Developing of approach for building extraction from open-sourced satellite image based on instance segmentation approach with watershed algorithm

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

Abstract: Accurate acquisition of building footprints from remote sensing data is essential for urban planning and construction. However, due to the diversity of buildings, such as scale, colour, and texture, it is challenging for automatic building extraction from complex scenes. This paper proposed a deep-learning-based framework to obtain a detailed building distribution map in instance-level from open-sourced satellite images. The approach extracts the building footprints from the fused data based on an improved Mask R-CNN network, followed by a watershed algorithm for separating connected buildings. We chose three different 1-km^2 areas (urban, suburban and rural) as the test area to verify the proposed approach's effectiveness. Compared with other input sources, the watershed algorithm, which used distance transform with the local maxima method, could effectively separate connected buildings while avoiding over-segmentation.

Keywords: building extraction, remote sensing, instance segmentation, watershed algorithm

1. Introduction

The building footprint is a primary dataset in an urban geographic information system (GIS) database, which is essential for urban planning, map production, and population estimation. Due to the background's complexity (e.g., shadows and other artificial objects such as buildings) and diversity in the remote sensing image sources (e.g., different spatial resolution and capture time), establishing a reliable building database poses significant challenges.

In recent years, with the rapid development of deep learning technology (DL) in computer vision, deep learning models are gradually introduced into remote sensing. Many researchers have used semantic segmentation approaches to achieve efficient and automatic building extraction from remote sensing images (Shao et al., 2020). However, the study of building extraction methodologies is not limited to semantic segmentation. Most practical applications focus on whether a pixel is a building and classify different buildings individually to obtain the distribution and number of buildings, which is a typical instance segmentation task (Wagner et al., 2020).

Besides, in terms of data sources, high-resolution (HR) remote sensing imagery is not always available due to the limitation of technology and hardware costs. In that case, people have to use open-sourced satellite images alternative. In earlier research, we used the Mask R-CNN model to extract building footprints in instance-level for a wide area from 0.6-1 m open sourced satellite images. However, due to the low image resolution and vast variation of building in different regions, multiple buildings were detected as a single one, especially in dense residential and urban building areas (Chen, Ogawa, Zhao, & Sekimoto).

For the two-stage instance segmentation task, the semantic segmentation in the second stage is usually not accurate for locating regional boundaries, tending to cause edge bonding of adjacent areas. Traditional hybrid image segmentation algorithms have been shown to integrate different segmentation methods to help improve results in the edge part (Merabet, Meurie, Ruichek, Sbihi, & Touahni, 2015). Among them, the watershed algorithm is a mathematical morphology-based region segmentation method, which can provide closed contour lines when the target region has low resolution and weak edges. Therefore, based on this idea, we tried to use the watershed algorithm as post-processing to improve the problem of building adhesion. First, colour normalization and super-resolution were applied to improve open-source

images' quality. Then, an enhanced Mask R-CNN with a Multi-path Vision Transformer (MPViT) as backbone was used to extract the building footprint. Finally, the watershed algorithm was applied to separate connected buildings. To demonstrate the validity of the proposed framework, we chose three different areas with one km^2 (urban, suburban, and rural) and evaluated the extraction accuracy.

2. Related works

Since Ji, Wei, and Lu (2018) realized the extraction of buildings based on a Mask R-CNN model and achieved better pixel-wise accuracy than the FPN, the instance-based segmentation approach has proven its effectiveness for building extraction. There are two primary approaches for instance segmentation for building extraction: (1) semantic segmentation based, and (2) end-to-end instance segmentation. In (1), semantic segmentation first classifies each image pixel as either a building or non-building; the subsequent post-processing extracts the instances (Wagner et al., 2020). For example, Sirko et al. (2021) extracted 516M building footprints of Africa based on the U-Net model. They achieved the performance of instance segmentation by improving network architecture, loss functions and training strategy.

