INSTANCE SEGMENTATION OF 3D MESH MODEL BY INTEGRATING 2D AND 3D DATA

W.X. Wang[1], G.X. Zhong[1], J.J. Huang[1], X.M. Li[1], L.F. Xie[1,*]

[1] Research Institute for Smart City, School of Architecture and Urban Planning, Shenzhen University – (wangwx, lixming, linfuxie)@szu.edu.cn, (2200325009, huangjunjie2020)@email.szu.edu.cn

Commission II, WG II/3

KEY WORDS: Instance segmentation,3D urban scene, Multi-view clustering, Mask-RCNN, 2D-3D Integrated Data.

ABSTRACT:

Buildings are an important part of the urban scene. In this paper, a novel instance segmentation framework for 3D mesh models in urban scenes is proposed. Unlike existing works focusing on semantic segmentation of urban scenes, this work focuses on detecting and segmenting 3D building instances even if they are attached and occluded in a large and imprecise 3D surface model. Multi-view images are first enhanced to RGBH images by adding a height map and are segmented to obtain all roof instances using Mask R-CNN. The 2D roof instances are then back-projected onto the 3D scene, the accurate 3D roof instances are obtained using a novel 3D clustering method and two post-processing steps which preserve the largest connected region and remove the model ambiguity. Finally, the 2D convex hull of each 3D roof instance is calculated and the model is divided within the range into building instances. The performance of the proposed methods is evaluated using real UAV images and the corresponding 3D mesh models qualitatively and quantitatively. Results revealed that the proposed method could effectively segment the model of the urban scenes and building instance is obtained, the over-segmentation masks can be clustered correctly into roof instances and the under-segmentation masks caused by image segmentation errors are eliminated.

1. INTRODUCTION

Buildings are an important dataset and foundation for the study of urban scenes. 3D reconstruction and modeling from images or range data of buildings, the most prominent man-made objects on the Earth's surface, has been a very active research area for the past three decades(Song et al., 2021; Haala and Kada, 2010; Shephard and Georges, 1992). Digital building models have a wide range of applications in urban planning, population density analysis, mobile communications, solar energy potential assessment, disaster management, 3D GPS navigation, and environmental simulation.

However, 3D models of urban scenes without semantic information will greatly limit their application scope. We focus on instance segmentation of buildings, rather than semantic segmentation, because it separates different building instances even if they are attached. Therefore, this paper aims to accurately and automatically segment all building instances in a large 3D urban scene.

In recent years, deep learning has relatively mature technology and framework in the field of image instance segmentation (Yekeen et al., 2020; Chen et al., 2023; Chen et al., 2017; Tajbakhsh et al., 2016; Liu et al., 2022; Oba and Ukita, 2020). Similarly, the direct application of deep learning technology to object instance segmentation of 3D scenes has also become a research hotspot(Yasir et al., 2022; Sanchez et al., 2020; Shen and Stamos, 2021; Qi et al., 2017b; Qi et al., 2017a; Xiong et al., 2015). However, the above papers are applied to object instance segmentation of indoor scenes and the data form is point cloud. The proposed method(Huang et al., 2022) uses LiDAR data to segment building instances. It first obtains the roof boundary, and then constructs the wall surface perpendicular to the roof boundary and the ground to obtain the building instance. Although a good result for building instances can be obtained in the end, the instance lacks realism, and the complex wall surface of the building instance is replaced by a simple vertical plane. Therefore, building instance segmentation in large urban scenes is still challenging.

Instead of directly segmenting 3D models, segmenting images first and projecting them to the 3D models is a potential alternative, as it can utilize powerful neural networks for image segmentation. At present, (Leotta et al., 2019) use multi-view satellite images to reconstruct the scene and obtain the orthophoto map and point cloud model, then perform building semantic segmentation on the orthophoto map and then project back to the point cloud model, finally generate the mesh model of the building instance. However, segmentation of building instances only on the orthophoto map often leads to under-segmentation, such as connected buildings cannot be segmented. (Yu et al., 2021) uses multi-view images to generate DSM, DOM, and orthophoto map, and use them to extract building boundaries with deep learning, depth map, and point cloud are used to obtain building elevation to construct large scene-building instances. However, two very close individual buildings in the extraction of building boundaries may be under-segmented into the same one, which will affect the generation of 3D building instances. Therefore, how to solve the over-segmentation and under-segmentation in the segmentation process will be an inevitable research focus.

We develop a novel framework for instance segmentation of urban buildings in large scenes. Based on the multi-view images captured by the Unmanned Aerial Vehicle (UAV), the roof instances are directly segmented. Through the spatial clustering algorithm proposed in this paper and other instance optimization processing, the roof instances and building instances on the 3D model are obtained. The technical framework of this paper is as follows：

[*] Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-1-W2-2023-1677-2023

1. The RGB image is enhanced to RGBH image, the instance segmentation is performed on the multi-view images, and the instance segmentation result is back-projected onto the 3D model. The roof semantic segmentation mask is generated by the instance segmentation result, and the 3D semantic mask is obtained for the subsequent roofing instance error elimination and optimization.
2. A novel clustering algorithm is applied to roof clustering to obtain original roof instances. The clustering algorithm can effectively eliminate the under-segmented instance masks, such as multiple roofs identified as one roof, and perform spatial clustering for over-segmented masks. The original roof instance results are clustered from multiple over-segmented roof masks.
3. Due to the error of the image segmentation mask, the basic units on the 3D model are ambiguous. Therefore, the original roof instances obtained by clustering will have obvious regions due to error accumulation. By retaining the largest connected area of the roof instance and filtering with the roof semantic mask, the non-roof regions can be eliminated to obtain the roof instances with better results.
4. Extract the 2D convex hull of the refined roof instance, segment the area within the convex hull, and the segmentation result is the building instance.

