EasyChair Preprint

№ 7726

Multi Deep Learning Model for Building Footprint Extraction from High Resolution Remote Sensing Image

Ho Trong Ánh, Tran Anh Tuan, Hoàng Phi Long, Lê Hai Hà and Tran Ngoc Thăng

April 6, 2022

Multi Deep Learning Model For Building Footprint Extraction From High Resolution Remote Sensing Image

Ho Trong Anh[1], Tran Anh Tuan[1†], Hoang Phi Long[1†], Le Hai Ha[1] and Tran Ngoc Thang[1*]

[1]School of Applied Mathematics and Informatics, Hanoi University of Science and Technology, No. 1 Dai Co Viet, Ha Noi, Viet Nam.

*Corresponding author(s). E-mail(s): thang.tranngoc@hust.edu.vn;
Contributing authors: Anh.HT211306M@sis.hust.edu.vn; trananhtuan23012000@gmail.com; longhoangphi225@gmail.com; ha.lehai@hust.edu.vn;
†These authors contributed equally to this work.

Abstract

3D city modeling is a new development trend in cartography that has a lot of practical and scientific value. The project necessitates the extraction of a building footprint using remote sensing images. This research examined how to solve the Building Footprint problem using automatic segmentation methods. We reviewed popular segmentation models as Mask-RCNN, U-net, and U2-net, and developed two multi-models that generated more stable and good results than the single models.

Keywords: segmentation, convolution neural networks, building footprint, remote sensing

1 Introduction

Building footprint extraction from huge remote sensing images is a challenging endeavor. In recent years, high-tech businesses have made significant investments in surveillance equipment with high resolution, providing a rich supply of image data for remote sensing. Automatically detecting the footprint of a structure has never been easier thanks to recent developments in computer vision.

In recent years, many powerful segmentation models, such as U-net, U2-net, and Mask-RCNN, have been created. Through experiments [1], U-net has demonstrated its potential in the Building Footprint problem. Wei et al. [2] advocated using the U2-net model to create an accurate position and exact building outline. To tackle the problem, Mask-RCNN has been combined with post-processing [3] or preprocessing [4] stages. These studies are helping to solve the problem of the Building Footprint, but more research is needed to improve and validate their accuracy and usefulness.

In remote sensing images, a building's pixels are substantially larger than the 28×28 pixels of a mask produced by Mask-RCNN, hence the sampling results are less detailed. When we replaced the Mask-RCNN branch with U-net and U2-net, we were able to generate two multi-models and compare the results to single segmentation models.

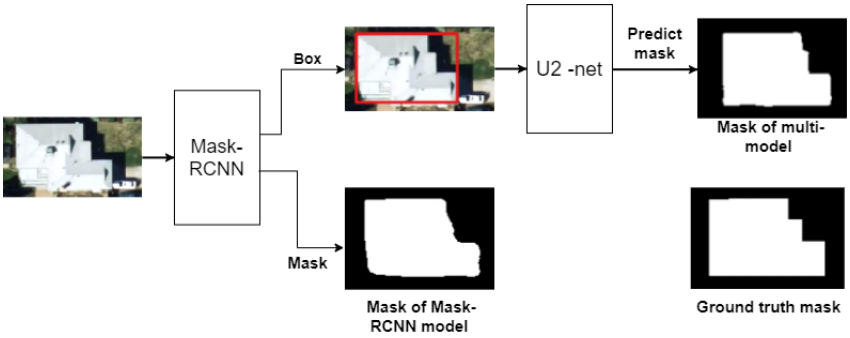


Fig. 1: Mask-R CNN and U2-net combined to form a generalized multi-model. The similarity between the Multi-model mask and the ground truth mask was obvious, which confirms the effectiveness of the Multi-model.

2 Method

2.1 U-net

The U-net model developed by Olaf Ronneberger et al. aims to partition cells from biomedical imaging. The model architecture consists of two main branches: the downsampling path and the upsampling path in the same way as

the encoding and decoding process. The downsampling path is responsible for learning the features from the input image like a conventional CNN network, with the key trait that each time it reduces, the number of filters rises correspondingly such as Resnet, VGG16, etc. As for the upsampling path, its job is to restore the size after the downsampling path to its original size. When it comes to the upsampling approach, its primary function is to restore the size to its original size after downsampling. The model's downsampling path may help reduce information loss by reusing the data that was previously used in the downsampling process. We chose available architectures such as Resnet50 from the pre-trained model ImageNet when choosing the initialization architecture for the research process with a change in activation function, that was ReLU.

2.2 Mask R-CNN

Recently, many deep learning models have been presented, mainly for object classification. These models have many different approaches, such as SSD bounding box approach, box-based object search on different aspect ratios, and scale per image location. Or U-net approaches in the direction of finding masks of objects in the image. These methods each have their advantages and disadvantages that seem to complement each other. There is a model that approaches object classification in both ways and is called Mask R-CNN. This is an extension of the Fast R-CNN model by adding a branch to predict the object's mask based on parallelism with bounding box detection. This model is quite simple to train, and the cost is not too much greater than the Fast R-CNN model.

