Building Instance Extraction Method Based on Improved Hybrid Task Cascade

Xiaoxue Liu, Yiping Chen, *Senior Member, IEEE*, Mingqiang Wei, Cheng Wang, *Senior Member, IEEE*, Wesley Nunes Gonçalves, *Member, IEEE*, José Marcato Junior, *Member, IEEE*, and Jonathan Li, *Senior Member, IEEE*

***Abstract*— Automatic building extraction from remote sensing imagery is crucial to urban construction and management. To address the main challenges of diverse building scale and appearance, this letter proposes an automatic building instance extraction method based on an improved hybrid task cascade (HTC). Our method consists of three components by obtaining high-resolution representation, defining guided anchor, and forming focal loss to boost the adaptability of automatic building instance extraction. Comprehensive experimental results on WHU aerial building data set demonstrated that compared with the mainstream Mask R-CNN method, our method increased AP and AR in bounding box branch and mask branch by 9.8%–6.5% and 10.7%–8.0% respectively, especially AP_S and AP_L in the two branches by 10.1%–6.9% and 3.4%–2.4%, respectively. We evaluated the effectiveness and complexity of these components separately and discussed the universality and practicability of deep learning method in automatic building extraction.**

***Index Terms*— Aerial imagery, building extraction, deep learning, hybrid task cascade, instance segmentation.**

I. Introduction

AUTOMATIC extraction of large-scale, high-precision, and periodic building rooftop information from remote sensing imagery is an urgent need for urban planning, disaster response, environmental monitoring, and other application research. In view of the challenges of different remote sensing imaging principles, abundant types and details of ground objects, complex scene structure and distribution, diverse building scale and appearance, how to accurately and efficiently automate building extraction have always been a frontier topic in the field of remote sensing image analysis.

Traditional methods require experts to design appropriate building feature representations (edge and shadow [1], spectrum and context [2], color and shape [3], semantics and height [4], etc.) based on experience and then combine with corresponding algorithms to identify and extract pixel-level buildings. Since many complicated and changeable factors (light, atmospheric condition, season, sensor quality, building scale and appearance, environment, etc.) may affect empirical design features, traditional methods usually can only deal with specific issues with specific data, and their accuracy and efficiency are difficult to meet the needs of different tasks and practical applications.

In recent years, deep learning methods have surpassed and gradually replaced traditional empirical design feature methods by virtue of the ability of convolutional neural network (CNN) to automatically learn multilevel feature representations [5], [6]. A large number of deep learning literature studies are devoted to semantic segmentation methods for extracting pixel-level buildings. These methods mainly improve multiscale inference, enrich context information, alleviate data class imbalance, optimize building boundaries, eliminate salt and pepper noise and fill holes, fuse multisource data, to make fully convolutional network (FCN) [7], [8] (including U-Net [9], [10], DeconvNet [11], SegNet [12], and other variants) models more suitable for complex remote sensing image background and small-sized building targets. A small number of deep learning literature studies focus on emerging instance segmentation methods for obtaining object-level buildings (i.e. building instances). Compared with the semantic segmentation that only judges the semantic category of each pixel, the instance segmentation that finely distinguishes each building (including location, contour, area, and other information) has more application value and urgently needs technical expansion. Related instance studies based on the Mask R-CNN [13] model mainly regularize mask contour [14], ameliorate mask of building edge [15], expand mask receptive field [16], design and adjust the rotation angle and aspect ratio of anchors [16], but these methods still have obvious limitations in adapting to the extreme scale and heterogeneous appearance of buildings. The application and development of existing deep learning methods are generally restricted by the limited amount of labeled remote sensing data. Semi-supervised learning (e.g., MixMatch [17]), unsupervised learning (e.g., MoCo [18]), data synthesis (e.g., GAN [19]), and other research to mitigate the data reliance of

Manuscript received November 8, 2020; revised January 28, 2021; accepted February 17, 2021. Date of publication March 12, 2021; date of current version December 16, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 41871380 and Grant U1605254. *(Corresponding author: Yiping Chen.)*

Xiaoxue Liu, Yiping Chen, and Cheng Wang are with the Fujian Key Laboratory of Sensing and Computing for Smart Cities, School of Informatics, Xiamen University, Xiamen 361005, China (e-mail: liuxiaoxue@stu.xmu.edu.cn; chenyiping@xmu.edu.cn; cwang@xmu.edu.cn).