In (2), the problem is decomposed into feature extraction, boundary box regression, and mask prediction stages. Approach (2) is direct and flexible, allowing the algorithm to obtain boundary boxes and segmentation masks and perform better (Wen et al., 2019). Most state-of-the-art instance segmentation methods are end-to-end models, such as Mask R-CNN, MS R-CNN, and RefineMask. Therefore, some researchers have designed novel algorithms to extract buildings from remote sensing images based on state-of-the-art instance segmentation methods. Wu, Hu, Peng, and Chen (2020) proposed an anchor-free instance segregation method based on CenterMask and balanced building extraction accuracy and efficiency. Y. Wang et al. (2022) fused RGB images from an unmanned aerial vehicle (UAV) with visible light difference vegetation index (VDVI) features and Sobel edge detection features to improve the recognition accuracy of the Mask R-CNN model for rural building roofs. Liu, Liu, Ning, and Li (2022) proposed an MS-CNN model combing ResNet and Mask R-CNN, improving multi-scale feature extraction capability by using a fusion enhancement strategy and feature segmentation mechanism.

In summary, in order to obtain building footprint in instance-level, researchers often increase the complexity of the model or fusing multiple channels data to extract as much feature information as possible. However, it also increases the computational cost, while reducing the generality of the algorithm. In contrast, we went to the basics of image processing and morphological operations to disjoint the connected parts. The proposed method is independent of deep learning model architecture, which is simple to realize and can be integrated into any instance segmentation model.

3. Dataset

We selected representative areas from other parts of Japan as the training set, primarily located in Shinjuku, Setagaya, Hachioji, and Susono City. The data source of training data is Google Earth (GE), with a spatial resolution of 0.3 m. We collected 1,222 aerial ortho-colour images of dimensions 1,024 × 1,024 pixels and manually labelled over 186,000 high-quality building footprints. All photos were in TIF format with three channels (RGB), and the corresponding labels were annotated according to the COCO annotation format (Lin et al., 2014). The data were assigned to the training and validation sets in the ratio of 70%:30%, respectively. The examples of training data for different areas are shown in Figure 1. The number of image data for other regions is shown in Table 1. Parts of the training data could be available publicly (Miyazaki, 2022).

We selected the open-sourced seamless satellite image dataset for test data from the Geospatial Information

Authority of Japan (GSI)[1]. The GSI seamless photo dataset contains PNG images obtained from multiple sources with a resolution ranging from 0.6–1 m. All test images were mosaicked together and cropped to the dimensions of 1024 × 1024 pixels with 20% overlapping to avoid splitting buildings at the image edges. The defective parts were filled using black pixels.

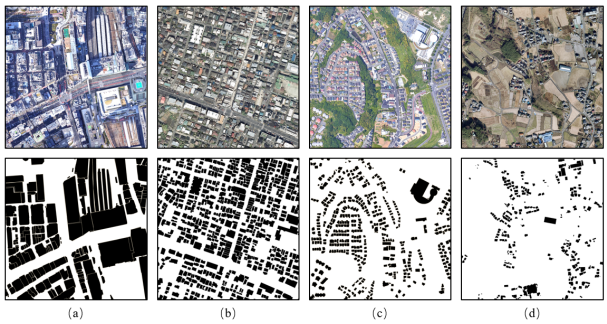


Figure 1. Example of training data in different regions, including a) high-rise, b) urban, c) suburban, and d) rural areas.

Table 1. Number of image data in different regions.

Area	Shinjuku	Setagaya	Hachioji	Susono
Training set	32	240	269	319
Validation set	8	103	115	136

4. Methodology

Figure 2 shows the three procedures constituting the proposed framework: (1) image pre-processing, (2) building extraction, and (3) result post-processing. First, test images are performed colour normalization and super-resolution to eliminate colour differences and enhance texture details. The pre-processed patches are subsequently input into the instance segmentation model to predict the building footprint. Finally, the prediction results in raster format are post-processed to separate connected objects through a watershed algorithm and generate a distribution map in vector format.

