Building Area Estimation in Drone Aerial Images Based on Mask R-CNN

Jun Chen, *Member, IEEE*, Ganbei Wang, Linbo Luo, Wenping Gong, and Zhan Cheng

***Abstract*— In rural areas where disasters occur frequently, the calculation of building areas is crucial in property assessment. In the segmentation algorithm, Mask R-CNN can distinguish the adjacent objects and extract the outline of an object. Based on this observation, we propose a novel method to calculate the building areas based on Mask R-CNN and adopt the concept of transfer learning to train our model, which can achieve good results with a small number of drone aerial images as training samples. The proposed method involves three main steps: 1) pretraining using open-source satellite remote sensing images; 2) fine-tuning with a small number of drone aerial images; and 3) testing with new images and area calculation based on the number of building pixels. The experiments show that the proposed method can achieve good results in terms of F1 score and intersection over union.**

***Index Terms*— Building area, drone aerial images, Mask R-CNN, transfer learning.**

I. Introduction

IN RECENT years, with the change of climate, geological disasters occur more frequently, especially collapses, mudslides, floods, and so on. These disasters seriously threaten to human life and property safety, so it is necessary to take appropriate preventive measures (e.g., antislip piles) in advance, especially in areas prone to disasters, such as landslide belts. Generally, the value of property is assessed according to the size of buildings, and hence, the building area needs to be calculated. The traditional methods are manually operated, which consumes a lot of time. In recent years, satellite technologies have been developed rapidly and applied to many remote sensing applications [1]–[4]. Therefore, the use of satellite images to extract buildings has become a trend [5]–[7]. However, the ground resolution of satellite images is relatively low, and large errors are caused when calculating the area, which will affect the property value assessment. Therefore, this letter considers using drones to obtain the images of the specific area. There are two main reasons why we use drone images for property evaluation. First, being equipped with a high-definition camera, we can easily obtain high-resolution images even at the centimeter level. Second, in some disaster-prone areas, using drones to collect data can reduce the risk. Therefore, it is desirable to develop a new method to automatically calculate the building areas based on drone images.

Recently, convolutional neural networks (CNNs) have become very popular in the applications of computer vision, such as target detection, image classification, and super-resolution and fusion [8]–[10]. Inspired by this, CNN was quickly introduced into image segmentation and achieved good results. Nevertheless, these methods still have restrictions. Generally speaking, satisfactory performance is difficult to achieve in case of a small amount of training data. As supervised methods, the methods based on CNN require a large amount of data to train. However, in our mission, the scale of data is typically small. Fortunately, some satellite remote sensing images with buildings can be easily obtained from open-source websites. Therefore, based on the concept of transfer learning, we can use a large number of satellite images for pretraining and then utilize a small number of drone aerial images for fine-tuning to achieve satisfying performance [11], [12].

Many approaches can be used to extract buildings from given scenes via deep learning methods. These approaches can be divided into two categories. One is semantic segmentation, such as FCN [13] and Deeplab [14], and the other is instance segmentation, such as fully convolutional instance-aware (FCIS) [15] and Mask R-CNN [16]. Semantic segmentation typically only judges if a pixel in a scene belongs to a certain class, while instance segmentation can be regarded as an extension of semantic segmentation, which further distinguishes each individual object in a scene. Instance segmentation is similar to the combination of object detection and semantic segmentation [17], [18]. Given that we have to separate the adjacent buildings from a specific area, the latter is more suitable in our task [19].

Mask R-CNN is a new network structure proposed by He *et al.* [16] recently. Without any trick, it outperforms all existing individual models in instance segmentation, including the winning model of the common objects in context (COCO) 2016 Challenge. Mask R-CNN has the advantages of simple structure, good flexibility, and remarkable effect. Therefore, it is preferable to be used for building segmentation in our task. After executing Mask R-CNN, we can extract the outline of each building and obtain the number of pixels in the contour.

Manuscript received February 1, 2019; revised July 11, 2019 and October 19, 2019; accepted April 3, 2020. Date of publication May 6, 2020; date of current version April 22, 2021. This work was supported by the National Natural Science Foundation of China under Grant 41977242 and Grant 61973283. *(Corresponding author: Jun Chen.)*

Jun Chen and Ganbei Wang are with the School of Automation, China University of Geosciences, Wuhan 430074, China, and also with the Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, Wuhan 430074, China (e-mail: chenjun71983@163.com; wangXWL86@163.com).

Linbo Luo is with the School of Mechanical Engineering and Electronic Information, China University of Geosciences, Wuhan 430074, China (e-mail: luolb@ige-live.com).

Wenping Gong and Zhan Cheng are with the Faculty of Engineering, China University of Geosciences, Wuhan 430074, China (e-mail: wenpinggong@cug.edu.cn; zhancheng2018@foxmail.com).

