3D Object Classification and Segmentation from Large Scale Area Reconstruction

Hye Sun Kim
CG/Vision Research Lab.
ETRI
Daejeon, Korea
hsukim@etri.re.kr

Yun Ji Ban
CG/Vision Research Lab.
ETRI
Daejeon, Korea
banyj@etri.re.kr

Chang Joon Park
CG/Vision Research Lab.
ETRI
Daejeon, Korea
chjpark@etri.re.kr

Ho Won Kim
CG/Vision Research Lab.
ETRI
Daejeon, Korea
hw_kim@etri.re.kr

***Abstract*—3D reconstruction is difficult to use in general applications because it treats objects as a whole without distinction. Post-processing to segment them into individual objects is essential. However, when the reconstruction target is a large-scale area, it is difficult to segment directly from the 3D geometry because it is divided into small rectangular tiles for convenience. We propose a method that can accurately classify and segment objects even in the case of locally divided 3D tiled geometry. Instead of using 3D segmentation, we adopt the method of 2D instance segmentation of a global multi-view image and then projecting it onto the 3D geometry. This solves the problem that objects located on the boundaries of tiles are fragmented and not properly segmented.**

Keywords—3D object segmentation, object classification, multi-view reconstruction, large-scale reconstruction

I. INTRODUCTION

The 3D reconstruction pipeline processes from 2D or 3D captured information without identifying or recognizing objects, resulting in a single all-inclusive mesh. To increase the usefulness and convenience of the result across a wider range of 3D applications, the first step is to separate and categorize the significant mesh into small, individual object meshes. Individual building meshes need to be accessed instead of a large mesh of the region to analyze the structure and distribution of buildings in a 3D digital cloned world. To be used in 3D content, the building mesh should be separated from doors and windows that the user can interact with.

Various methods have been proposed for segmenting 3D reconstruction results for individual objects. Historically, this has been accomplished through the use of color or shape continuity information of the object. Recently, object segmentation accuracy has significantly improved through the use of deep learning techniques. This paper proposes a deep-learning based method for 3D object classification and segmentation that can be used to reconstruct large-scaled regions using multi-view images.

II. RELATED WORKS

Object segmentation techniques can be classified based on their input as either 2D images or 3D geometries. 2D image segmentation has been studied for a longer period and has direct applications in various areas compared to the other methods.

Image segmentation can be classified into three categories: semantic segmentation labels each pixel with a class; instance segmentation identifies and segments individual objects; and panoptic segmentation merges the latter two based on their purpose[1]. In this paper, we utilized the instance segmentation method to distinguish individual objects. Extensive research has been conducted on deep learning models for instance segmentation. CNN models have been leading this research[2-4]. Recently, several transformer models have been published and they are competing for segmentation accuracy[5,6].

Several studies exist about segmenting scanned 3D point clouds[7,8] Experimental results suggest that these studies primarily aim to distinguish predefined object units or semantic parts within a room. Universal utilization of 3D point segmentation appears difficult due to limited training data.

III. 3D OBJECT CLASSIFICATION AND SEGMENTATION

Commercial 3D reconstruction software applications, such as VRICON and Context Capture, utilize segmentation techniques for 3D point clouds. However, it would be impossible to apply the same technique to reconstruct large-scale areas. Because of several computing limitations, we use a tiling process method to reconstruct a large-scale area. A tile-by-tile local segmentation method is not feasible since a single object is often divided into multiple tiles.

It has been determined that, for global object segmentation segmenting objects using multi-view images that were used as input to the reconstruction pipeline is the most suitable approach. Below is a figure that shows the entire process of object segmentation combined with 3D reconstruction.

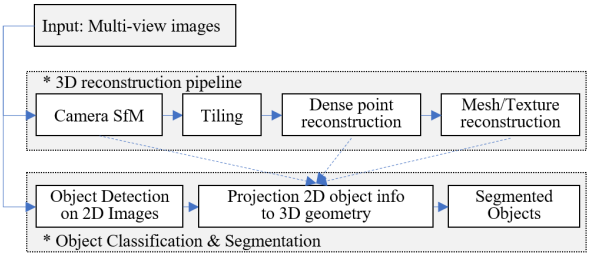

Fig. 1. System overview: 3D reconstruction pipeline and Object Classification/Segmentation

A. Large-scale Scene Reconstruction with Multi-view Images

Multi-view image-based 3D reconstruction generally includes the following steps[9]: the Camera Structure-from-Motion(SfM) method estimates the camera parameters of captured images; Multi View Stereo generates a dense point cloud; surface reconstruction meshes the point cloud; and mvs-texturing creates a texture map with images.

To reconstruct a large-scale area, an additional stage involves dividing the space into smaller tiles. Due to the size constraints, it is impossible to process a large, unbounded area

in one operation, so it is necessary to divide it into smaller, more manageable tiles. In this process, many times an arbitrary object is fragmented into multiple tiles, making it impossible to identify the object by looking at just one tile and extract it correctly. Figure 2 illustrates the process of dividing a large-scale area into 8x5 tiles, facilitating easier processing. The diagram also depicts a divided single tile. It's worth noting that the borders of the tiles are associated with fragmented objects.