[1] URL: https://www.gsi.go.jp/tizu-kutyu.html (Accessed on 03/25/ 2022)

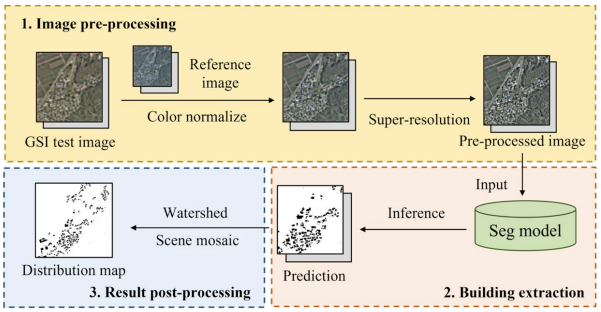


Figure 2. Proposed framework for building extraction with watershed algorithm.

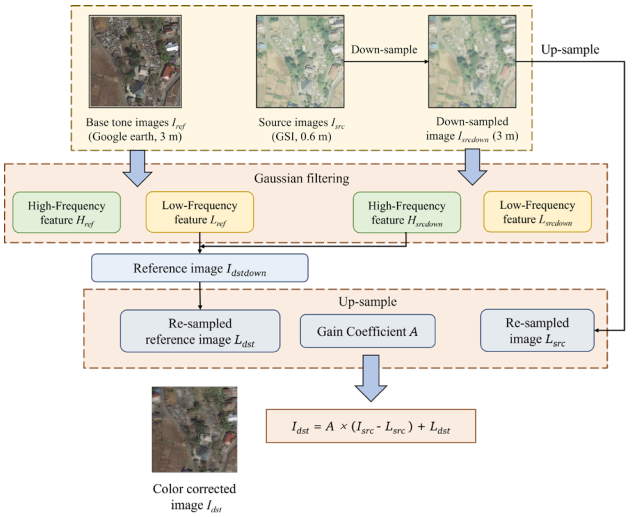
4.1 Image pre-processing

The GSI test images were sourced from multiple sources with a resolution range from 0.6-1 m. The color difference disrupts the visual continuity of the images and impedes the generalization ability of the building extraction model. Moreover, the lower image resolution makes the textures and outline details of the buildings blurry. Therefore, we first performed colour normalization and super-resolution to improve the quality of the test images.

Figure 3 illustrates the framework of the colour normalization algorithm based on a reference map. Analyzed from a spectral perspective, remote sensing images can be divided into high- and low-frequency information. High-frequency information is caused by the sharp excess of grayscale, primarily including the texture and edges of buildings. The low-frequency information is related to the grayscale component of image variations, which reflects the trend of image colour variations. As the low-frequency component of the image is smooth, the replacement of the low-frequency information will not interfere with the high-frequency information of the picture. Therefore, we considered the low-frequency information of the reference image as a standard to correct the tones of the test image (Cui, Zhang, Wang, Li, & Qi, 2021). We selected the 3 m resolution GE image as the reference map in this study.

For super-resolution, we adopted the Real-ESRGAN

network. (X. Wang, Xie, Dong, & Shan, 2021), which proposes a high-order degradation approach by introducing two times the number of degradation parameters as traditional degradation (blur, noise, resize, and image compression). Figure 4 shows the architecture of the Real-ESRGAN network, which contains a generation network and a u-net discriminator. The generative network is the same as that of ESRGAN. In addition, the U-Net discriminator can judge the authenticity of generated images from the pixel perspective, focusing on the details of the generated image while ensuring overall authenticity. Figure 5 shows an example of a test image after pre-processing. It is shown that the tone of the processed image is close to that of the reference image, and the texture detail of the building has been significantly improved.



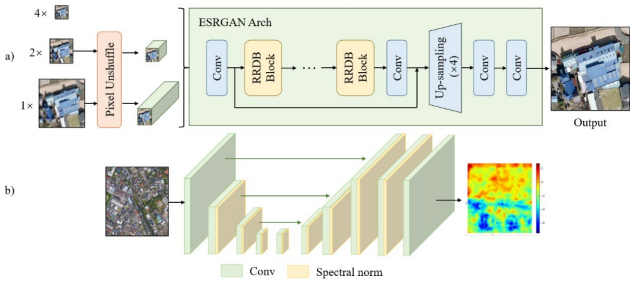


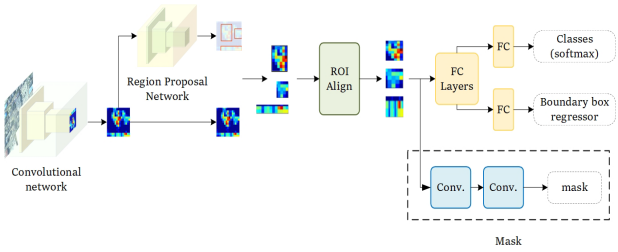
Figure 5. Example of pre-processed image.