Color versions of one or more of the figures in this letter are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2020.2988326

According to the unit area represented by each pixel, we can finally estimate the area of each building.

The contributions of this letter involve the following three aspects. First, we propose a new method to calculate the building areas based on Mask R-CNN, which can accurately extract the outline of each building in a drone aerial image. Second, we adopt the concept of transfer learning to train our model, which can achieve good results with a small number of drone aerial images as training samples. Third, we collect a data set of drone images and test our method on it with comparison to several alternatives; the qualitative and quantitative results demonstrate the advantages of our method in terms of F1 score and intersection over union.

II. Method

We introduce the structure of Mask R-CNN and describe the method of calculating the building area based on Mask R-CNN, which is composed of three main parts: 1) pretraining with satellite images and fine-tuning with drone aerial images; 2) testing new drone images with the trained model to extract the contour of building and calculating the number of pixels within the outline; and 3) determining the unit area represented by each pixel and estimating the building area accordingly together with the number of building pixels.

A. Structure of Mask R-CNN

In order to correctly detect all objects in an image while precisely segmenting each instance, He *et al*. [16] proposed the Mask R-CNN, which combines the object detection algorithm, Faster R-CNN [20] with the semantic segmentation algorithm FCN [13]. The goal of Faster R-CNN is to classify individual objects and localize each with a bounding box, while the aim of FCN is to classify each pixel into a fixed set of categories without differentiating objects' instances. Specifically, a lot of region of interest (RoI) will be generated after an image passing through the Faster R-CNN, and then, the mask branch called FCN is applied to each RoI, predicting a segmentation mask in a pixel-to-pixel manner.

In particular, the Mask R-CNN uses RoI align instead of the RoI pool used in Faster R-CNN, which can reduce errors. RoI pool is a standard operation for extracting a small feature map (e.g., 7×7) from each RoI. RoI pool first quantizes a floating-number RoI to the discrete granularity of the feature map, and this quantized RoI is then subdivided into spatial bins that are themselves quantized. Finally, feature values covered by each bin are aggregated (usually by max pooling). Quantization is performed, e.g., on a continuous coordinate x by computing $[x/16]$, where 16 is a feature map stride and $[\cdot]$ is rounding; likewise, quantization is performed when dividing into bins (e.g., 7×7). These quantizations introduce misalignments between the RoI and the extracted features. While this may not impact classification, which is robust to small translations, it has a large negative effect on predicting pixel-accurate masks. To address this, Mask R-CNN proposes an RoI align layer that removes the harsh quantization of the RoI pool, properly aligning the extracted features with the input. The purpose is to avoid any quantization of the RoI boundaries or bins (i.e., $x/16$ instead of $[x/16]$). Bilinear interpolation is used to compute the exact values of the input features at four regularly sampled locations in each RoI bin and aggregate the result (using max or average).

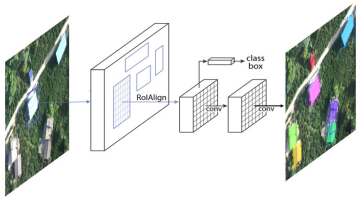

Fig. 1. Framework of Mask R-CNN for instance segmentation.

Fig. 2. Examples of (Top) drone images and (Bottom) satellite images.

In addition, using binary loss instead of multinomial loss can eliminate the competition among different types of masks. Another improvement of Mask R-CNN is the use of an enhanced backbone network using the residual neural network (ResNet) [21] instead of visual geometry group neural network (VGGNet) [22] to enhance the performance of feature extraction. ResNet was proposed by He *et al.* [16] and won the championship in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2015 competition. Compared with VGGNet, the effect is better and the parameters are less. In general, RoI align uses bilinear interpolation to avoid quantization operations, which is an important improvement and has been applied to many networks.

Mask R-CNN is a flexible framework and can be applied to various tasks with the state-of-the-art performance, including target detection, image segmentation, and human pose recognition [23]. For the COCO challenge with all tasks, Mask R-CNN performs better than the previous model. At the same time, multiple architectures are used as the backbone network for Mask R-CNN, including Resnet-101 or Resnet-101-Feature Pyramid Networks (FPN) and ResNeXt-101-FPN [24], where FPN is applied to extract multiscale features by using a top–down architecture with lateral connections to build an in-network feature pyramid from a single-scale input.

Mask R-CNN performs detection first and then segmentation. In our task, Mask R-CNN identifies each object with a bounding box. Each bounding box is then segmented into building and nonbuilding regions. The structure of Mask R-CNN is shown in Fig. 1.

B. Implementation of the Proposed Method

Although Mask R-CNN can achieve good results in instance segmentation, it requires a large number of data for training. When the scale of training data is small, how to get a better segmentation performance becomes a problem. In this letter, we solve this problem by the technique of transfer learning.