2.3 Multi-model

The U-net model's identification of dense objects is quite difficult, but it has high accuracy when detecting sparse objects. In contrast, with the Mask R-CNN model, we get the ability to identify dense areas accurately. Therefore, when combining the two models, that will give us an expectation of the model's efficacy for the Building Footprint problem.

To accomplish this, we combined two models: the U-net and the Mask R-CNN. During the training phase, we merged the Mask-RCNN model outputs into the identified object branch, which were utilized as inputs to the U-net model. Although we expected a multi-model between the Mask-RCNN model and the U-net model, which was made effective, the evaluation metrics are contrary. So, we replaced the U-net model with the U2-net model, which improved the disadvantages that the two original models (the U-net model and the Mask-RCNN model) have. When compared to applying each model independently, the results on the test set demonstrated that the two integrating models considerably modify the AP value at the IoU thresholds, as shown in Figure 3. The difference between the mask of the Mask-RCNN model and the multi-model is shown in Figure 2.

Fig. 2: Top left: Input image. Top right: Ground truth mask. Lower left: Mask was predicted by Mask-RCNN model. Lower right: Mask was predicted by multi-model (Mask-RCNN model combined with U2-net model).

3 Experiments

3.1 Datasets

This research was conducted on three different datasets: SUNNY VALE, USA UAV, and VN UAV. The SUNNY VALE dataset [5] contains High-Resolution Orthoimagery images with a resolution of 30 cm. For the training/test/validation set, we picked 16 images with a size of 5000×5000, split them into images with sizes of 256×256, and randomly divided them with a scale of $6 : 2 : 2$. Furthermore, the data label was obtained from the building dataset on the website OSM (Open Street Map) [6]. The following dataset, USA UAV, is a collection of UAV images from several locations in the United States, including Santa Ana, Visalia, California, and Orem, Utah. They were labeled by the Skymap Global company's label and were utilized for research purposes only. These images were separated into 512×512 images that included 3600 images for the training set, 1200 images for the validation set, and 1200 images for the test set, with resolution quality ranging from 7.5 cm to 15 cm. Finally, the VN UAV (Viet Nam UAV) dataset was also provided by the Skymap Global company. They were split into images with dimensions of 512×512, which included 3000 images for the training set, 1000 images for the validation set, and 1000 images for the test set, with a resolution quality of 10 cm.

3.2 Evaluation Metrics

Before calculating evaluation metrics, we excluded objects that were located on the contour of the image and were recognized more than once by the non-maximum-suppression method. We used Kaggle's contest-based evaluation metrics [7] including mAP and mAR. Both mAP and mAR were averaged by AP values at IoU thresholds, which were in $[0.55 : 0.05 : 0.95]$. At the same

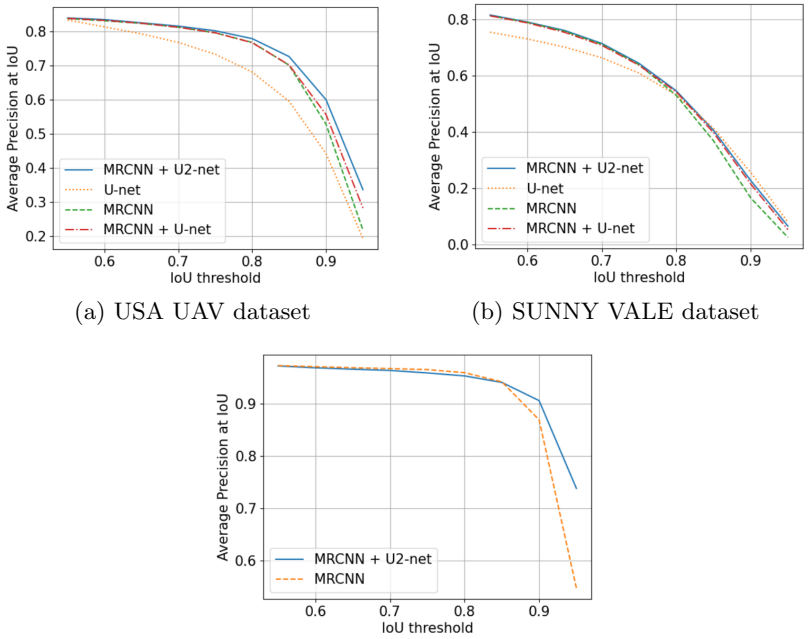
AP's IoU criteria, the accuracy and recall were likewise calculated with a batch size of 4.

4 Results

Table 1 show evaluated results on HP-Z800 workstation with configuration: 02 CPU Intel Xeon Processor X5650, 32GB RAM, GPU Nvidia GTX 1080 Ti 11GB.