4.2 Instance segmentation model

This study combines Mask R-CNN and a multi-path vision transformer (MPViT) as the instance segmentation model for building extraction. (He, Gkioxari, Dollár, & Girshick, 2017; Lee, Kim, Willette, & Hwang, 2021) Mask R-CNN is a widely used, powerful deep learning model. Figure 6 shows the network architecture of Mask R-CNN. The input images are first sent to ResNet for feature extraction. The backbone feature map is passed through the region proposal network (RPN) to extract the possible ROI. This ROI is mapped into a 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 fully convolutional layer upsamples the other branch to obtain the segmented region image.

MPViT follows a unique approach to multi-scale patch embedding and multi-path structures compared to other Transformers (Figure 7). A four-stage feature hierarchy to generate feature maps at various scales is constructed. The blocks of the proposed multi-scale patch embedding (MS-PatchEmbed) and multi-path Transformer (MP-Transformer) are stacked at each stage. By applying convolution operations with overlapping patches, the MS-PatchEmbed layer uses fine and coarse-grained visual tokens at the feature level. Then, the tokens embedding features of varied sizes are fed into the MP-transformer encoder separately. In this transformer block, global self-attention is performed by each transformer encoder with various patch sizes. The generated features are aggregated to allow for fine and coarse feature representations. In the feature aggregation step, a global-to-local feature interaction (GLI) process is introduced to connect the local convolutional features to the global features of the

Transformer, leveraging the local connectivity convolution and the global connectivity context of the Transformer.



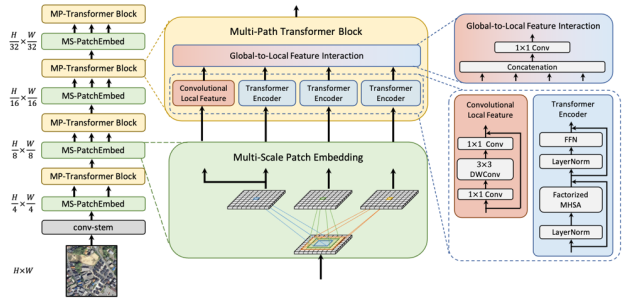


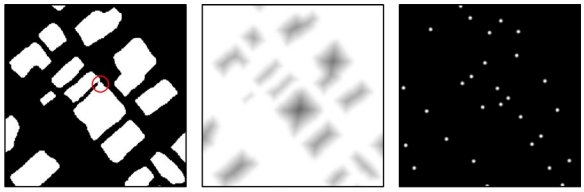
Figure 7. Architecture of the MPViT.

4.3 Results post-processing

The output of the instance segmentation model is the mask binary map in raster format. Besides, to avoid edge effects in the prediction results, there was a 20% overlap between the test images. Therefore, in post-processing, the raster data is used as input and stitches predictions into a vector format building distribution map.

4.3.1 Watershed algorithm

Figure 8a) is smaller than the area of the building roof or the entire building shape itself, it is possible to ensure that the local maximum is located within the building area and not within the connection section. The distance transform is used as the cost map (Figure 8b), and the local maximum of the distance (Figure 8c) is used as the sink points/markers to separate connected buildings.



(a) Connection between buildings (b) Distance transform (c) Local maxima

Figure 8. Illustration of the improved watershed algorithm.

4.3.2 Scene mosaicking

Figure 9. The method first sorts the masks in the overlap according to their areas, calculates the intersection-over-union (IoU) between them, and eliminates the masks with smaller areas if the IoU is more significant than a threshold.