Mingqiang Wei is with the School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China (e-mail: mqwei@nuaa.edu.cn).

Wesley Nunes Gonçalves and José Marcato Junior are with the Faculty of Engineering, Architecture and Urbanism and Geography, Federal University of Mato Grosso do Sul, Campo Grande 79070-900, Brazil (e-mail: wesley.goncalves@ufms.br; jose.marcato@ufms.br).

Jonathan Li is with the Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L 3G1, Canada, and also with the Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: junli@uwaterloo.ca).

Digital Object Identifier 10.1109/LGRS.2021.3060960

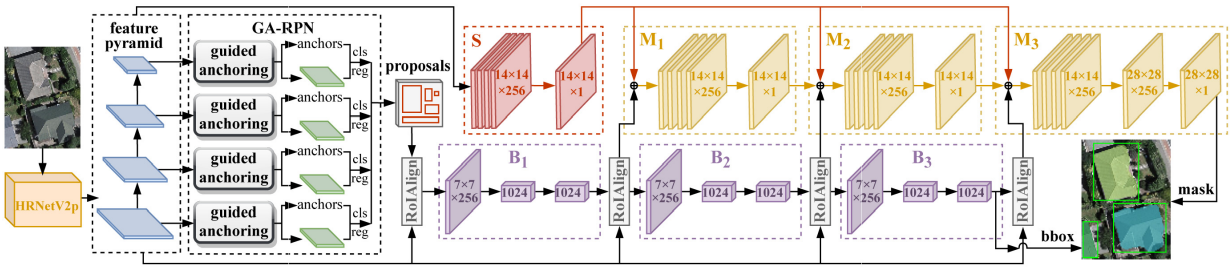

Fig. 1. Model architecture of our method.

CNN or reduce the cost of data labeling will be conducive to break through the above dilemma.

To address the main challenges of diverse building scale and appearance, this letter proposes an automatic building instance extraction method from high-resolution aerial imagery based on improved hybrid task cascade [20] architecture. Our main contributions are summarized as follows: 1) leverage HRNetV2p [21] instead of ResNet and FPN as backbone network to obtain high-resolution representation; 2) define a guide anchoring [22] mechanism to generate dynamically predicted anchors; and 3) form the classification loss of guided anchoring region proposal network (GA-RPN) [22] from cross entropy (CE) to focal loss (FL) [23] to alleviate data class imbalance and pay attention to the difference in the difficulty of sample classification. Comprehensive experiments based on WHU aerial building data set [6] show that our method outperforms state-of-the-art instance segmentation methods for accuracy and smaller module scale in the complex urban environment.

II. Method

The model architecture of our method is shown in Fig. 1: First, input aerial images into HRNetV2p [21] to extract rooftop features and construct a feature pyramid. Second, guided anchoring region proposal network (GA-RPN) [22] leverages semantic features in the feature pyramid to find possible image regions with buildings and generate anchors according to the location and then further classifies (cls) and regresses (reg) anchors to screen out proposals (i.e. candidate building bounding boxes). Third, these proposals are dynamically mapped to the corresponding feature maps to extract regions of interest (RoIs) of various sizes. Each RoI is pooled into a fixed size feature map (e.g., 7×7 or 14×14) by RoIAlign [13] layer. Finally, fully connected bounding box branch (B) and fully convolutional mask branch (M) are interleaved cascade in three stages to predict the bounding box (bbox) and mask of each RoI. The regression results of B_{i-1} are mapped to the feature pyramid to regenerate RoIs for training of B_i and M_{i-1}. The mask features of M_{i-1} are embedded in M_i and then fused with the backbone features by elementwise sum. Fully convolutional semantic segmentation branch (S) predicts the semantic category of all pixels in the whole image. The semantic segmentation features with encoded spatial contexts are fused with the mask features of each stage by element-wise sum. In summary, this model architecture integrates the features of each branch in each stage to gradually ameliorate bbox regression and mask prediction.