Transfer learning is a machine learning technique, in which a model trained on one task is reused on another related task to improve the model accuracy for that task. In particular, it first trains a model using a lot of data that are similar to the target task and then readjust or migrate the learned features to the

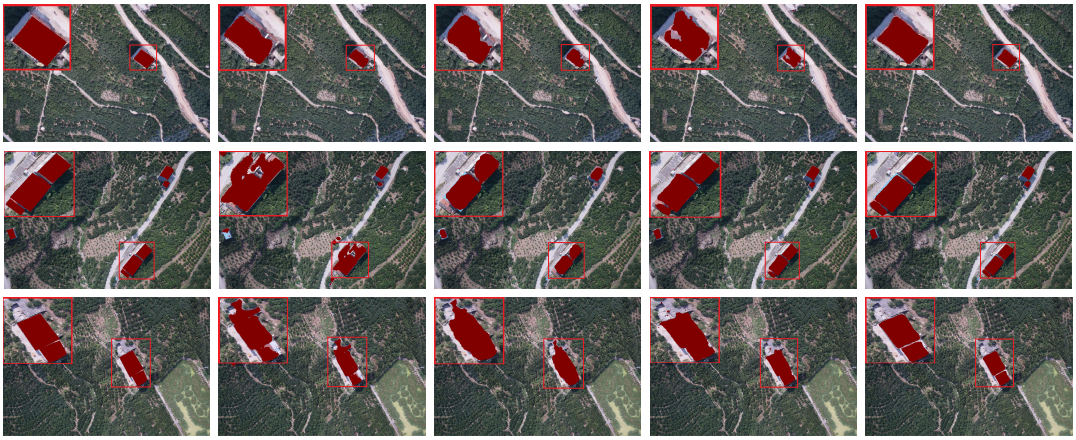

Fig. 3. Qualitative illustration of different methods on three typical drone images. (From Left to Right) Original drone images with ground truth, result of FCN [13], deeplab [14], SegNet [25], and our used Mask R-CNN [16].

target task, that is to say, we do not need to train a model from zero. In general, it can overcome the problem of overfitting caused by a too-small amount of data and improve the model accuracy. The main steps of our method are as follows.

First, we use the publicly available satellite remote sensing data set for pretraining. Given the differences between drone aerial images and satellite remote sensing images, we cannot directly test with the drone aerial images due to their differences in scenes and ground resolutions. Therefore, we use drone aerial images for fine-tuning. Second, we select 870 drone aerial images for fine-tuning and obtain our final model. During testing, the learned model will output the outline of the building region for an input image. Finally, we determine the area of one single pixel according to the camera intrinsics and flight parameters of the drone and then calculate the number of pixels in the building region. Subsequently, the area of building is obtained by the product of these two values.

III. Experimental Results

In this section, we test the performance of the proposed method on drone aerial images and compare it with other classic methods. Implementation is based on TensorFlow. Experiments are performed on two NVIDIA GeForce 1080 Ti GPUs, which takes about two days to train our network.

A. Data Sets

The Mapping Challenge encourages people to automatically extract the outline of a building from remotely sensed images. This challenge provides a large number of images,[1] including 280 000 training images and 60 000 verification images. The data set is publicly available, which only have one class of buildings, and the sizes of images are all 300×300 (pixels). In addition, we collect 937 drone aerial images of size 1024×682 with a ground resolution of 3.1 m $\times$ 3.1 m and manually label the buildings on these images using LabelMe, a software for annotations. Among the 937 images, 870 of which are used for training, and the rest 67 images are used for testing. Examples of satellite and drone images are shown in Fig. 2, where all of them are RGB images. The drone aerial images were captured in Zigui, China, by ourselves in 2018.

[1]Available at: https://www.crowdai.org/

Fig. 4. Building area estimation using Mask R-CNN. There are seven individual buildings segmented by Mask R-CNN marked in different colors. The value beside each bounding box is the estimated area of corresponding building.

B. Result Analysis

We use three classic models for comparison, including FCN [13], SegNet [25], and deeplab [14]. Using the transfer learning method, for all the methods, the satellite images are first used for pretraining, and then, the final model is obtained by fine-tuning on the small number of drone images. Some representative qualitative results are provided in Fig. 3. From the results, we see that the Mask R-CNN can achieve the best performance. In particular, the buildings are basically completely divided and it can also distinguish the adjacent buildings (see Fig. 4 for example). In densely populated areas, this is important for calculation the area of an individual building.