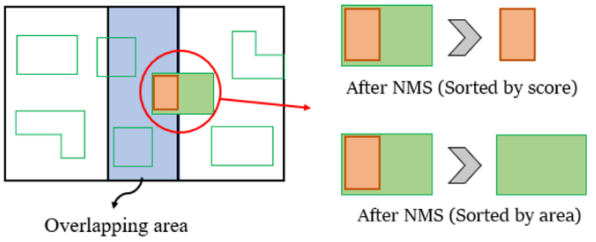


Figure 9. Effect of NMS strategies in scene mosaicking.

5. Experiment

5.1 Implement details

The training of the Mask R-CNN model with an MPViT backbone was implemented through Detectron2 (Wu et al., 2019). The maximum number of iterations was set to

30000 using a multi-scale training strategy, which resizes the input such that the shorter side is between 900–1100, whereas the longer side is ≤ 1333. Additionally, the AdamW optimizer was used for network optimization with an initial learning rate of 0.0001 and weight decay of 0.05. We selected a batch size of 16 and an initial learning rate of 0.0001, which decays 10 times at the iterations of 24000 and 27000. All the parameters were initialized according to the orthogonal distribution. Furthermore, we employed random cropping, flipping, and random rotation to avoid overfitting the augmentation model. The experiments were implemented on a large-scale platform called mdx in 4 Nvidia A100 GPU (40 GB) (Suzumura et al., 2022).

5.2 Evaluation Criteria

To quantitatively evaluate the performance of building extraction and usage classification algorithms, this study uses the mean recall, precision, F1 scores and accuracy to assess the prediction results, as shown in Equation (1). The TP (True Positive) and FP (False Positive) indicate the correct and incorrect building detections, respectively. As illustrated in Figure 10, we used an IoU threshold of 0.5 to classify the predictions as correct (TP).

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN} \qquad (1)$$

$$\text{F1} = \frac{2 \times Precison \cdot Recall}{Precision + Recall}$$

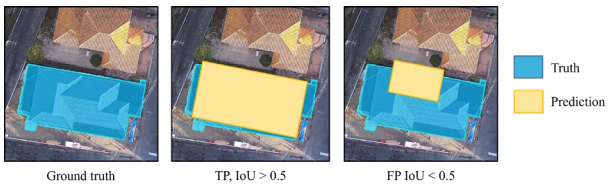


Figure 10. Schematic representation of the object-wise metric.

5.3 Results of building extraction

Buildings and other essential features in different areas show different grey values, texture, density, and size characteristics. Therefore, to evaluate the accuracy and robustness of the prediction results, we randomly selected three 1-km^2 areas for rural, suburban, and urban regions in Hyogo Prefecture and manually created ground truth as test data. All test images were upsampled 2× through super-resolution. To compare the effectiveness of the applied watershed algorithm, the result of the model without post-processing was applied for comparison.

Figure 11 shows the predicted results of two models and the corresponding RGB image and ground truth of the three areas. For the sake of clarity, some areas have been selected for enlargement. The watershed algorithm refines the segmented building boundaries, leaving watershed lines inside the buildings' segmented results. The building adhesion problem is somewhat alleviated, as shown in the yellow circles in Figure 11a, b. This illustrates that the watershed algorithm is effective in the segmentation of structured buildings or low-density built-up areas. For the high-density building areas in Figure 11c, the model can identify most buildings and separate them with slightly lower accuracy. However, it is challenging to separate small buildings from large connected areas. This is because we assume that the connected parts are usually smaller than the building itself. This assumption does not hold when the buildings are small and dense, especially in urban areas. This can also be obtained from the quantitative results in

Table 2; the best result is shown in bold. The model with a watershed algorithm can achieve a slightly better performance than the baseline model in terms of recall and F1 in rural and suburban areas. However, the precision of all regions is lower than the baseline model. This is because the watershed method is based on the distance transformation of the building. Building with complex structures or small internal connections will be mistakenly divided into multiple parts, as indicated by the red circles in Figure 11. In urban areas, this phenomenon is more common. Over-division of buildings is more frequent compared to the correct division of connected parts, which causes the accuracy in urban areas to decrease after applying the watershed algorithm.