A. High-Resolution Representation

The architecture of HRNetV2p [21] is shown in Fig. 2: 1) One set of high-resolution convolution and three sets of low-resolution convolution are connected in four stages in parallel to maintain high-resolution representation and repeatedly fuse high-to-low resolution representations; 2) low-resolution representations are upsampled to high-resolution representations by bilinear interpolation, and four representations are then fused by 1-strided 1×1 convolution; and 3) hybrid representation is downsampled to multiple levels by average pooling to construct a feature pyramid. Our method sets the number of channels of high-resolution convolution to 32 (i.e. HRNetV2p-W32 [21]) and the 2nd, 3rd, and 4th stages to 1, 4, and 3 repeated multiresolution convolution blocks, respectively.

B. Guided Anchor

In building extraction task, the disadvantages of the traditional sliding window mechanism are as follows: dense anchors evenly distributed in the background region waste many computing resources; manually predefined anchor shapes (aspect ratio and size) are not necessarily suitable for buildings with extreme aspect ratio or size. Our method defines a guided anchoring [22] mechanism to generate sparse and variable-shaped anchors to solve the above problems.

The joint conditional probability formula of guided anchor is defined as

$$p(x, y, w, h \mid I) = p(x, y \mid I)p(w, h \mid x, y, I) \tag{1}$$

where I is given image feature, (x, y) is the anchor center location, (w, h) is the anchor shape (width, height). For each feature map F_I output in the feature pyramid, the guided anchoring process based on above formula principle is shown in Fig. 3: 1) N_L (a 1×1 convolution + sigmoid) and N_S (a 1×1 convolution + nonlinear transformation of w and h) branches in anchor generation module output single-channel and dual-channel maps with the same resolution as F_I, respectively, representing the center location probability and the optimal shape with the highest overlap with the nearest ground truth bounding box; 2) N_T (a 1×1 convolution + a 3×3 deformable convolution) branch in feature adaptation module applies deformable convolution according to each position offset to make the feature map perceive and adapt various anchor shapes ($F_I \longrightarrow F_I'$) for subsequent classification and regression of anchors. Our method shares anchor generation parameters among all involved feature levels and only uses 300 proposals.

C. Focal Loss

Our improved GA-RPN [22] loss ($\mathcal{L}_{\text{ga-rpn}}$) is defined as

$$\mathcal{L}_{\text{ga-rpn}} = \lambda_1 \mathcal{L}_{\text{loc}} + \lambda_2 \mathcal{L}_{\text{shape}} + \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{reg}} \tag{2}$$

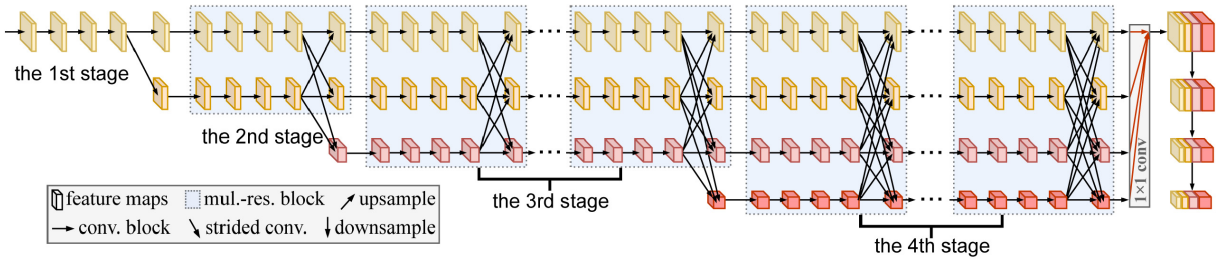

Fig. 2. Architecture of HRNetV2p [21]. The 1st stage is high-resolution convolutions. The 2nd, 3rd and 4th stages are composed of repeated multiresolution convolution blocks. The number of channels and resolutions of the four types of convolutions increase by two times and decrease by 0.5 times in turn.

