City-wide building footprint extraction from satellite images based on deep instance segmentation model

Shenglong Chen*, Yoshiki Ogawa**, Yoshihide Sekimoto**

Abstract: Building footprint is one of the primary data in the urban geographic information database and is critical in the applications such as urban planning, population estimation, and disaster prevention. In recent years, with the development of machine learning technology and the application of high-resolution remote sensing images, automatic extraction of building footprints from remote sensing images based on a deep learning algorithm is currently the most popular method. However, due to the diversity of feature textures and scale, discriminating adjacent buildings over a large area and applying the model to a different region remain significant challenges. As a result, this study attempted to propose a framework for city-wide building footprint extraction at the instance level using the Mask R-CNN model and test the model generalization ability for different remote sensing images and regions.

Keywords: Building extraction, Remote sensing, Deep learning, Instance segmentation.

1. Introduction

The building footprint is one of the primary data in the urban geographic information database and is essential to urban planning, map update, and disaster prevention. With the development of remote sensing technology, high-resolution remote sensing images with rich feature information have been widely used in ground targets extraction. Therefore, the study of automatic building footprints extraction from remote sensing images for a wide area is significant in academic and practical terms.

In recent years, machine learning technology represented by deep learning algorithms has ushered in a new wave of research, which has achieved excellent results in the field of computer vision. Therefore, many researchers have accordingly tried to apply deep neural networks to building extraction. It is shown that deep neural networks can automatically learn target features from many high-resolution remote sensing images with efficient feature representation (Deshapriya et al., 2020). However, due to the diversity of building features and scales, deep learning methods suffer from connected adjacent buildings and boundary integrity. The mainstream semantic segmentation models do not have high accuracy, for instance, level extraction, with a typical object-wise recall below 0.7 (Weir, 2019). Also, the model generalization ability for remote sensing images from diverse sources and buildings in different regions needs to be improved.

Based on the above background, in this study, an instance segmentation model based on Mask R-CNN was trained on high-resolution aerial images using transfer learning, aiming to improve the object-wise recall above 0.8. Besides, for test regions different from the training set, 10% training data of the target region was produced to fine-tune the model. Finally, based on the above training strategy, the model performance is compared and analyzed for various sources of remote sensing images (0.25m resolution aerial images and 0.6m resolution satellite images) and different regions (Setagaya, Hachioji, and Mashiki) to verify the model generalization ability.

2. Related work

With the widespread use of deep learning in remote

* Student member Institute of Industrial Science, University of Tokyo
〒153-8505 4-6-1 Komaba, Meguro City, Tokyo Tel: 03-5452-6406
** Regular member Center for Spatial Information Science, University of Tokyo
〒277-8568 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba Tel: 04-7136-4291

sensing, some deep learning-based methods have established traditional ones in related fields. (Ball et al., 2017). The convolutional neural networks (CNN) (Krizhevsky et al., 2013) have risen to prominence in deep learning. They are popular in building extraction applications, which are summarized in the following three approaches.

The first method is the sliding window method based on the image classification task (Farabet et al., 2012). The method uses a sliding window to traverse the entire remote sensing image in a specific step to obtain a fixed size tile. Then the tile is input into the CNN network to predict the class of the central pixel to obtain the segmentation result of the whole image. Mnih used the method to conduct experiments on the Massachusetts buildings dataset and compared the base model, the model with conditional random fields, and the model with post-processing. The highest accuracy of 92.03% was achieved for extracting buildings (Minh, 2013). However, this method results in a large number of repeated calculations, which hurts image segmentation efficiency.

The second method combines image segmentation with neural network classification and object-oriented semantic segmentation with CNN (Zhao et al., 2017). This method consists of two steps. Firstly, the image is segmented into potential object patches using traditional image segmentation methods and compressed, stretched, and filled with meeting the input size of the neural network. Secondly, these image patches are input into the neural network for training and classification. However, since deep learning methods are not used in the image segmentation process, the feature representation is insufficient to describe some complex geometric features and obtain global-level contextual information. (Zhao et al., 2018).