2. METHODOLOGY

2.1 Experimental Area and Data

The area selected for this experiment is an urban village area in Longhua District, Shenzhen, China, and the size of the experimental area is about 0.12 km. The multi-view images acquisition is obtained by DJI Phantom 4 RTK, with a total of 127 photos, and the Context Capture software is used to reconstruct the 3D model scene. The 3D model format used here is the Mesh model. The experiment scene is shown below. It can be seen from the reconstruction results that the Mesh model is degraded to a certain extent, such as incorrect adhesion and holes between buildings. The main reason is that the surround shooting is not specified when collecting the images of this area, so not enough images are collected, resulting in the lack of multi-angle image data of buildings in the scene, and the final scene modeling effect is poor. For the subsequent image mask back projection to the Mesh model, the Aero Triangulation file (AT file) generated by Context Capture software is used to recover the position and orientation of the multi-view images to establish the correspondence between the images and the 3D model.

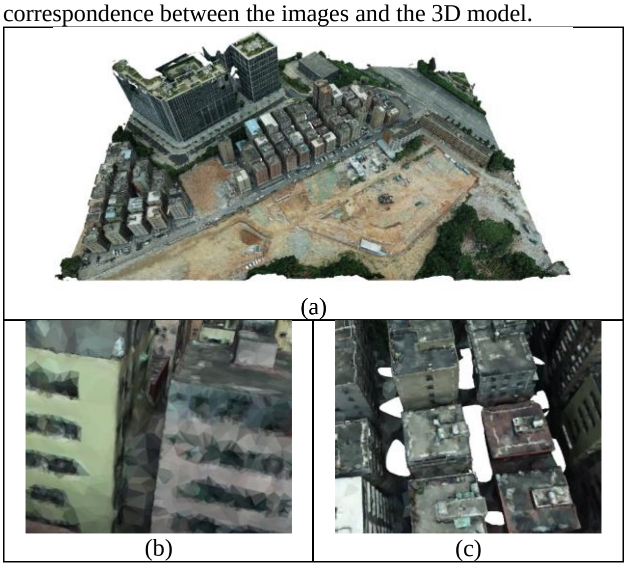
(a)
(b)
(c)

Figure 1. Experimental area and model errors

Vertices	Triangles	Area (km^2)	Images (resolution)
1,691,201	3,370,714	0.12	127 (5472×3648)

Table 1. Statistics on the 3D model

2.2 2D Roof Instance Segmentation with Height Map

Instead of directly segmenting 3D models, segmenting images first and back-projecting them to the 3D models is a potential alternative because it can leverage powerful neural networks for image segmentation. Orthophoto maps could be the first candidate because their projection direction is unity. However, buildings in orthophoto maps have severe self-occlusion, e.g. walls cannot be seen. Therefore, in the process of 2D-3D projection, its inaccurate segmentation will affect the classification of the 3D Mesh model. For this sake, we employ a multi-view 3D segmentation framework in this paper.

In this paper, we only segment the roof instance from the multi-view images. The main reason is that the buildings in the experimental area are densely distributed and the overall visibility of the buildings is low, this situation can be seen from Figure 2. It is impossible to directly segment the entire building instance through multi-view images and back-project it onto the Mesh model to obtain the entire building instance. Therefore, we first choose the roof for its good visibility to segment the building, and then obtain the roof instance by back-projecting it back onto the Mesh model, ultimately segmenting the complete building instance.

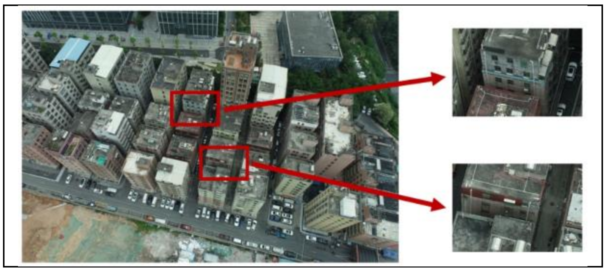

Figure 2. Building distribution in the experimental area

Here, the image segmentation method we use is the mask regional convolutional neural network (Mask R-CNN). Mask-RCNN model was developed in for semantic segmentation, object localization, and object instance segmentation. To avoid dividing other objects into roof instances, the multi-view RGB image is enhanced to an RGBH image by adding an additional channel to encode height information. The geometric information is a very important supplement that can improve segmentation accuracy. The specific Mask R-CNN framework is shown in Figure 2. After the RGB image is enhanced to an RGBH image, it can effectively eliminate the ground area being divided into roof instances.

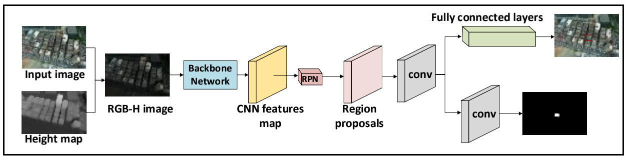

Figure 3. Mask R-CNN architecture with RGBH images

To avoid dividing other objects into roof instances, the multi-view RGB image is enhanced to an RGBH image by adding an additional channel to encode height information. The geometric information is a very important supplement that can improve segmentation accuracy. The specific Mask R-CNN framework is shown in Figure 3.

In this experiment, there are total of 127 images, 22 images are randomly selected as the training set of the model training, and 15 images are used as the test set of the model training. After the RGB image is enhanced to an RGBH image, it can effectively eliminate the ground object being divided into roof instances.