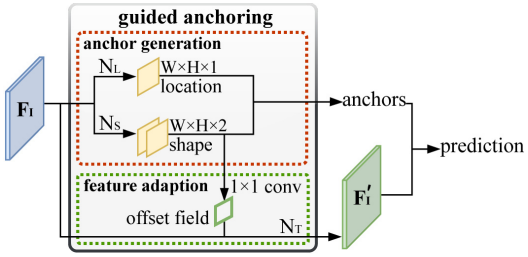

Fig. 3. Architecture of guided anchoring module.

$$\mathcal{L}_{\text{shape}} = \mathcal{L}_1\left(1 - \min\left(\frac{w}{w_g}, \frac{w_g}{w}\right)\right) + \mathcal{L}_1\left(1 - \min\left(\frac{h}{h_g}, \frac{h_g}{h}\right)\right) \tag{3}$$

where λ_1 and λ_2 are parameters to balance the location loss ($\mathcal{L}_{\text{loc}}$) and shape loss ($\mathcal{L}_{\text{shape}}$) of anchors; $\mathcal{L}_{\text{loc}}$ and classification loss ($\mathcal{L}_{\text{cls}}$) are FL [23]; regression loss ($\mathcal{L}_{\text{reg}}$) and $\mathcal{L}_1$ are smooth L1 loss; (w, h) and (w_g, h_g) represent the predicted anchor shape and the shape of corresponding ground truth bounding box.

In building extraction task, the pixels of background class are usually far more than those of building class, and there are usually significant differences in the number of building samples with different scales and appearances, especially for buildings with extreme scales and heterogeneous appearance. Our method forms FL [23] based on CE as the classification loss in GA-RPN [22] to alleviate extreme imbalance between building and background classes and reduce the weight of easy samples to make the model focus more on few and hard samples during training.

FL [23] is defined as

$$p_t = \begin{cases} p, & \text{if } y = 1 \\ 1 - p, & y = -1 \end{cases} \tag{4}$$

$$\alpha_t = \begin{cases} \alpha, & \text{if } y = 1 \\ 1 - \alpha, & y = -1 \end{cases} \tag{5}$$

$$\text{CE}(p_t) = -\log(p_t) \tag{6}$$

$$\text{FL}(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \tag{7}$$

where $y \in \{\pm 1\}$ is the ground truth class (1: building and -1: background); $p \in [0, 1]$ is the estimated probability for the class with label $y = 1$; $\alpha \in [0, 1]$ is a weight factor to balance positive/negative samples; $(1 - p_t)^\gamma$ is a modulation factor to reduce the weight of easy samples (as $p_t \to 1$, $(1 - p_t)^\gamma \to 0$), and $\gamma \geq 0$ is a focus parameter used to smooth the weight adjustment process. Our method sets $\lambda_1 = 1$, $\lambda_2 = 0.1$, $\alpha = 0.25$ and $\gamma = 2$ for the best performance benefit.

TABLE I
COMPARISON BASED ON PIXEL-LEVEL EVALUATION METRICS

Method	Backbone	IoU (%)	Precision (%)	Recall (%)
FCN [7]	ResNet	85.2	95.2	89.0
U-Net [9]	ResNet	85.9	90.6	94.2
Our method	HRNetV2p-W32	**92.1**	**97.3**	**94.5**

III. EXPERIMENTS AND RESULTS

A. Data Set

The data set used in experiments comes from WHU aerial building dataset [6]. This data set crops an entire aerial image into 8188 high-resolution orthophoto tiles (512 × 512 pixels in size, 0.3 m in spatial resolution, RGB bands), and sets 5772 tiles (145 000 buildings) and 2416 tiles (42 000 buildings) as training samples and testing samples respectively. These samples cover a variety of building types located in different urban areas such as administrative, residential, commercial, industrial, and suburban areas. Our method converts the label data format from .TIF to .JSON and clears the samples without buildings to match various deep learning methods.

B. Implementation Details

Training and testing are based on PyTorch deep learning framework and NVIDIA Tesla V100 GPU (a GPU, 16 GB) hardware environment. Configurations and parameters mainly include: 1) pretrained weights based on COCO 2017 data set; 2) four images per GPU; 3) SGD optimizer, initial learning rate of 0.0025, momentum of 0.9, weight decay of 0.0001; and 4) restricted by GPU performance, these models only trained basic training schedule: 20 epochs.