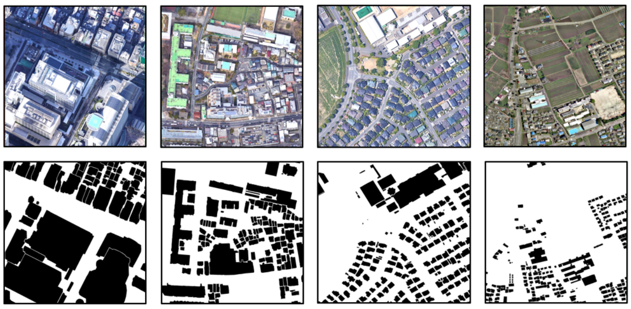
The third method is semantic segmentation based on a fully convolutional neural network (FCN) (Long et al., 2015). FCN is an end-to-end deep learning network. The idea is to replace all the last few fully connected network layers with convolutional layers, resulting in a two-dimensional feature map. Most current research in building extraction has used semantic segmentation methods based on FCN and its variants (Ji et al., 2018), such as BRRNet (Shao et al., 2020), Arc-net (Liu et al., 2020), and De-net (Liu et al., 2019). However, the experiments in these papers consider only pixel-wise classification accuracy.

The goal of the building extraction is not to focus on whether a pixel is a building or not but more on the building object itself. It is a typical instance segmentation task. The most popular model is the region-based model, such as Mask R-CNN (He et al., 2017). However, current research on CNN-based building instance segmentation is still scarce and urgently needs to be populated. Thus, it remains a challenge to obtain satisfactory building extraction results at the instance level.

3. Methodology

3.1. Dataset

The data source of the remote sensing images used in this study is from Google Earth (GE). The training set contains a total of 528 aerial ortho-color images of size 1024 pixels × 1024 pixels with a spatial resolution of 0.25 m and more than 36,000 high-quality, manually labeled building footprints. The whole dataset covers three different areas, ranging from the high-rise area of Shinjuku City to the suburban residential area of Hachioji City. Besides, to evaluate the model's generalization ability, another training set of Mashiki town was created for fine-tuning, containing 26 satellite images with 0.6 m resolution from the tilemap of Geospatial Information Authority of Japan (GSI). The example of remote sensing images and labeled buildings of different areas are shown in Figure 1.



(a) Shinjuku (b) Setagaya (c) Hachioji (d) Mashiki

Figure 1. Example of training data of different areas

In order to meet the input requirements of neural networks and the efficient use of video memory, the size of remote sensing images in the dataset needs to be unified. In this study, all images are cropped to 512 pixels by 512 pixels, and the defective parts are filled using black pixels. Finally, the cropped data are assigned to the training set, validation set, and test set in the ratio of 60%, 20% and 20%, as shown in Table 1.

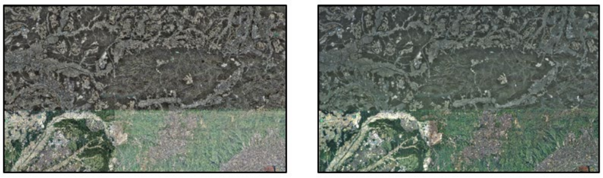
Table 1. Number of image data in different regions

Area	Image source	Training set	Validation set	Test set
Shinjuku	GE	131	33	0
Setagaya	GE	144	48	48
Hachioji	GE	1366	171	171
Mashiki	GSI	62	21	21

3.2. Image pre-processing

Remote sensing images are subject to errors in data acquisition due to various factors, so it is necessary to pre-process the images before training, including color balance, linear stretching, and bilateral filtering (Elad, 2002).

Due to the influence of image acquisition time, external lighting, and other factors, there are color differences in the acquired images, especially for large areas. In order to eliminate the color difference, the mask dodging algorithm (Zhang et al., 2011) is adopted. The target color of each pixel is picked up from the third-order target surface. As shown in Figure 2, the color transition between adjacent images becomes more seamless, and the overall tone is unified after color balancing.



(a) original images (b) after color balance

Figure 2. Effect of color balance

Image stretching aims to enhance image contrast, reduce image data volume, and transform the data into an 8-bit image suitable for deep neural network processing. In this study, the 2% minimum-maximum linear contrast stretch based on a grayscale histogram is applied. Moreover, to extract the target features while eliminating image noise, the bilateral filtering algorithm is adopted to perform spatial filtering and retain the target edge contour information while smoothing the noise. The image after pre-processing is shown in Figure 3.