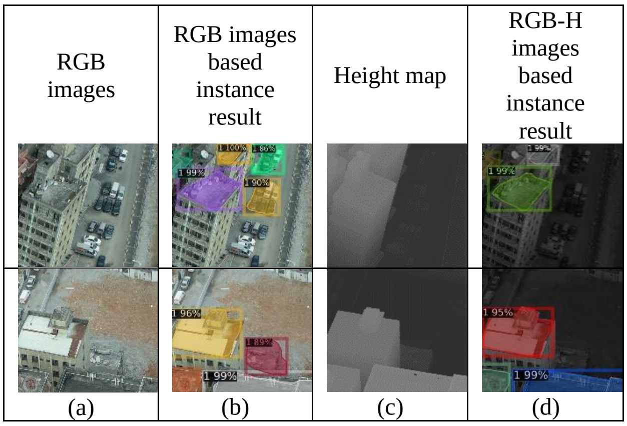
RGB images
RGB images based instance result
Height map
RGB-H images based instance result
(a)
(b)
(c)
(d)

Figure 4. Comparison of the segmentation results without and with height information

Figure 4. shows a visual comparison. Ground objects like cars and other things are successfully separated from the roofs, although some of them have a visually indistinct roof texture. The Mask R-CNN also computes a probability for each instance mask to represent its prediction confidence. To avoid masks with low confidence, only roof instance masks with predicted confidence higher than 70% were used.

2.3 Instance Ambiguity Removal

Although enhancing RGB images to RGBH images can effectively distinguish roofs from other objects, it is still inevitable to divide other areas into roof instances, e.g. building walls. Besides, the result of roof instances segmented using Mask R-CNN still have errors, with some instance results containing a small portion of the walls connected to them. Due to the small number of images used for training and testing of the instance segmentation model, there are errors in the roof instances judged and output by the Mask R-CNN model. Therefore, we need to calculate the ambiguity of the Mesh model triangles and remove triangles area that doesn't belong to the roof in the roof mask obtained by back-projection.

Firstly, the roof semantic segmentation images are obtained from the Mask R-CNN instance segmentation results, and the semantic segmentation mask of each multi-view image is back-projected onto the Mesh model. For each multi-view image, the ambiguity of each triangle on the Mesh model can be divided into three situations:
(1) Roof triangle: The triangle on Mesh which is the back-projected intersecting unit of the roof semantic image (foreground).
(2) Non-roof triangle: The triangle on Mesh which is the back-projected intersecting triangle of the non-roof semantic area (background).
(3) Unknown triangle: The triangle on Mesh which is not within the back-projection range of the roof semantic image (background).

Figure 5. shows the instance (a) and semantic (b) results of one image, and corresponding back-projection results (c) for one image. The Red triangles represent the result of the roof mask back-projection of this image, the green triangles represent the result of the non-roof mask back-projection of this image, and the remaining gray triangles represent regions that are not in the back-projection range of the image.

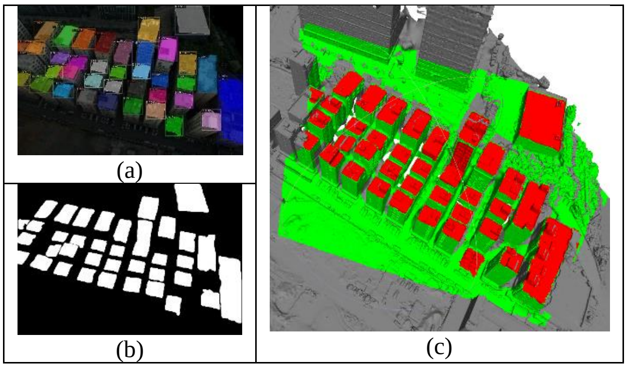

Figure 5. Instance results, semantic results, and back-projection results for one image

We back-project the roof semantic segmentation results of all multi-view images onto the Mesh model, recording the ambiguity of each triangle. For one Mesh triangle, it will be finally recorded as a roof triangle if the number of roof triangle records exceeds the number of non-roof triangle records; otherwise, it will be recorded as a non-roof triangle. In the end, the triangle on the Mesh model will be divided into two categories: the roof semantic triangle and the non-roof semantic triangle. Figure6. show the final semantic result of the roof after traversing through all the multi-view images. This work is aimed at optimizing the subsequent results of the roof instance mask, mainly solving the ambiguity problem of the Mesh triangle through the 2D-3D projection relationship.

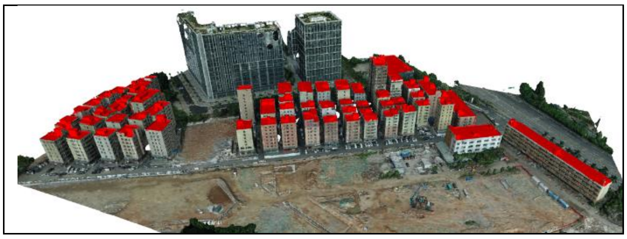

Figure 6. The semantic segmentation result of the Mesh model

2.4 3D Roof Instance Segmentation

In this paper, a unique roof instance is defined as the top cover outside a house or structure, that is each building corresponds to one roof. Subsequent definitions of building instances are synonymous. If two attached buildings have two roofs, they are considered to be two individual building instances.