Table 2. Object-wise metric of different test areas

Area	Model	Precision	Recall	F1
Rural	MPViT-b + WS	0.748	**0.652**	**0.697**
	MPViT-b	**0.750**	0.647	0.695
Suburban	MPViT-b + WS	0.836	**0.755**	**0.793**
	MPViT-b	**0.840**	0.748	0.791
Urban	MPViT-b + WS	0.767	0.625	0.689
	MPViT-b	**0.780**	**0.639**	**0.694**

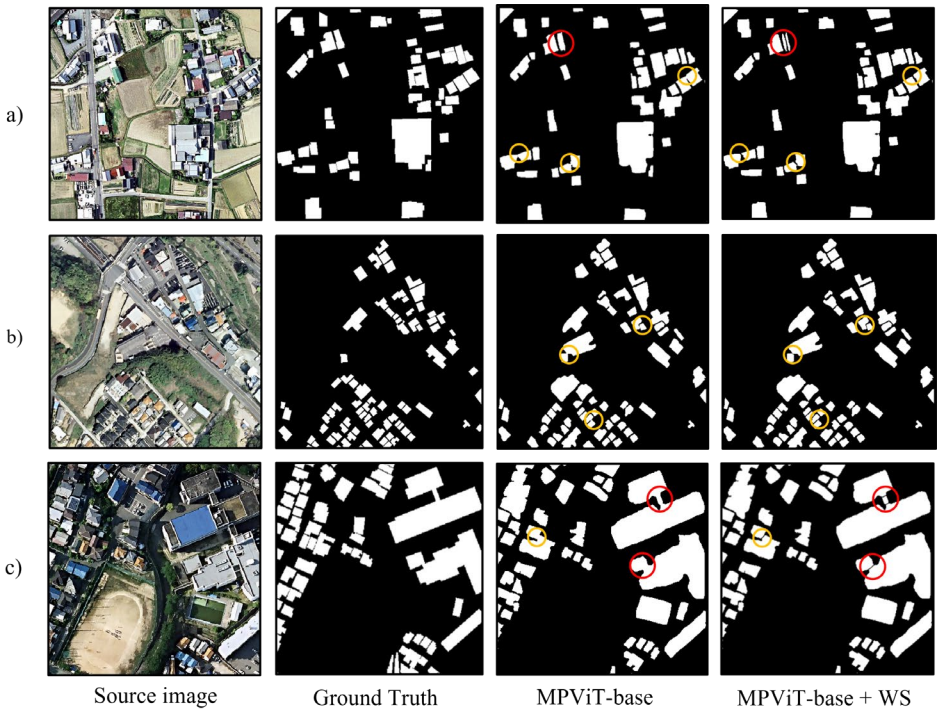


Figure 11. Prediction results of two models in (a) rural, (b) suburban, and (c) urban areas

6. Discussion

6.1 Cost map of watershed algorithm

The watershed algorithm segments the gradient image based on the input cost map. The catchment basin is the dark area, and the watershed line will stop at the outline of the dark part of the gradient image. In the traditional watershed algorithm based building extraction, the grayscale map of remote sensing image or Digital surface model (DSM) is generally chosen as the input image. The grayscale map of the image provides the true outline and brightness information of the building while DSM data provides direct access to the height information of the building. In this paper, in order to improve the drawback that the watershed algorithm is easy to over-segment, the distance transform diagram is chosen as the cost map. To verify the effectiveness of this improvement, we tested the effect of different input sources on the test data of suburban area, which are 1) Grayscale image, 2) DSM data, and 3) distance transform diagram. The DSM data were acquired from the Hyogo prefecture-wide DSM dataset released by the Ministry of Land, Infrastructure, Transport and Tourism of Japan, with a resolution of 1m [2]. Table 3 shows the quantitative results of different cost maps, where the distance transform diagram has the best result of all indicators. It could be also be illustrated in . Since the local maxima of grayscale images depend on the roof colour, the same building may have multiple ones. The watershed algorithm not being able to give the actual boundary of each building but lies somewhere between the boundaries. This can lead to over-segmentation even for

[2] URL: https://www.geospatial.jp/ckan/dataset/2010-2018-hyogo-geo-dsm (Accessed on 07/28/2022)

small buildings. The results using DSM data successfully eliminated small connected parts, but for roofs with significant slopes, it is easy to split a circle around the extreme value point into islands. This may be because the DSM data is not accurate enough, or due to the obscuration of objects such as trees. Also, some buildings may have adjacent roof ridges, which results in no local maximum inside the building, only a large local maximum. This is also not conducive to the division of adjacent buildings.