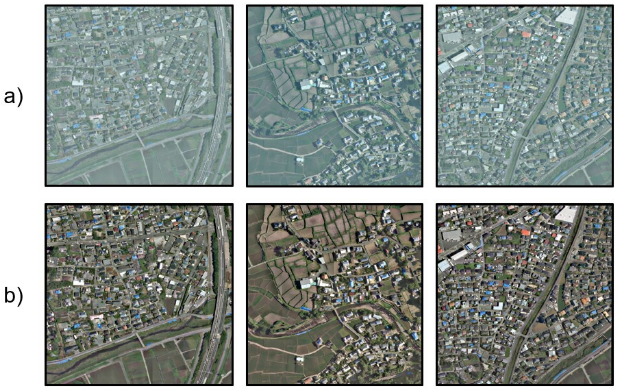


Figure 3. Result of linear stretching and bilateral filtering.

(a) Original images; (b) After pre-processing.

3.3. Instance segmentation model

This study's deep learning segmentation algorithm is based on the Mask R-CNN instance segmentation network (He et al., 2017). The network is derived from Faster R-CNN and Fully Convolutional Network (FCN), to which a new task branch is added to complete the pixel instance segmentation of the target object. The architecture of the network is as shown in Figure 4. The input images are first sent to ResNet for feature extraction. The obtained backbone feature map is passed through Region Proposal Network (RPN) to extract the possible target regions (ROI). These ROIs are mapped

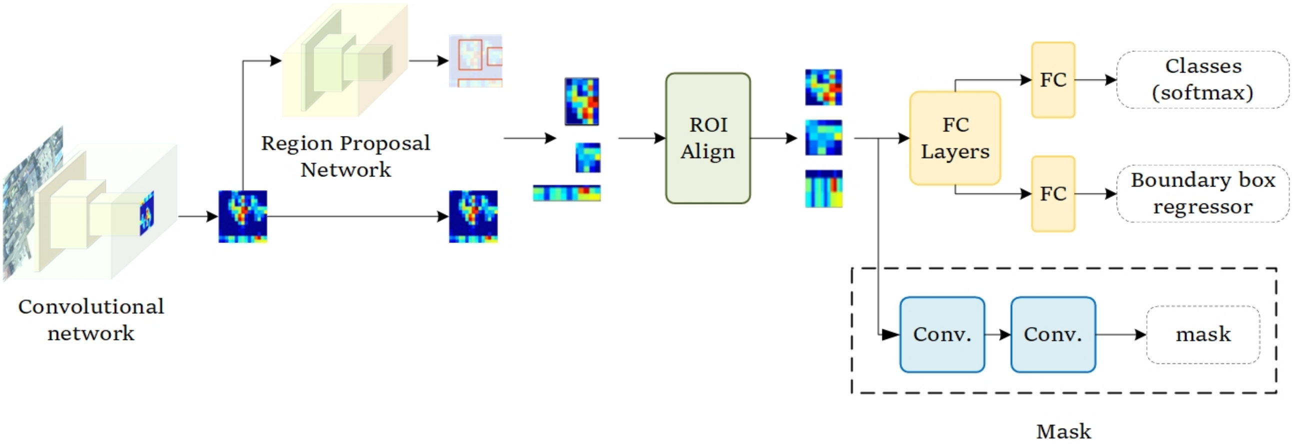


Figure 4. Architecture of Mask R-CNN.

into fixed dimensional feature vectors by the ROIAlign Layer. Two branches are for classification and regression of the target boundary box through the Fully Connected Layer. The other branch is up-sampled by the Fully Convolutional Layer to obtain the segmented region image.

3.4. Results post-processing

The raw output of the Mask R-CNN network contains three task branches: the target class, the coordinates of boundary boxes, and the binary segmentation mask of target regions. Since building extraction is a single-class segmentation task, we are interested in the mask of the building footprint. However, the model may return multiple polygons for the same building, especially as a tiling side effect. For this case, a non-maximum suppression algorithm (Hosang et al., 2017) is applied to remove the feature with the lower confidence value if two features overlap more than a given maximum ratio, as shown in Figure 5. After that, the new output without duplicate features is given geographic coordinates and mosaicked together to obtain the final prediction results in GeoTIFF format.