Since roofs of the same building in multiple views have been segmented independently, the correspondences between roof instance masks are not known. This results in the number of instance masks being much larger than the number of roofs in the scene. Moreover, due to the problem that the dataset used for training and testing the image instance segmentation model is too small, the image roof instance mask has over-segmentation and under-segmentation. So we need to find the spatial relationship between instance masks and obtain the roof instance mask. There are two main steps. The first step is the spatial clustering of instance masks. The second step is to further generate realistic 3D roof instance masks from the over-segmented instance masks obtained in the first step.

2.4.1 Instance Mask Clustering

We propose an instance mask clustering method that divides instance masks into different groups each corresponding to a unique roof instance of an individual building. Representative masks are first selected from the segmented instance masks, and the remaining masks are merged with them according to mask similarity measures. For clarity, all roof instance masks in multi-view images are referred to as local masks, while representative masks selected for clustering are referred to as global masks since they represent unique building roofs in different images.

We first build a similarity matrix M to measure the spatial overlap for each pair of local masks. For the ith local mask, we record a set of triangles S_i whose centers are projected within this local mask region. A similarity matrix $M_{n\times n}$ is then computed to quantify the spatial overlap between every pair of local masks, where n is the number of all local masks. The similarity element m_{ij} measures the intersection over union (IoU) between the ith and the jth local masks, i.e.,

$$m_{ij} = A(S_i \cap S_j)/A(S_i \cup S_j), \quad (1)$$

where $A(S)$ is the surface area of the triangles in the set S. $M_{n\times n}$ is a symmetric matrix as $m_{ij} = m_{ji}$.

Generally, an ideal global mask should overlap most with the local masks corresponding to the same roof and least with the local masks corresponding to the roofs of different buildings. However we need to consider another situation that local masks with larger area are not always the ideal global masks. Segmentation errors of image instances may result in abnormally large areas of back-projection on the mesh model, such as two adjacent roofs being divided as the same roof instance. So we refer to the method (Chen et al., 2022) to estimate a confidence value C for each local mask to evaluate the overall overlap with all other local masks in the scene. The β parameter in the following formula is set to 0.5. More details about the evaluation of the parameter β can be found in the implementation details in Subsection 3.4.

$$C_i = P_i \cdot \sum_{j=1}^{n} \delta\left(m_{ij} - \beta\right) \cdot P_j \cdot m_{ij}, \quad (2)$$

where δ(·) is the delta function:

$$\delta(x) = \begin{cases} 0, & if\ x \le 0 \\ 1, & if\ x > 0 \end{cases} \quad (3)$$

Figure 7 shows that C values calculated from the above formulas for three examples of local masks, it can be found that the higher the completeness of the local mask, the larger the calculated C value, and the lower C value is generally the local mask with wrong image segmentation, such as containing walls.

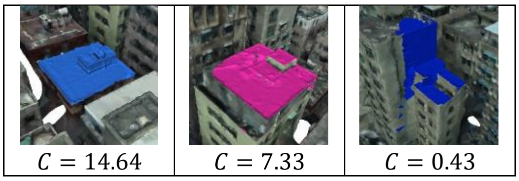

Figure 7. Samples of C values obtained from three local masks

2.4.2 3D Roof Instance Generating

One key observation of this work is that local masks with higher confidence values are consistent with other masks and thus should have higher priority to be selected as global masks. Based on the mask confidence, we employ a simple yet efficient order-based mask clustering. We first sort all local masks according to their confidence values C and then traverse them in descending order to select global masks. In the traversing loop, if a local mask has not been marked, we mark it as a new global mask, and other non-marked local masks whose similarities with this global mask are higher than β are considered consistent with this global mask， i.e, $\delta(m_{lg} - \beta) = 1$ where l and g are the indices of the local mask and this global mask, respectively. If a local mask has been already marked, we traverse to the next local mask. In general, a certain local mask will have a spatial intersection with other local masks. Therefore, we decide that the global mask should be clustered by the spatial intersection of at least two local masks and meet the criteria $\delta(m_{lg} - \beta) = 1$. The proposed method can spatially cluster the normal over-segmented local masks and eliminate the abnormal masks, in other words, the cases where the local mask does not have a spatial intersection with the rest of the local masks or does not satisfy $\delta(m_{lg} - \beta) = 1$， such as under-segmented masks are excluded here.

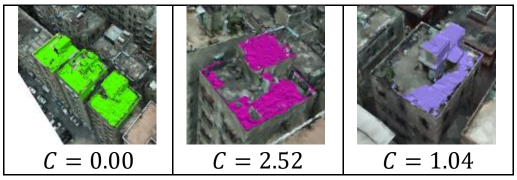

Figure 8. Samples of local masks that do not satisfy the clustering condition

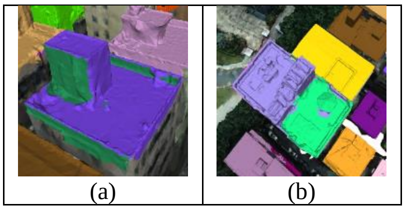

Figure 9. Samples of global mask result

Figure 8 shows local masks that do not satisfy the clustering conditions. The left sample shows the wrong segmentation (under-segmentation) of the roof instance. Three adjacent roof instances are divided into the same roof instance on the image, resulting in the local mask corresponding to three actual roof instances, although the local mask intersects with many other local masks, the C value is 0 due to the abnormally large local mask. The small roof integrity (over-segmentation) represented by the local mask in the middle and right results in an extremely small intersection area with the remaining local masks that have a spatial intersection, i.e. $\delta(m_{lg} - \beta) = 0$. So the C value is low. With the pre-computation of mask confidence values, the traversal is required only once. After one global local mask traversal, the global mask set is obtained. Due to the setting of the threshold β, there will still be cases of multiple global masks on a real roof instance, as shown in Figure 9, and this kind of case can be considered as over-segmentation of the roof instance mask. Therefore, we perform spatial clustering of global masks again, and the principle is to merge global masks if there is a spatial intersection, in other words, they contain the same triangle. In addition, to improve the clustering efficiency, we first calculate the bounding box(bbox) of each global mask, and only when the bbox of two global masks intersects, do we consider whether two global masks have spatial intersection to cluster. So far, the original roof instances are obtained.