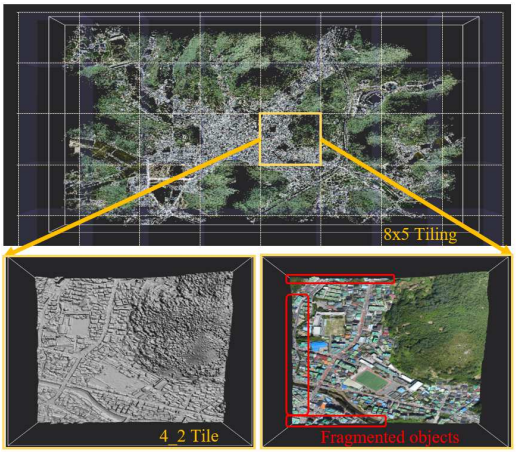

Fig. 2. Tiling and fragmented objects in each tiles

B. *Object Detection and Segmentation on 2D Images.*

Fragmented objects in neighboring tiles can impede 3D point-based segmentation methods, making it challenging to identify them as parts of the same object. Our method used a global approach to acquire detection information for object segmentation on multi-view images. Detectron2, developed by Facebook AI Research, is a software system that implements state-of-the-art object detection algorithms. It can proficiently segment and classify objects in 2D images based on the learned objects and classification schemes. Detectron2 provides numerous segmentation models, and we chose the Instance Segmentation model(Mask R-CNN)[2]. The model predicts class labels for all objects in the image and assigns a unique ID to each object. Since our objective is to segment 3D geometry, it is essential to have precise boundary information for each object.

The output in Figure 3 presents the result of a deep learning model trained on a specific object type and classification scheme. When analyzing all of multi-view images, the obtained detection and segmentation results for identifying objects are as follows.

Fig. 3. Object detection result images

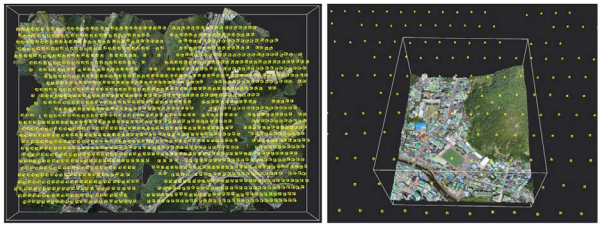

Fig. 4. 3D render view of camera SfM and selected cameras on a tile

C. *3D Object Segmentation by Projection to Images*

To integrate the 2D image segmentation information into the 3D reconstructed geometry, a projection method that exploits the camera parameters is used. The first step in the 3D reconstruction pipeline is camera SfM, which already has the camera's position and pose at the time of image capture, along with the intrinsic parameters required for projecting the image. Figure 4 displays the results of camera SfM using yellow 3D objects, indicating the position and pose of each input image's camera. The right image displays the selected cameras with the tile visible.

The projection of 3D geometry, such as points or meshes, onto the image, enables the computation of 2D coordinates based on the camera projection matrix. This is the computation result of the camera SfM that integrates the camera's position and orientation, along with the internal parameter information to convert 3D information into 2D coordinates in the camera's perspective. The values obtained from object detection on the projected 2D image coordinates can be integrated into the segmentation and classification information of the 3D geometry. This process is shown in Figure 5.

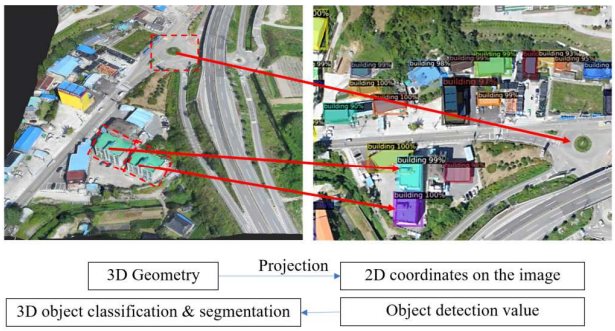

Fig. 5. Project 3D geometries onto 2D images and obtain the object detection information

3D geometry projection is performed on all visible multi-view images. Since a single object has been captured by a large number of cameras at different angles, the 3D segmentation of the object can be obtained. This process has the advantage of correcting for unrecognized object information due to errors in the 2D detection phase of some images.

The decision to use 3D points or meshes depends on the purpose of segmentation. In this project, meshes were chosen since they are easier to edit and manage in later stages. The triangular faces of the mesh are projected onto the image and the object detection information of the faces is obtained. Faces classified as belonging to the same object are collected together, and if they belong to different objects, they are separated and cut into different meshes.

IV. RESULTS

The experiment was conducted on a large-scale area measuring 3.59 km x 2.58 km and was divided into 8x5 tiles. 1659 4K images captured by a drone were utilized. Reconstructing and segmenting all at once is challenging due to the sizeable area and volume of images.

The primary objective of the application is to differentiate between the reconstructed large-scale terrain mesh of buildings and the ground. Therefore, the object detection has solely concentrated on detecting buildings. Figure 6 highlights the meshes categorized as ground in green and those categorized as buildings in blue.