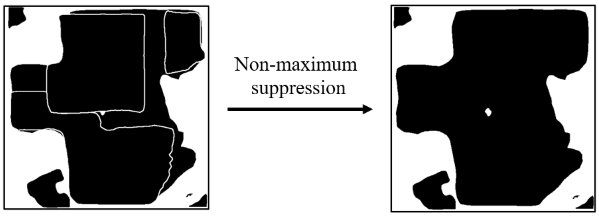


Figure 5. Effect of non-maximum suppression algorithm.

4. Experiment

4.1. Training process

The algorithmic network was implemented on the Detectron2 (Wu et al., 2019), an open-sourced platform for state-of-art detection and segmentation algorithms provided by Facebook. The hardware environment is an EC2 P2 instance of Amazon web service (AWS) with NVIDIA K80 GPU graphics (12GB) and 64-bit Ubuntu 18.04 operating system.

Since the dataset is not large enough to train a Mask R-CNN model end-to-end from the start, a model pre-trained on the COCO dataset with a ResNet-101 FPN model (X-101-FPN) backbone was used for transfer learning. The pre-trained model is available from the Detectron2 Model Zoo. Transfer learning reduces the training data and effectively improves the overall accuracy and generalization ability of the model.

The dataset with GE images (Shinjuku, Setagaya, and Hachioji) described in Section 2.1 is used for training. The max number of iterations is set to 5000; the batch size is set to 8; the initial learning rate is set to 0.0001. Also, the Adam algorithm is used for network optimization. All the parameters are initialized according to the orthogonal distribution. After about 18 hours of training, the loss function stops decreasing and converges to 1.420, as shown in Figure 6.

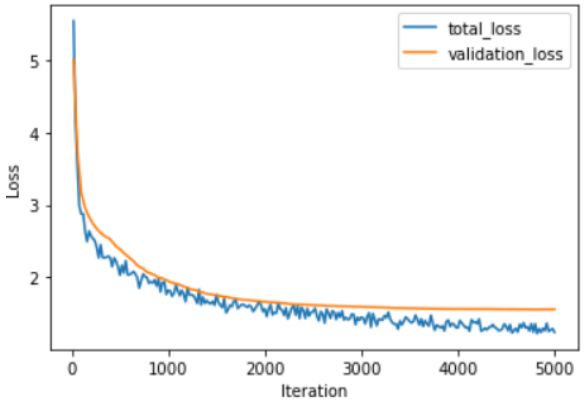


Figure 6. Learning curve of model training.

4.2. Evaluation Criteria

In order to quantitatively evaluate the performance of building extraction algorithms, this study uses the mean recall (mRecall), mean precision (mPrecision), and F1 scores to evaluate the prediction results, as shown in Equation (1) to Equation (3). These object-wise metrics are often used for remote sensing segmentation evaluation.

$$mPrecision = \frac{1}{k+1}\sum_{i=0}^{k}\left(\frac{TP}{TP+FP}\right) \tag{1}$$

$$mRecall = \frac{1}{k+1}\sum_{i=0}^{k}\left(\frac{TP}{TP+FN}\right) \tag{2}$$

$$F1 = \frac{2 \times mPrecision \times mRecall}{mPrecision + mRecall} \tag{3}$$

Where k is the randomly selected k sets of test images; TP (True Positive) means the correct building detection; FP (False Positive) means the error building detection; FN (False negative) means error background detection. This study uses an intersection over union (IoU) threshold of 0.5 to classify predictions as TP.

4.3. Experiment result

In order to examine the performance of the trained model, building extraction was performed on the test set of Setagaya City. Also, a U-Net model with a VGG-16 encoder (Ronneberger et al., 2015) is applied on the same test set to compare the effectiveness of the method used in this study with the mainstream model. The used U-Net model was pre-trained on the SpaceNet Nadir Imagery Dataset (Weir et al., 2019) and transfer-learned on the same training set of the Mask R-CNN model.