However, the obtained roof instances still have errors, such as the current roof instance containing the triangles of ground, adjacent building wall and roof, and the wall where the building itself contains, which is mainly caused by the error accumulation of the previous local mask clustering. Therefore, we need to optimize the roof instance. Firstly, the largest connected area of the roof instance is retained to remove the triangles of the ground and adjacent building roof and wall. Then, according to the previously obtained roof semantic mask, the error triangles of the roof instance containing the wall of the buildings itself are removed. This optimization flow is illustrated in (a) to (c) of Figure 11. And the result of all refined roof instances in the scene is shown in Figure 11(d).

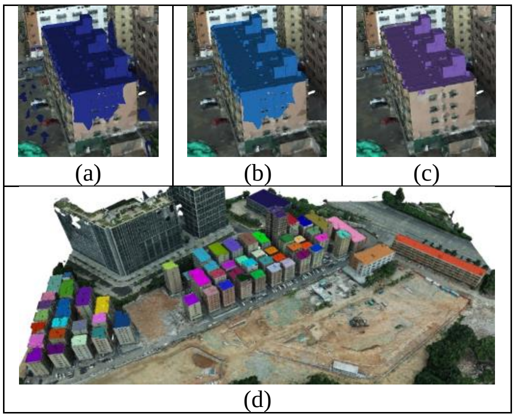

Figure 10. Sample of the optimization process for one roof instance and the result of all roof instances

2.5 Convex Hull-based 3D Building Segmentation

Based on 3D roof segmentation, the next step is to segment the entire 3D buildings. Here, we segment the whole building instance by constructing the 2D convex hull of the roof instance. The function of convex hull is that given a point set in a 2D plane, convex hull is a convex polygon formed by connecting the outermost points, which can contain all the points in the point set. Compared with constructing the oriented bounding box(obbox), the 2D convex hull can contain all the points more compactly and reduce the redundancy of the segmentation region. Here, we show a comparison of the results of the 2D convex hull of a roof instance with obbox, as shown in Figure 11. The black point set is the 2D point set of the extracted roof instance, the purple point set is the 2D convex hull point set of the roof instance, while the green rectangle range is the roof instance obbox. It can be found that the convex hull points can better wrap the roof point set. The main process of obtaining a building instance is as follows: Firstly, the 2D point set of the roof instance is extracted and the 2D convex hull is constructed. To segment an entire building from a 3D scene, we expand the boundary of a 2D convex hull by a certain offset value (1.5 meters in all of our experiments), is shown as the red point set in Figure 11 (b). After that, all triangles of the Mesh model are traversed to check whether the center points coordinates x and y of the triangle are within the range of the 2D convex hull of the roof instance. Finally, the segmentation and output of building instances were carried out.

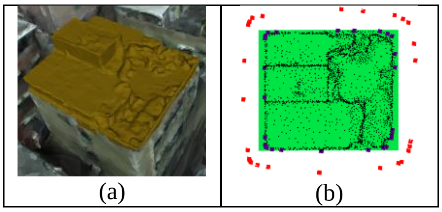

Figure 11. Sample of the roof instances and its corresponding 2D convex hull, oriented bounding box

3. RESULT

3.1 The Result of Building Instance Segmentation

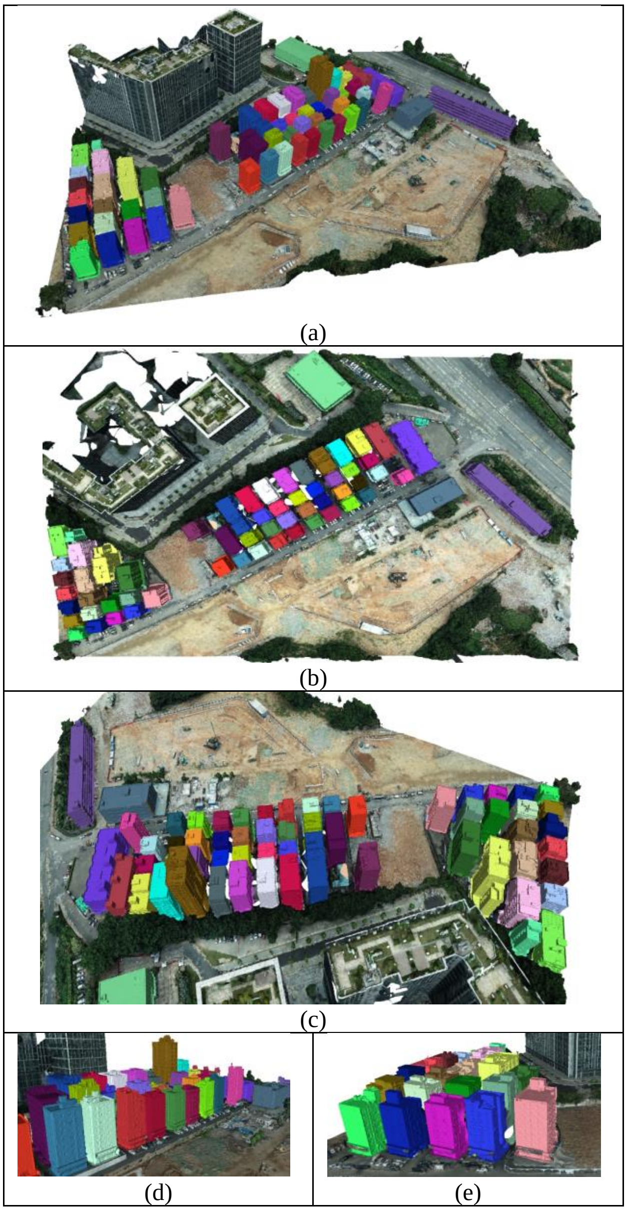

Figure 12. The result of building instance segmentation of the experimental scene