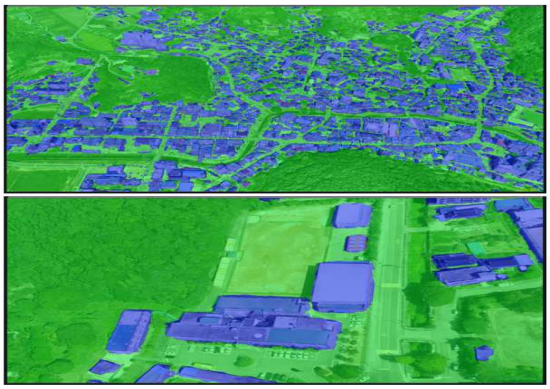

Fig. 6. The results of classifying 3D meshes into ground and building

As shown in Figure 7, the complete reconstructed mesh is partitioned into individual building and ground meshes. The upper image in Figure 7 shows the result of rendering only the building meshes, each segmented separately, while the lower image shows the result of segmenting the ground mesh into tiles and then smoothly interpolating the holes where the buildings used to be.

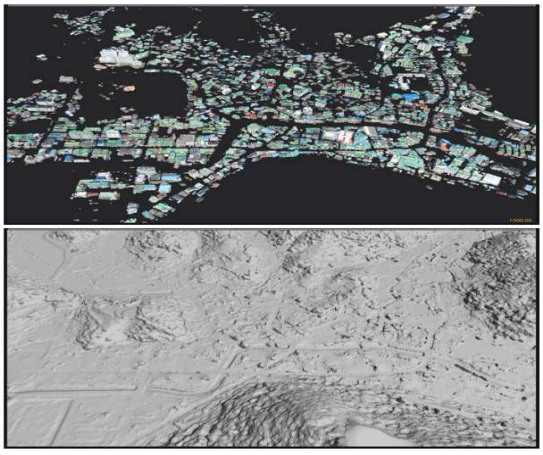

Fig. 7. The results of segmenting 3D meshes into ground and building

3D classification and segmentation accuracy were calculated for every face of the mesh. We calculated the ratio of the total number of correctly segmented mesh faces to the total number of mesh faces based on the true values, which have been segmented by hand. The segmentation accuracy level reached 90.4%. To enhance accuracy, we must precisely extract not just object detection but also object shape boundary information. While we continue working on improving accuracy, we have included the option to edit mesh face segments in a 3D mesh painting style.

V. CONCLUSION

This paper presents a technique to segment a tiled mesh acquired through 3D reconstructing a large-scale area into ground and building objects. To reconstruct an unbounded large area, it is necessary to divide it into smaller tiles that can be computed simultaneously. Applying a 3D segmentation method directly, however, is challenging due to the object fragmentation at the tile boundaries. To overcome this problem, we decided to use an instancing segmentation method to first segment all multi-view images and then project them onto the 3D geometry. Multi-view images can deal with the boundary fragmentation problem because they offer global information that is independent of the tile division. Additionally, the image segmentation approach benefits from simpler training data and provides higher accuracy.

ACKNOWLEDGMENT

This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2022-2023(Project Name: Development of large-scale space cloning and high quality real-time reconfiguration technology for realistic content, Project Number: R2020040207, Contribution Rate: 50%) (Project Name: Development of NeRF-based Hyper-Realistic 3D Character Creation Technology Project Number: RS-2023-00227819, Contribution Rate:50%).

REFERENCES

[1] X. Feng, Y. Jiang, X. Yang, M. Du and X. Li, “Computer vision algorithms and hardware implementations: A survey” in Integration, vol 69, November 2019, pp. 309-320

[2] K. He, G. Gkioxari, P. Dollar and R. Girshick, “Mask R-CNN” Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2961-2969.

[3] D. Bolya, C. Zhou, F. Xiao and Y. Lee, “YOLACT: Real-time Instance Segmentation” Computer Vision and Pattern Recognition(CVPR), 2019

[4] X. Wang, T. Kong, C. Shen, Y. Jiang and L. Li, “SOLO: Segmenting Objects by Locations” Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2020

[5] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin and B. Guo “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows” Computer Vision and Pattern Recognition(CVPR), 2021

[6] J. Jain, J. Li, M. Chiu, A. Hassani, N. Orlov and H. Shi, “OneFormer: One Transformer to Rule Universal Image Segmentation” ” Computer Vision and Pattern Recognition(CVPR), 2022

[7] C. R. Qi, L. Yi, H. Su and L. J. Guibas, “PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space” Computer Vision and Pattern Recognition(CVPR), 2017

[8] M. Xu, J. Zhang, Z. Zhou, M. Xu, X. Qi and Y. Qiao, " Learning Geometry-Disentangled Representation for Complementary Understanding of 3D Object Point Cloud”, Computer Vision and Pattern Recognition(CVPR), 2018

[9] Z. Liu, Z. Xu, C. Diao, W. Xing and D. Lu “Benchmarking large-scale multi-view 3D reconstruction using realistic synthetic images”, Proc. SPIE 11373, Eleventh International Conference on Graphics and Image Processing (ICGIP 2019), 113732N (3 January 2020)