The confusion matrix and accuracy of the prediction results of the two models are shown in Table 2 and Table 3. The accuracy of the Mask R-CNN model is lower than that of the U-Net, especially the precision drops by 17%. Figure 7 randomly shows three sets of buildings extraction results on the Setagaya test set of two models and corresponding source images and ground truth. The figure shows that the extraction rate and contour integrity of the Mask R-CNN model is better than the U-Net model for some large stand-alone buildings. However, for general buildings, the results of the Mask R-CNN model have more missing, the footprint contours are coarser, and the adjacent buildings are not segmented.

Overall, the result was not as good as expected, probably due to insufficient training sets for transfer learning. On the other hand, the SpaceNet dataset, on which the pre-trained model of the comparative U-Net trained, contains data from over 120,000 buildings.

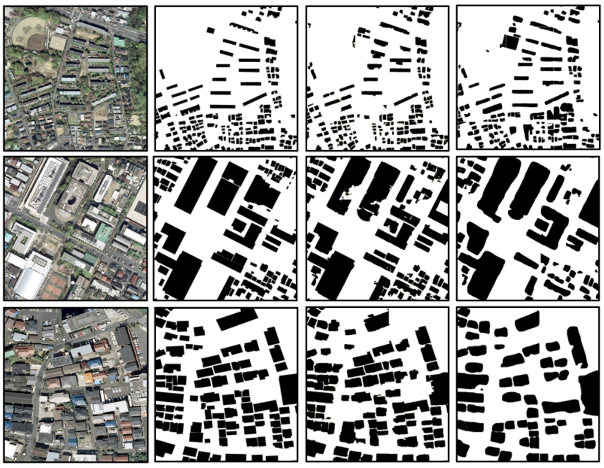

Source image Ground truth U-Net Mask R-CNN

Figure 7. Prediction results of two models

Table 2. Confusion matrix of two models.

Model	Confusion matrix		Actual Value	
			Positive	Negative
U-Net	Predicted values	Positive	3147	667
		Negative	2015	-
Mask R-CNN	Predicted values	Positive	2691	1345
		Negative	2471	-

Table 3. Object-wise accuracy of two models

Model	mPrecision	mRecall	F1
U-Net	0.82	0.61	0.70
Mask R-CNN	0.67	0.52	0.58

5. Discussion

5.1. Comparison of different source images

For building extraction tasks, the higher the resolution of remote sensing images, the richer the feature details, making it easier to perform detection. However, very high-resolution (VHR) remote sensing images tend to be more expensive and may not be available in some areas. Therefore, to make the model applicable to lower resolution images, this study compares two different training strategies. The first one (model A) is to train the basic model on high-resolution images (GE images) and then fine-tune it using low-resolution images (GSI images). The second one (model B) is to train the model directly using low-resolution images (GSI images). Both models are trained with the same parameter settings and pre-trained models.

Table 4 shows the accuracy of the prediction results for the two models on the Setagaya test set. First, from the results of model A, the fine-tuned model has a negligible effect on the accuracy of the high-resolution images. However, the accuracy is reduced by about nearly 30% on the low-resolution images. In comparison, the results of model B at different resolutions are lower than those of model A. In addition, from the prediction results shown in Figure 8, Model A can identify more buildings for images of different resolutions compared to Model B. The results show that the model trained on high-resolution images can thoroughly learn the target features and, therefore, has a higher generalization ability for image resolution.

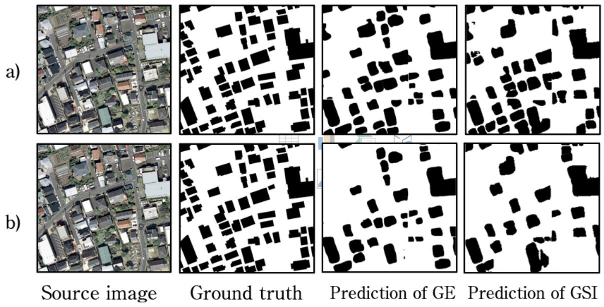


Figure 8. Prediction results of different training strategies.

(a) Prediction of model A. (b) Prediction of model B

Table 4. Object-wise accuracy of different training strategy.