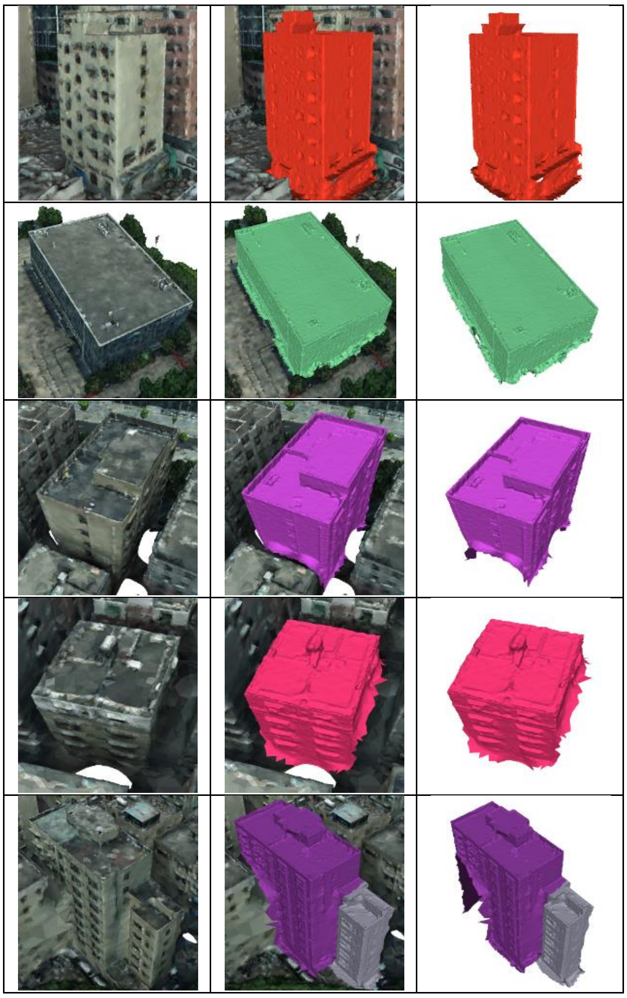

Figure 13. Samples of Building instance

Intermediate and Final Results	
Local Mask	1499
Global Mask	150
Original Roof Instance	62
Refined Roof Instance	59
Final Building Instance	59
Actual Building Instance	61

Table 2. Intermediate and Final Results of the experimental scene

It can be seen from the examples that, on the whole, the roof part of the building is well segmented, while some buildings cannot be completely segmented. For example, part of the walls of the third and fourth samples in Figure 13 are not segmented well. The main reasons are limited by the degradation of the Mesh model, and there are holes or incorrect connections between the buildings. At the same time, when a building is connected by two sub-buildings, this method can also segment the whole building well, it can be shown in the last sample of Figure 13.

3.2 The Necessity of Instance Ambiguity Removal

If the ambiguity of the Mesh triangles is not removed, there will be errors in the results of the generated roof instance. That is, the surrounding triangles area that does not belong to the roof will also be included, so that the 2D convex hull range of the generated roof instance in the later generation is incorrect, resulting in the final segmentation result of the building instance. As shown in Figure 14 (a) and (c), the segmentation of the roof and building instance without removing ambiguity is wrong, part of the adjacent building is also segmented into it. In contrast, Figure 14 (b) and (d) respectively represent the result of the roof instance and building instance with removing ambiguity.

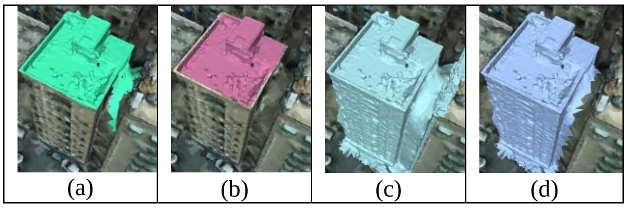

Figure 14. Comparison of instance segmentation results between removed and unremoved instance ambiguity

3.3 Instance Segmentation Limitation

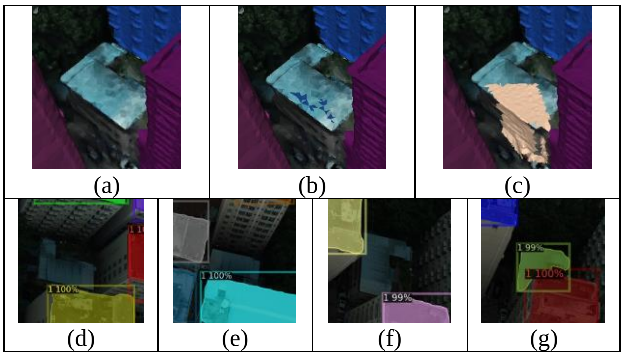

Figure 15. Sample of building Instance error. (a) Sample building in the scene. (b) Refined roof instance with removing ambiguity. (c) The result of building instance. (d)~(e) Recognition results of the building in multi-view images

The problem with this building instance is that instance has incomplete segmentation. As shown in Figure 15, the main reason for this wrong segmentation is that the location is recognized as a roof less times than it is recognized as a non-roof in the multi-view images, so the calculated semantic mask of the roof does not fully cover the roof, resulting in incomplete final roof segmentation and building segmentation error.