Table 3. Object-wise metric of different cost maps.

Cost map	Precision	Recall	F1
Grayscale image	0.743	0.718	0.730
DSM	0.775	0.745	0.760
Distance transform	**0.836**	**0.755**	**0.793**

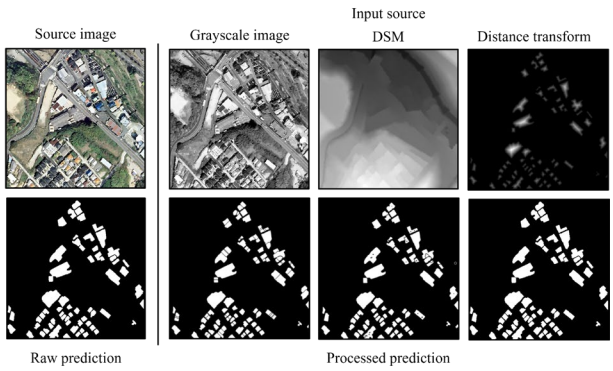


Figure 12. Processed prediction of different cost map of the watershed algorithm.

6.2 Marker of watershed algorithm

The marker is the initial area of the catchment basin, which is forced as a minimum value. By providing a marker for each object and background and making it the catchment area base, the desired object can be split from the background. The marker-based watershed algorithm utilizes the gradient information, the building itself and the background information of the detected building area, and can effectively extract the closed contour of the building, which has the advantages of continuous and closed boundaries. In this section, we compared two marker determination strategies, a thresholding-based method and a local-maxima-based method. The thresholding method sets the part greater than the threshold as the foreground which does not participate in the segmentation. Thus, the watershed operation is performed on the unknown region of the part smaller than the threshold. The local maxima method finds the local maxima in the building and uses it as the starting point of the watershed algorithm.

Table 4 and Figure 13 shows the quantitative and qualitive results of different marker generation strategy. The distance transformation with the thresholding method brings worse results. This is because the distance transform assigns more weight to large patches and less weight to smaller patches and edge pixels. Therefore, for large connected buildings, even the weight of the connection is greater than the weight of the small building itself. Therefore, applying thresholding removes not only connections but also small buildings. In contrast, the distance transform with the local maximum method used in this paper effectively separates connected buildings while avoiding over-segmentation, proving the method's effectiveness. However, this method does not take full advantage of the RGB data of the original image. And it is challenging to separate small buildings with large connected areas. We plan to overcome these problems in future work.

Table 4. Object-wise metric of different markers.

Marker	Precision	Recall	F1
Thresholding	0.695	0.629	0.660
Local maxima	**0.836**	**0.755**	**0.793**

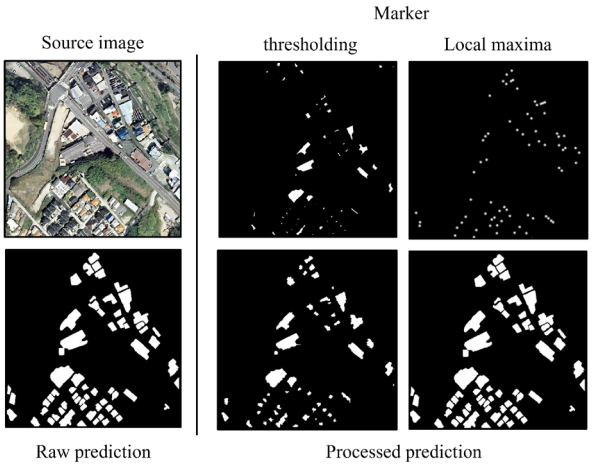


Figure 13. Processed prediction of different marker of the watershed algorithm.