C. Evaluation

1) Qualitative Evaluation: The qualitative evaluation results of several instance segmentation methods are compared, as shown in Fig. 4: 1) the image in row 1 contains buildings with obvious differences in aspect ratio. [13] and [20] repeatedly mistook a building for several buildings, and both mistook three small ground objects in the top right corner of the scene for buildings; 2) the image in row 2 contains buildings with obvious differences in size. [13] and [20] confused the edges

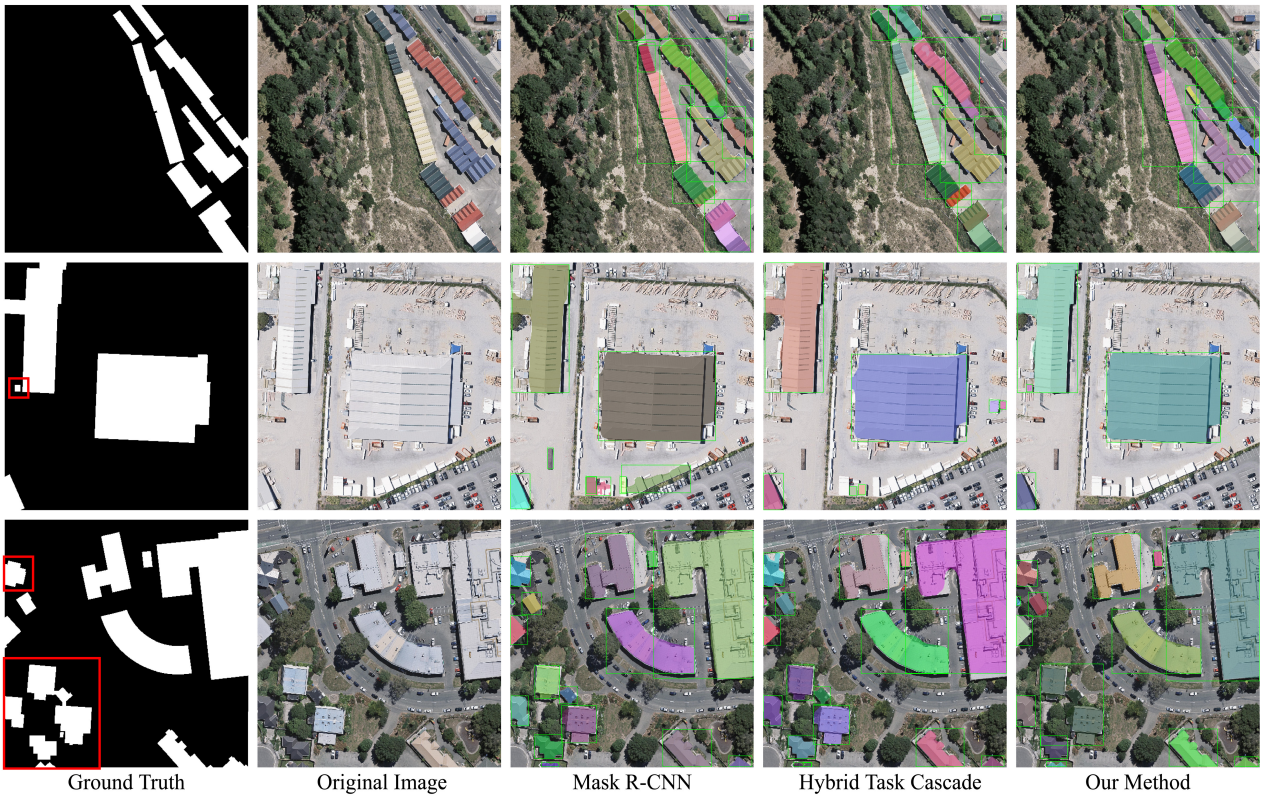

Fig. 4. Comparison of building instance extraction. The red boxes in the ground truth indicate some areas with major errors in the extraction results.