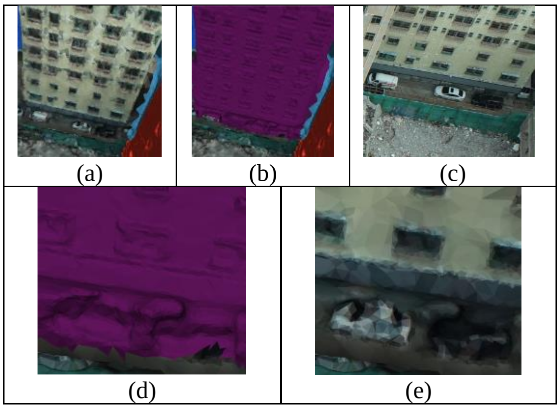

Figure 16. Sample of building Instance flaw. (a) Sample building in the scene. (b) The segmentation result. (c) Examples of images corresponding to the building. (d)~(e) Local detail segmentation result of the building.

As shown in Figure 16, this building instance contains part of the vehicles on the ground. The reason is that the first floor of the building is smaller than the 2D convex hull of the extended roof instance corresponding to the building, and the vehicle stops in the internal area of the building, so the final result includes the ground vehicles.

3.4 Effects of parameters

The method involves several parameters, of which β in the mask clustering step is the only parameter left adjustable to the user. This subsection will discuss how this parameter affects mask clustering.

Intuitively, the meaning of the β parameter in our work is very similar to the threshold parameter for IoU in many existing target detection efforts, where a mask is considered to be correctly predicted when the IoU between the detection mask and the true value is greater than this threshold. As a rule of thumb, this threshold is initially set to 0.5. Similarly, in mask clustering, two local masks should be considered as belonging to the same group if their IoU is greater than β. That is, they represent the same roof instance. In this work, we initially set $\beta = 0.5$ in all experiments. to determine the optimal value of β, we conducted experiments with different values. As shown in Table 3.

β-value	global masks	original roof instance	refined roof instance	precision
0.0	61	51	44	0.7213
0.1	75	61	53	0.8688
0.2	94	62	57	0.9344
0.3	99	63	59	0.9672
0.4	120	63	59	0.9672
0.5	**150**	**62**	**60**	**0.9836**
0.6	174	64	57	0.9344
0.7	219	63	58	0.9508
0.8	172	60	56	0.9180
0.9	11	9	7	0.1147
1.0	0	0	0	0.0000

Table 3. Influence of different β

This table shows the impact of different β values on the number of generated global masks, original roof instances, refined roof instances, and then the final building instance generation. The optimized judgment criterion for the number of roof instances here excludes the case of wrong segmentation of roof instances, as shown in Figure 15. The precision is calculated by dividing the number of segmented building instances by the number of actual buildings. Here, the number of actual buildings is counted by hand and the number is 61. We can see that the precision is highest when β equals 0.5. With the higher value of β, the number of global masks increases, which mainly increases the difficulty of merging local masks on the same roof, resulting in the situation that one actual building is segmented to two or more roof instances at the same time.

4. CONCLUSION AND FUTURE WORK

This paper presents a novel method for instance segmentation of 3D buildings based on Mesh model. Firstly, the 2D multi-view images with added height information are used to segment the roof instance, and then the 3D mask set is obtained by back-projecting the images' roof instance segmentation results onto the Mesh model. By constructing the mask clustering method, the 3D mask set is clustered to obtain the roof instances on the Mesh model. The spatial clustering method can still obtain the 3D correct roof instances despite the wrong segmentation results of the image. In addition, due to the errors in the instance segmentation results of the image, the roof instances obtained by the spatial clustering method will have some non-roof triangles due to the accumulation of errors, which will affect the subsequent extraction of building instances, so it is necessary to optimize the roof instances. By preserving the maximum connected area of the roof instance, the error triangles that are not connected with the roof triangles are eliminated, such as the roof of the adjacent building, the ground area, etc. Then, the 2D semantic mask of the roof of each multi-view images is obtained according to the segmentation results of the roof instance, and the 3D roof semantic mask is calculated by back-projection on the Mesh model. The semantic mask is used to eliminate the non-roof triangles connected with the roof instance, such as the wall surface connected with the roof. So far, the roof instance is refined. Finally, by calculating the 2D convex hull of the refined roof instance, the model in the range is segmented, and the building instance is finally obtained. The experimental results show that the proposed method can effectively segment accurate roof instances on 3D models with low accuracy, even though the building has two attached sub-buildings. However, limited by the accuracy of the Mesh model itself, the accuracy of some final building instances will be incomplete segmentation, which is inevitable.

4.1 Future Directions

In this paper, the structure of the buildings in the experimental area is relatively simple, therefore a good result can be obtained by constructing the convex hull for in-range model segmenting. However, further work is still needed for the complex structure of urban scene buildings such as classical buildings. In addition, the results of cutting through the convex hull still need to be optimized. The essence is to classify all triangles within the convex hull as building instances, so the non-building triangles in the segmented building instances need to be subsequently eliminated, such as using the Markov random field method. Finally, applying the method to 3D point clouds of urban scenes could also be an interesting future direction.