Table 1: Results of models on 3 datasets with mAP, mAR in IoU range [0.55 : 0.05 : 0.95]. The entries are blank because of the non-converging model.

	USA UAV		SUNNY VALE		VN UAV	
Model	mAP	mAR	mAP	mAR	mAP	mAR
U-net	0.65	0.615	0.526	0.52	-	-
MRCNN	0.701	0.787	0.533	0.54	0.906	0.903
MRCNN+U-net	0.712	0.795	0.544	0.55	-	-
MRCNN+U2-net	**0.728**	**0.813**	**0.551**	**0.56**	**0.930**	**0.927**



(a) USA UAV dataset

(b) SUNNY VALE dataset

(c) VN UAV dataset

Fig. 3: Visualizing the AP value at the IoU thresholds on 3 datasets.

Construction works are typically spaced and moderate in countries with strong construction planning, such as the United States, allowing the U-net model to function well on this data. In contrast, in countries with poor building planning, such as Vietnam, construction activities are near and dense, making the combination of the Mask R-CNN and the U-net ineffective. The adoption of the U2-net model to replace the U-net model, later on, improves identification in instances when buildings are hidden by trees or are close together. The predicted mask of the Multi-model, as shown in Figure 4, exhibits accuracy when compared to the mask of Mask R-CNN model, which substantially assists further work when using geometry correction algorithms to obtain the best linear outlines for objects.

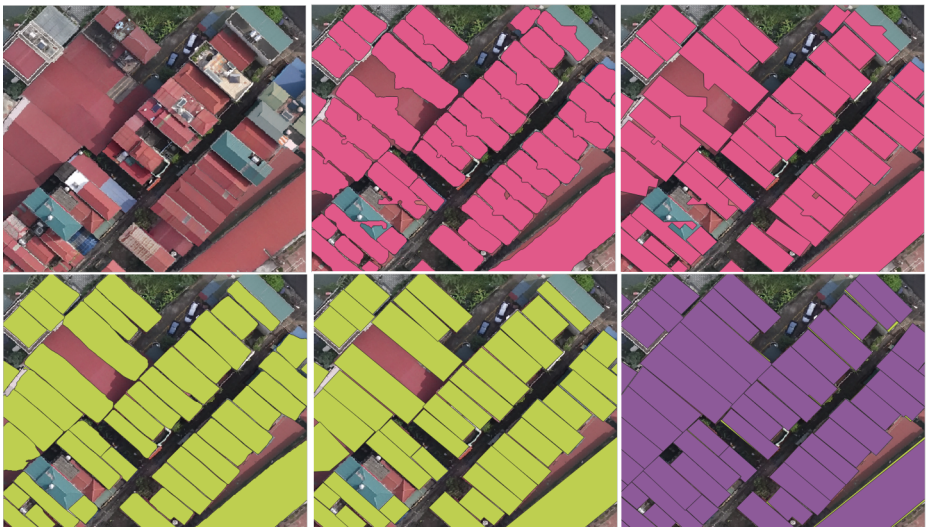


Fig. 4: The first top-left image is the input image, second and third is the Mask R-CNN predictions and geometry correction algorithm predictions respectively. The first lower-left image and second are the Multi-model predictions and corresponding geometry correction algorithm predictions, and the third image is the expected output image.

5 Conclusion

Except for U-net, all of the empirical models perform well on various data sets. Multi-models provided significant improvements, according to the findings. As a result, we have proposed an effective solution to the Building Footprint problem, which may be used as a foundation for post-processing procedures such as geometry correction and proving its multi-model potential application for other tasks.

References

[1] Emek, R.A., Demir, N.: Building detection from sar images using unet deep learning method, 215–218 (2020). https://doi.org/10.5194/isprs-archives-XLIV-4-W3-2020-215-2020

[2] Wei, X., Li, X., Liu, W., Zhang, L., Cheng, D., Ji, H., Zhang, W., Yuan, K.: Building outline extraction directly using the u2-net semantic segmentation model from high-resolution aerial images and a comparison study. Remote. Sens. **13**, 3187 (2021)

[3] Zhao, K., Kang, J., Jung, J., Sohn, G.: Building extraction from satellite images using mask r-cnn with building boundary regularization. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 242–2424 (2018). https://doi.org/10.1109/CVPRW.2018.00045

[4] Qinzhe, H., Yin, Q., Zheng, X., Chen, Z.: Remote sensing image building detection method based on mask r-cnn. Complex & Intelligent Systems (2021). https://doi.org/10.1007/s40747-021-00322-z

[5] USGS: Sunnyvale uav images. https://earthexplorer.usgs.gov/

[6] OSM: Sunnyvale uav labels. https://www.openstreetmap.org/

[7] Kaggle: 2018 data science bowl (2018). https://www. kaggle.com/c/data-science-bowl-2018