Model	Image	mPrecision	mRecall	F1
Model A	GE image	0.65	0.52	0.57
	GSI image	0.51	0.37	0.43
Model B	GE image	0.53	0.32	0.40
	GSI image	0.45	0.28	0.34

5.2. Comparison of different areas

In order to test the extraction performance of the model for buildings in different regions, the prediction results of Setagaya, Hachioji, and Mashiki were compared. These three areas represent Japan's urban areas, suburban areas, and rural areas, respectively. Due to the lack of high-resolution remote sensing images in the Mashiki area, GSI images were used for the test set in all three areas.

The precision accuracy is shown in Table 5. As can be seen, the Setagaya has the highest accuracy, followed by Hachioji, and the Mashiki is the worst. The overall accuracy decreases as the density of buildings in the area decreases. Although the number of training data of Hachioji in the training set is the most, it contains mainly urban areas such as the station front, which means the model learns the features of urban buildings. The test set of Hachioji wanted to serve as a representative of suburban areas in Japan, so the images selected were mainly of more remote areas, which led to poor accuracy of Hachioji's prediction results. Besides, the poor results of Mashiki are due to the insufficient number of training sets in the rural area.

Figure 9 shows the results of building extraction for each region randomly. The Hachioji area is heavily vegetated, and the buildings are scattered and small and medium-sized buildings. Since the buildings are confused with the surrounding information, the vegetation cover significantly influences the building extraction. The building coverage in the Setagaya is dense and concentrated, with less influence of vegetation cover. Although most of the building areas were extracted correctly, the building boundaries were not finely divided. Buildings are scarce in Mashiki areas, but farmland can easily be confused with buildings. For example, open space and a swimming pool are mistakenly extracted as buildings in the figure.

In order to further explore the extraction of buildings in different types of areas, the buildings in the ground truth are divided into three categories according to the area: 1) small building: Area ≤ $150m^2$ 2) medium building: $150m^2 \leq$ Area $\leq 450m^2$ 3) large building: Area ≥ $450m^2$. Figure 10 shows the detection ratios for various sizes of buildings in different areas. The inability of the model to identify adjacent groups of buildings as one building resulted in a detection rate greater than one for large buildings in each region and medium-sized buildings in Setagaya. Similarly, the model has a low recognition rate for small and medium-sized buildings, especially non-urban areas. This is in urgent need of improvement in the future.

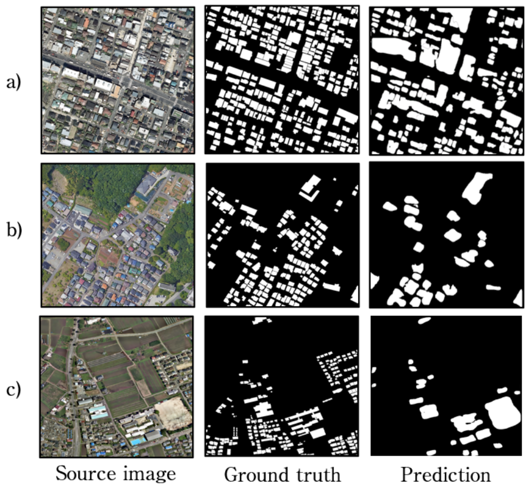


Figure 9. Prediction results of the different areas.

(a) Setagaya. (b) Hachioji. (c) Mashiki

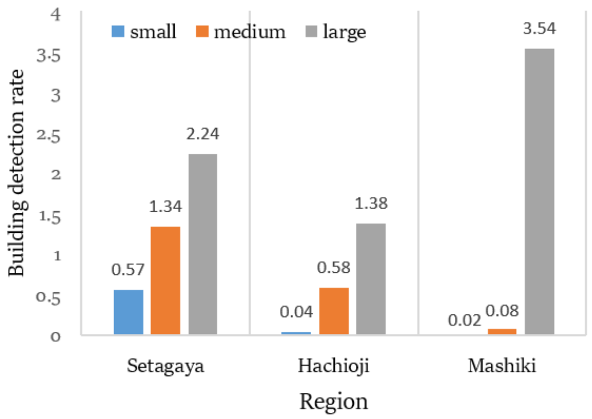


Figure 10. Various sizes of buildings detected in different regions

Table 5. Object-wise accuracy of different areas.