7. Conclusion

We proposed a framework for building extraction based on an instance segmentation approach with a watershed algorithm suitable for open-sourced satellite images. The framework consisted of three main steps: image pre-processing, building instance segmentation and result post-processing. For open-sourced satellite images, pre-processing using colour normalization based on reference maps and super-resolution could improve the image quality and enhance textural details. Then, the Mask R-CNN model with an MPViT backbone was applied to extract the building footprint at the instance level. Finally, the watershed algorithm, which used distance transformation with the local maxima method, was applied post-processing for segmenting connected buildings. Compared with other input sources, the proposed watershed approach could effectively separate connected buildings while avoiding over-segmentation. However, it is hard to separate small buildings with large connected areas. Besides, buildings with complex structures in urban area tend to mistakenly be divided into multiple parts. In the future, we intend to overcome the above issues.

References

Bandara, R. (2018). Image segmentation using unsupervised watershed algorithm with an over-segmentation reduction technique. *arXiv preprint arXiv:1810.03908*.

Chen, S., Ogawa, Y., Zhao, C., & Sekimoto, Y. LARGE-SCALE BUILDING FOOTPRINT EXTRACTION FROM OPEN-SOURCED SATELLITE IMAGERY VIA INSTANCE SEGMENTATION APPROACH.

Cui, H., Zhang, G., Wang, T., Li, X., & Qi, J. (2021). Combined Model Color-Correction Method Utilizing External Low-Frequency Reference Signals for Large-Scale Optical Satellite Image Mosaics. *IEEE Trans. Geosci. Remote. Sens., 59*(6), 4993-5007.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). *Mask r-cnn.* Paper presented at the Proceedings of the IEEE international conference on computer vision.

Ji, S., Wei, S., & 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.

Lee, Y., Kim, J., Willette, J., & Hwang, S. J. (2021). MPViT: Multi-Path Vision Transformer for Dense Prediction. *arXiv preprint arXiv:2112.11010*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., . . . Zitnick, C. L. (2014). *Microsoft coco: Common objects in context.* Paper presented at the European conference on computer vision.

Liu, Y., Liu, J., Ning, X., & Li, J. (2022). MS-CNN: multiscale recognition of building rooftops from high spatial resolution remote sensing imagery. *International Journal of Remote Sensing, 43*(1), 270-298.

Merabet, Y. E., Meurie, C., Ruichek, Y., Sbihi, A., & Touahni, R. (2015). Building roof segmentation from aerial images using a line-and region-based watershed segmentation technique. *Sensors, 15*(2), 3172-3203.

Miyazaki, H. (2022). A dataset for detecting buildings, containers, and cranes in satellite images. In.

Sahu, M., & Ohri, A. (2019). VECTOR MAP GENERATION FROM AERIAL IMAGERY USING DEEP LEARNING. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences, 4*.

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.

Sirko, W., Kashubin, S., Ritter, M., Annkah, A., Bouchareb, Y. S. E., Dauphin, Y., . . . Quinn, J. (2021). Continental-scale building detection from high resolution satellite imagery. *arXiv preprint arXiv:2107.12283*.

Suzumura, T., Sugiki, A., Takizawa, H., Imakura, A., Nakamura, H., Taura, K., . . . Kobayashi, H. (2022). mdx: A Cloud Platform for Supporting Data Science and Cross-Disciplinary Research Collaborations. *arXiv preprint arXiv:2203.14188*.

Wagner, F. H., Dalagnol, R., Tarabalka, Y., Segantine, T. Y., Thomé, R., & Hirye, M. (2020). U-net-id, an instance segmentation model for building extraction from satellite images—Case study in the Joanopolis City, Brazil. *Remote Sensing, 12*(10), 1544.

Wang, X., Xie, L., Dong, C., & Shan, Y. (2021). *Real-esrgan: Training real-world blind super-resolution with pure synthetic data.* Paper presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision.

Wang, Y., Li, S., Teng, F., Lin, Y., Wang, M., & Cai, H. (2022). Improved Mask R-CNN for Rural Building Roof Type Recognition from UAV High-Resolution Images: A Case Study in Hunan Province, China. *Remote Sensing, 14*(2), 265.

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.

Wu, T., Hu, Y., Peng, L., & Chen, R. (2020). Improved anchor-free instance segmentation for building extraction from high-resolution remote sensing images. *Remote Sensing, 12*(18), 2910.