TABLE II
COMPARISON BASED ON COCO OBJECT-LEVEL EVALUATION METRICS

Method	Backbone	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 [13]	ResNet101-FPN	59.6	83.4	69.4	62.3	**36.1**	25.9	64.8	57.7	83.6	68.3	60.1	**36.1**	25.0	62.1
HTC [20]	ResNet101-FPN	61.9	84.2	71.5	64.8	23.5	24.2	67.2	58.6	84.5	69.4	61.1	24.0	20.9	63.4
Our method	HRNetV2p-W32	**69.4**	**89.8**	**78.2**	**72.4**	23.7	**29.3**	**75.5**	**64.2**	**89.8**	**74.9**	**67.0**	23.9	**27.4**	**70.1**

AP (the average precision at IoU=0.50:0.05:0.95) is primary metric. Subscript 50 and 75 refer to AP at different IoU thresholds. Subscript S, M and L refer to AP for different objects scales differentiated by COCO standard. AR is the average recall given 100 detections per image.

of two large-sized buildings with the surrounding environment to varying degrees, more or less mistook some ground objects for buildings, and both omitted a tiny building; 3) the image in row 3 contains buildings with obvious differences in size and appearance. The results of [13] and [20] showed: some edge areas of large-sized buildings obviously had no mask; a townhouse with garage was mistaken for three adjacent buildings. In addition, [20] mistook a building fragment (caused by image cropping) at the bottom of the scene for two adjacent buildings, and even confused a cross-shaped building with the environment. Our method accurately identified and completely extracted various buildings in these above-mentioned scenes.

2) Quantitative Evaluation: The quantitative evaluation results based on pixel level of our method are all superior to several semantic segmentation methods, as shown in Table I.

The quantitative evaluation results based on object-level of several instance segmentation methods are compared, as shown in Table II: 1) Compared with Mask R-CNN [13], our method significantly increased AP and AR in bounding box branch and mask branch by 9.8%–6.5% and 10.7%–8.0% respectively, especially AP_S and AP_L in the two branches by 10.1%–6.9% and 3.4%–2.4% respectively and 2) Compared with HTC [20], our method significantly increased AP and AR in bounding box branch and mask branch by 7.5%–5.6% and 8.3%–6.7%, respectively, especially AP_S and AP_L in the two branches by 7.6%–5.9% and 5.1%–6.5%, respectively.

The quantitative evaluation results of our components are compared with HTC [20], as shown in Table III: 1) *Effectiveness of high-resolution representation.* AP and AR in the two branches were significantly increased by 4.6%–4.0% and 4.5%–3.8% respectively, indicating that high-resolution representation can enhance the ability of feature expression, and cascade learning can double the performance benefit, and interleaved execution can balance the benefit difference between different branches; and 2) *Effectiveness of guided anchor.* AP and AR in the two branches were further increased by 2.3%–1.0% and 3.1%–2.3% respectively, especially AP_S and AP_L in the two branches by 2.7%–1.6% and 6.6%–6.1%, indicating that learning to predict rather than manually predefining the aspect ratio and size of anchors can more effectively identify and completely extract buildings with extreme aspect ratio or size. ; 3) *Effectiveness of* FL. Although AP and AR in the two branches were only increased by 0.6% and 0.7%–0.6% respectively, AP_M was increased by 2.0%–1.3%,

TABLE III
EFFECTS OF EACH COMPONENT IN OUR DESIGN

High-resolution Representation	Guided Anchor	Focal Loss	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
			61.9	84.2	71.5	64.8	23.5	24.2	67.2	58.6	84.5	69.4	61.1	24.0	20.9	63.4
√			66.5	85.7	76.2	69.0	**39.7**	26.3	71.7	62.6	85.8	74.1	64.7	**39.6**	25.0	67.2
√	√		68.8	89.7	**78.3**	71.7	21.7	**32.9**	74.8	63.6	89.2	74.8	66.3	22.6	**31.1**	69.5
√	√	√	**69.4**	**89.8**	78.2	**72.4**	23.7	29.3	**75.5**	**64.2**	**89.8**	**74.9**	**67.0**	23.9	27.4	**70.1**

TABLE IV
COMPLEXITY OF EACH COMPONENT IN OUR DESIGN

High-resolution Representation	Guided Anchor	Focal Loss	Params (M)	Flops (GMac)	Mem (GB)	Inf time (fps)
			98.74	**325.16**	**9.93**	**8.9**
√			85.50	325.74	11.33	6.5
√	√		86.09	325.71	10.82	6.4
√	√	√	**86.09**	325.71	10.80	6.4

proving that FL [23] can improve the recognition ability of few and hard samples.