Area	Precision	Recall	F1
Setagaya	0.51	0.37	0.43
Hachioji	0.40	0.21	0.30
Mashiki	0.25	0.02	0.04

6. Conclusion

In order to expect a more efficient and robust implementation of city-wide building extraction, this study attempts to perform transfer learning to train a Mask R-CNN model, but the accuracy is unsatisfactory. Several problems include a wide range of missing detection, rough building outline, and undivided boundary, especially for small and medium-sized buildings. At the same time, the generalization ability of the model for different resolution images and buildings in different regions also needs to be enhanced. The main reasons for this are the insufficient training set, the influence of hyperparameter settings, the defects of the network model, which should be addressed in future research. We intend to use open-sourced building extraction datasets to train a pre-trained model for the network in the future. Then, we will review related literature and modify the default hyperparameter settings and model structure of Mask R-CNN to make it more adaptable to the multi-scale features of buildings and remote sensing images. Meanwhile, we will count the prediction results and visualize them in a 500-meter mesh so that the worst accuracy areas are identified for targeted expansion of the training set.

Acknowledgment

We are very grateful to the Dr. Hiroyuki Miyazaki of the Center for Spatial Information Science (CSIS), the University of Tokyo, for providing the training set data and part of the code used in this study.

References

Deshapriya, N. L., Dailey, M. N., Hazarika, M. K., & Miyazaki, H. (2020). Vec2Instance: Parameterization for Deep Instance Segmentation. *arXiv preprint,* arXiv:2010.02725.

Weir, N. (2019). The good and the bad in the SpaceNet Off-Nadir Building Footprint Extraction Challenge [Medium]. Retrieved from https://medium.com.

Ball, J. E., Anderson, D. T., & Chan Sr, C. S. (2017). Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community. *Journal of Applied Remote Sensing*, 11(4), 042609.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 1097-1105.

Farabet, C., Couprie, C., Najman, L., & LeCun, Y. (2012). Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(8), 1915-1929.

Mnih, V. (2013). Machine learning for aerial image labeling. University of Toronto (Canada).

Zhao, W., Du, S., & Emery, W. J. (2017). Object-based convolutional neural network for high-resolution imagery classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(7), 3386-3396.

Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., & Atkinson, P. M. (2018). An object-based convolutional neural network (OCNN) for urban land use classification. *Remote sensing of environment*, 216, 57-70.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431-3440).

Ji, S., Wei, S., and Lu, M. (2018). Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1), 574-586.

Shao, Z., Tang, P., Wang, Z., Saleem, N., Yam, S., & Sommai, C. (2020). BRRNet: A fully convolutional neural network for automatic building extraction from high-resolution remote sensing images. *Remote Sensing*, 12(6), 1050.

Liu, Y., Zhou, J., Qi, W., Li, X., Gross, L., Shao, Q., ... & Li, Z. (2020). Arc-net: An efficient network for building extraction from high-resolution aerial images. *IEEE Access*, 8, 154997-155010.

Liu, H., Luo, J., Huang, B., Hu, X., Sun, Y., Yang, Y., ... & Zhou, N. (2019). DE-Net: Deep encoding network for building extraction from high-resolution remote sensing imagery. *Remote Sensing*, 11(20), 2380.

Elad, M. (2002). *On the origin of the bilateral filter and ways to improve it*. IEEE Transactions on image processing, 11(10), 1141-1151.

Zhang, Z., & Zou, S. (2011). An improved algorithm of mask image dodging for aerial image. *In MIPPR 2011: Remote Sensing Image Processing, Geographic Information Systems, and Other Applications*, 80060S.

He K, Gkioxari G, Dollár P, et al (2017). "Mask R-CNN." *Proceedings of the IEEE international conference on computer vision*, 2961-2969.

Hosang, J., Benenson, R., & Schiele, B. (2017). Learning non-maximum suppression. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 4507-4515.

Wu Y, Kirillov A, Massa F, Lo W-Y, Girshick R (2019) Detectron2. https://github.com/facebookresearch/detectron2

Weir, N., Lindenbaum, D., Bastidas, A., Etten, A.V., McPherson, S., Shermeyer, J., Vijay, V.K., & Tang, H. (2019). SpaceNet MVOI: A Multi-View Overhead Imagery Dataset. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 992-1001.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *In International Conference on Medical image computing and computer-assisted intervention*, 234-241.