ACKNOWLEDGEMENTS

This work is supported by the National Key Research and Development Program of China (Projects Nos. 2022YFB3903700), the National Natural Science Foundation of China (42001407, 41971341, 41971354), the Shenzhen Science and Technology Program (Grant No. RCBS20221008093124064), the Guangdong Science and Technology Strategic Innovation Fund (the Guangdong–Hong Kong-Macau Joint Laboratory Program) (Project No.: 2020B1212030009).

REFERENCES

Chen, H., Qi, X. J., Yu, L. Q., Dou, Q., Qin, J., and Heng, P. A., 2017: DCAN: Deep contour-aware networks for object instance segmentation from histology images, Medical Image Analysis, 36, 135-146, 10.1016/j.media.2016.11.004.

Chen, J., Xu, Y., Lu, S., Liang, R., and Nan, L., 2022: 3-D Instance Segmentation of MVS Buildings, IEEE Transactions on Geoscience and Remote Sensing, 60, 1-14, 10.1109/TGRS.2022.3183567.

Chen, S., Ogawa, Y., Zhao, C., and Sekimoto, Y., 2023: Large-scale individual building extraction from open-source satellite imagery via super-resolution-based instance segmentation approach, Isprs Journal of Photogrammetry and Remote Sensing, 195, 129-152, 10.1016/j.isprsjprs.2022.11.006.

Haala, N. and Kada, M., 2010: An update on automatic 3D building reconstruction, Isprs Journal of Photogrammetry and Remote Sensing, 65, 570-580, 10.1016/j.isprsjprs.2010.09.006.

Huang, J., Stoter, J., Peters, R., and Nan, L., 2022: City3D: Large-Scale Building Reconstruction from Airborne LiDAR Point Clouds, Remote Sensing, 14, 10.3390/rs14092254.

Leotta, M. J., Long, C., Jacquet, B., Zins, M., Lipsa, D., Shan, J., Xu, B., Li, Z., Zhang, X., Chang, S.-F., Purri, M., Xue, J., Dana, K., and Ieee, 2019: Urban Semantic 3D Reconstruction from Multiview Satellite Imagery, 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, 2019
Jun 16-20, WOS:000569983600180, 1451-1460, 10.1109/cvprw.2019.00186.

Liu, Y., Wu, Y.-H., Wen, P., Shi, Y., Qiu, Y., and Cheng, M.-M., 2022: Leveraging Instance-, Image- and Dataset-Level Information for Weakly Supervised Instance Segmentation, Ieee Transactions on Pattern Analysis and Machine Intelligence, 44, 1415-1428, 10.1109/tpami.2020.3023152.

Oba, T. and Ukita, N., 2020: Instance Segmentation by Semi-Supervised Learning and Image Synthesis, Ieice Transactions on Information and Systems, E103D, 1247-1256, 10.1587/transinf.2019MVP0016.

Qi, C. R., Yi, L., Su, H., and Guibas, L. J., 2017: PointNet plus plus : Deep Hierarchical Feature Learning on Point Sets in a Metric Space, 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, 2017
Dec 04-09, WOS:000452649405018.

Qi, C. R., Su, H., Mo, K., Guibas, L. J., 2017, and Ieee: PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation, 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017
Jul 21-26, WOS:000418371400009, 77-85, 10.1109/cvpr.2017.16.

Sanchez, C. M., Zella, M., Capitan, J., and Marron, P. J., 2020: Semantic Mapping with Low-Density Point-Clouds for Service Robots in Indoor Environments, Applied Sciences-Basel, 10, 10.3390/app10207154.

Shen, X. K. and Stamos, I., 2021: 3D Object Detection and Instance Segmentation from 3D Range and 2D Color Images, Sensors, 21, 10.3390/s21041213.

Shephard, M. S. and Georges, M. K., 1992: Reliability of Automatic 3D Mesh Generation, Computer Methods in Applied Mechanics and Engineering, 101, 443-462, 10.1016/0045-7825(92)90033-g.

Song, J. W., Xia, S. B., Wang, J., 2021, and Chen, D.: Curved Buildings Reconstruction From Airborne LiDAR Data by Matching and Deforming Geometric Primitives, Ieee Transactions on Geoscience and Remote Sensing, 59, 1660-1674, 10.1109/tgrs.2020.2995732.

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., 2016, Hurst, R. T., Kendall, C. B., Gotway, M. B., and Liang, J. M.: Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?, Ieee Transactions on Medical Imaging, 35, 1299-1312, 10.1109/tmi.2016.2535302.

Xiong, B., Jancosek, M., Elberink, S. O., and Vosselman, G., 2015: Flexible building primitives for 3D building modeling, Isprs Journal of Photogrammetry and Remote Sensing, 101, 275-290, 10.1016/j.isprsjprs.2015.01.002.

Yasir, S. M., Sadiq, A. M., and Ahn, H., 2022: 3D Instance Segmentation Using Deep Learning on RGB-D Indoor Data, Cmc-Computers Materials & Continua, 72, 5777-5791, 10.32604/cmc.2022.025909.

Yekeen, S. T., Balogun, A.-L., and Yusof, K. B. W., 2020: A novel deep learning instance segmentation model for automated marine oil spill detection, Isprs Journal of Photogrammetry and Remote Sensing, 167, 190-200, 10.1016/j.isprsjprs.2020.07.011.

Yu, D., Ji, S., Liu, J., and Wei, S., 2021: Automatic 3D building reconstruction from multi-view aerial images with deep learning, ISPRS Journal of Photogrammetry and Remote Sensing, 171, 155-170, 10.1016/j.isprsjprs.2020.11.011.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-1-W2-2023-1677-2023 | © Author(s) 2023. CC BY 4.0 License.