Deformable ConvNets V2: More Deformable, Better Results. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog., pp. 9300-9308, doi: 10.1109/CVPR.2019.00953.

Zhu, Q., Liao, C., Hu, H., Mei, X., Li, H., 2021. MAP-Net: Multiple Attending Path Neural Network for Building Footprint Extraction From Remote Sensed Imagery. IEEE Trans. Geosci. Remote Sens. 59 (7), 6169–6181. https://doi.org/10.1109/TGRS.2020.3026051.

Zorzi, S., Bazrafkan, S., Habenschuss, S., Fraundorfer, F., 2022. PolyWorld: Polygonal Building Extraction with Graph Neural Networks in Satellite Images. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog., pp. 1938-1947, doi: 10.1109/CVPR52688.2022.00189.