To provide a quantitative evaluation, we consider two quantitative metrics, including F1 score and intersection over union (IoU). The definition of the F1 score is based on precision and recall, where the precision is defined as the ratio between the number of pixels in the identified true building area and the number of pixels in the whole identified building area, while recall is defined as the ratio between the number of pixels in the identified true building area and the number of pixels in the existing building area. The IoU is defined as the ratio between the intersection and the union of the target and the prediction. Specifically, the definitions of F1 and IoU are shown as follows:

$$F_1 = \frac{2\text{Precison} \cdot \text{Recall}}{\text{Precison} + \text{Recall}} \tag{1}$$

$$\text{IoU} = \frac{\text{Target} \bigcap \text{Prediction}}{\text{Target} \bigcup \text{Prediction}}. \tag{2}$$

TABLE I

PERFORMANCE COMPARISON OF FOUR MODELS. BOLD INDICATES THE BEST

	FCN [13]	deeplab [14]	SegNet [25]	Mask R-CNN
F_1	51.1%	64.8%	72.3%	**80.9%**
IoU	34.3%	47.9%	56.7%	**68.0%**

TABLE II

ESTIMATION OF BUILDING AREA USING MASK R-CNN. GT IS THE ABBREVIATION OF GROUND TRUTH (UNIT: m^2)

Building	A	B	C	D	E	F	G
GT	79.3	21.8	68.6	115.2	71.5	35.6	160.9
Ours	79.4	21.5	69.1	111.8	79.1	27.1	157.3

The quantitative results of different models on the whole 67 test drone images are reported in Table I. From the results, we see that for both the two metrics, Mask R-CNN with transfer learning can achieve the best results. In addition, we also test the performance changes by replacing the RoI align with the RoI pool in the Mask R-CNN architecture and find that the average IoU score will sharply drop about 10%.

C. Estimation of Building Area

After calculating the area of one single pixel and the number of pixels in the building region, we can finally obtain the area of building according to the product of these two values. We demonstrate some typical results in Fig. 4, where our method marks each individual building with different colors and calculates their areas. The area is given in the top-left corner of each bounding box. To provide a quantitative comparison, we also present the ground-truth area of each building, as shown in Table II. Clearly, the error rate is typically less than 5%, which is satisfying in practical use.

IV. CONCLUSION

In this letter, we present a novel method based on Mask R-CNN to estimate building areas. Through transfer learning, we can get an initial model and then fine-tune it with a small number of drone aerial images. The qualitative and quantitative results show that the proposed method can achieve better results on a small number of training data sets than the other classic alternatives.

Our method can also be applied to the area estimation of cultivated land, lakes, forest, and so on In our future work, we will apply our method to more remote sensing monitoring tasks using drone images. In addition, in order to keep the ground resolution consistent, the drone needs to be maintained a constant height during flight.

REFERENCES

[1] J. Yuan, "Learning building extraction in aerial scenes with convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2793–2798, Nov. 2018.

[2] J. Ma, J. Jiang, H. Zhou, J. Zhao, and X. Guo, "Guided locality preserving feature matching for remote sensing image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4435–4447, Aug. 2018.

[3] B. Bischke, P. Helber, J. Folz, D. Borth, and A. Dengel, "Multi-task learning for segmentation of building footprints with deep neural networks," 2017, *arXiv:1709.05932*. [Online]. Available: http://arxiv.org/abs/1709.05932

[4] J. Ma, H. Zhou, J. Zhao, Y. Gao, J. Jiang, and J. Tian, "Robust feature matching for remote sensing image registration via locally linear transforming," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6469–6481, Dec. 2015.

[5] 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.

[6] H. L. Yang, J. Yuan, D. Lunga, M. Laverdiere, A. Rose, and B. Bhaduri, "Building extraction at scale using convolutional neural network: Mapping of the united states," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2600–2614, Aug. 2018.

[7] 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, 2018.

[8] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4980–4995, 2020.

[9] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11–26, Aug. 2019.

[10] J. Ma, X. Wang, and J. Jiang, "Image super-resolution via dense discriminative network," *IEEE Trans. Ind. Electron.*, vol. 67, no. 7, pp. 5687–5695, Dec. 2020.

[11] T. Tian, C. Li, J. Xu, and J. Ma, "Urban area detection in very high resolution remote sensing images using deep convolutional neural networks," *Sensors*, vol. 18, no. 3, p. 904, 2018.

[12] Z. Yang *et al.*, "Deep transfer learning for military object recognition under small training set condition," *Neural Comput. Appl.*, vol. 31, no. 10, pp. 6469–6478, Oct. 2019.

[13] 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.

[14] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[15] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, "Fully convolutional instance-aware semantic segmentation," 2016, *arXiv:1611.07709*. [Online]. Available: http://arxiv.org/abs/1611.07709

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

[17] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. ECCV*, 2016, pp. 21–37.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[19] V. Iglovikov, S. Seferbekov, A. Buslaev, and A. Shvets, "TernausNetV2: Fully convolutional network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 233–237.

[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[23] R. A. Guler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7297–7306.

[24] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[25] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.