The complexity of our components is compared with HTC [20], as shown in Table IV: 1) high-resolution representation is the main reason that our model size is 14.69% smaller than HTC [20] and our computational complexity is only increased by 0.17%; 2) guided anchor [22] reduces GPU overhead by 4.50%; and 3) FL [23] has almost no effect. Our method can maintain a relative balance between parameters (Params), flops (input size is 512 × 512), training memory (Mem), and inference speed (Inf time).

IV. CONCLUSION

This letter proposes an automatic building instance extraction method with practical value and development prospects. Compared with the mainstream Mask R-CNN [13] method, our method increased AP and AR in bounding box branch and mask branch by 9.8%–6.5% and 10.7%–8.0% respectively, especially AP_S and AP_L in the two branches by 10.1%–6.9% and 3.4%–2.4% respectively. Better extraction results for buildings with extreme changes in scale and appearance verified the effectiveness and complexity of various components. In the future, we will develop and verify the generalization ability of the model and further compress the size of the model.

REFERENCES

[1] R. Chen, X. Li, and J. Li, "Object-based features for house detection from RGB high-resolution images," *Remote Sens.*, vol. 10, no. 3, p. 451, Mar. 2018.

[2] K. Stankov and D.-C. He, "Building detection in very high spatial resolution multispectral images using the hit-or-miss transform," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 1, pp. 86–90, Jan. 2013.

[3] S. Xu *et al.*, "Automatic building rooftop extraction from aerial images via hierarchical RGB-D priors," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7369–7387, Dec. 2018.

[4] B. Yang, W. Xu, and Z. Dong, "Automated extraction of building outlines from airborne laser scanning point clouds," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 6, pp. 1399–1403, Nov. 2013.

[5] E. Maltezos, A. Doulamis, N. Doulamis, and C. Ioannidis, "Building extraction from LiDAR data applying deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 1, pp. 155–159, Jan. 2019.

[6] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[8] W. Sun and R. Wang, "Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with DSM," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 3, pp. 474–478, Mar. 2018.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, Munich, Germany, 2015, pp. 234–241.

[10] K. Lu, Y. Sun, and S.-H. Ong, "Dual-resolution U-net: Building extraction from aerial images," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 489–494.

[11] Z. Huang, G. Cheng, H. Wang, H. Li, L. Shi, and C. Pan, "Building extraction from multi-source remote sensing images via deep deconvolution neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 1835–1838.

[12] Y. Sun, X. Zhang, X. Zhao, and Q. Xin, "Extracting building boundaries from high resolution optical images and LiDAR data by integrating the convolutional neural network and the active contour model," *Remote Sens.*, vol. 10, no. 9, p. 1459, Sep. 2018.

[13] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. ICCV*, 2017, pp. 2980–2988.

[14] K. Zhao, J. Kang, J. Jung, and G. Sohn, "Building extraction from satellite images using mask R-CNN with building boundary regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 242–244.

[15] L. Zhang, J. Wu, Y. Fan, H. Gao, and Y. Shao, "An efficient building extraction method from high spatial resolution remote sensing images based on improved mask R-CNN," *Sensors*, vol. 20, no. 5, p. 1465, Mar. 2020.

[16] Q. Wen *et al.*, "Automatic building extraction from Google Earth images under complex backgrounds based on deep instance segmentation network," *Sensors*, vol. 19, no. 2, p. 333, Jan. 2019.

[17] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," presented at the 33rd Int. Conf. NeurIPS, Vancouver, BC, Canada, 2019.

[18] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9726–9735.

[19] Y. Shi, Q. Li, and X. X. Zhu, "Building footprint generation using improved generative adversarial networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 4, pp. 603–607, Apr. 2019.

[20] K. Chen *et al.*, "Hybrid task cascade for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4969–4978.

[21] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Apr. 1, 2020, doi: 10.1109/TPAMI.2020.2983686.

[22] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin, "Region proposal by guided anchoring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2960–2969.

[23] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.