3D Point Cloud Segmentation Using 2D Image Segmentation

1st Dinesh Channabasappa Dhotrad
dept. of Computer Science
Case Western Reserve University
dxd539@case.edu

2nd Ravi Raj Kumar
dept. of Computer Science
Case Western Reserve University
rxk739@case.edu

3rd Praneeth Kollati
dept. of Computer Science
Case Western Reserve University
pxk505@case.edu

***Abstract*—This paper presents a novel approach to 3D point cloud segmentation by leveraging 2D image segmentation techniques. Traditional methods that directly process point clouds are computationally intensive due to the unstructured nature of the data. Our method utilizes the spatial correspondence between 2D image pixels and 3D point coordinates, facilitated by accurate camera calibration parameters, to efficiently segment 3D point clouds. We employ the state-of-the-art OneFormer model for 2D image segmentation into instance, semantic, and panoptic layers. The segmented 2D masks are then back-projected onto the 3D point cloud using a voting-based approach to infer 3D segment labels. Our method demonstrates significant reductions in computation time and resource usage compared to traditional deep learning techniques, without compromising segmentation accuracy. Experiments across multiple indoor and outdoor scenes showcase the robustness of our system and its potential for real-time applications**

I. Introduction

The segmentation of 3D point clouds into meaningful objects and regions is a fundamental task in computer vision with wide-ranging applications in autonomous driving, robotics, augmented reality, and beyond. However, traditional point cloud segmentation methods that directly operate on the 3D data are often computationally expensive due to the inherently unstructured nature of point clouds. Recent advances in deep learning, such as PointNet and PointFormer, have improved performance but still require substantial computational resources and large training datasets.

In this work, we propose a novel approach that leverages state-of-the-art 2D image segmentation techniques to indirectly infer 3D point cloud segmentation. Our key insight is to utilize the spatial correspondence between 2D image pixels and 3D point coordinates by incorporating accurate camera calibration parameters. Specifically, we employ the OneFormer model, a Transformer-based architecture that achieves state-of-the-art performance on 2D image segmentation benchmarks.

II. Related Work

In this section, we discuss related works in the field of semantic segmentation of 2D and 3D datasets. Other related works which employ 2D semantics for 3D segmentation are also discussed which is crucial to understand our work.

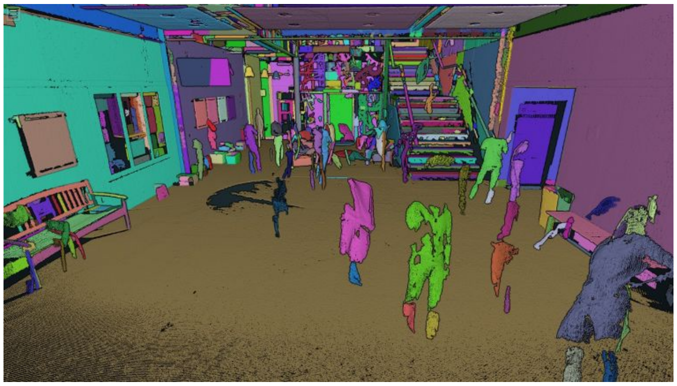


Fig. 1. 3D Segmented Point Cloud

A. Point Cloud Segmentation

Early work on point cloud segmentation relied on geometric primitives and hand-crafted features. The advent of deep learning ushered in a new era of data-driven approaches, with PointNet being one of the pioneering works to directly process raw point clouds using deep neural networks. Subsequent methods like PointFormer further improved performance by incorporating attention mechanisms. However, these approaches are computationally intensive and require large labeled datasets for training.

B. 2D Image Segmentation

Tremendous progress has been made in the field of 2D image segmentation, with techniques like Mask R-CNN and Transformer-based models achieving impressive results on benchmarks like COCO and Cityscapes. These models excel at segmenting both object instances and semantic regions, providing rich information that can be leveraged for 3D perception tasks.

C. 3D from 2D

There have been several efforts to infer 3D information from 2D data sources like images and videos. For instance, researchers have explored techniques for 3D object detection, 3D reconstruction and semantic scene completion by leveraging 2D cues and geometry. Our work is inspired by these concepts but focuses specifically on 3D point cloud segmentation using 2D image segmentation as an intermediate representation.

III. Methods

In this section, we will discuss the methodology employed in this project. The architecture shown in the figure we use the input Data in the form of RGB images, Depth map, and camera extrinsic values. RGB data undergoes segmentation from one-former, yielding both Semantic and Panoptic segmentation outputs, subsequently utilized for mask generation. The Depth map and camera extrinsic values inputs are used in 3D Processing block, responsible for generating an enhanced 3D point cloud. Subsequently, the 3D voting block leverages inputs from the mask block and the enhanced point cloud from the fusion block to establish correspondences between 2D image pixels and 3D points within the point cloud to generate 3d point segmentation.

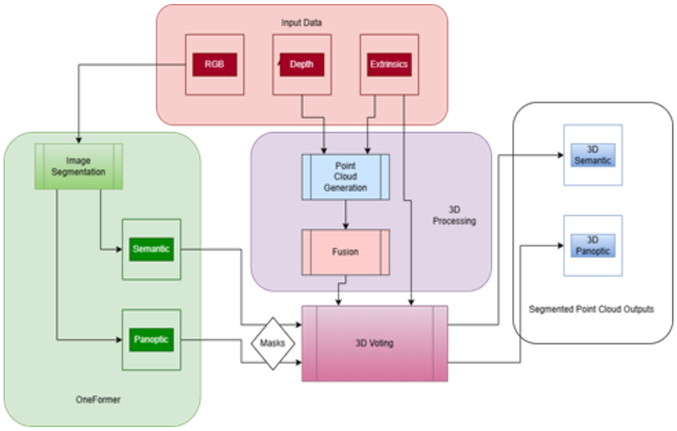


Fig. 2. Architecture Pipeline

A. Data Collection

Our data collection pipeline involves capturing RGB images, depth maps, and camera calibration parameters using an Apple iPhone 13 Pro equipped with a LiDAR sensor. The RTAB (Real-Time Appearance-Based Mapping) framework is utilized for real-time extraction and synchronization of these data streams. The iPhone 13 Pro's advanced camera system, including the LiDAR sensor, allows for accurate depth perception and 3D mapping capabilities. The RGB images capture the visual appearance of the scene, while the depth maps provide precise spatial information about the objects and surfaces within the environment. Additionally, the camera calibration parameters ensure accurate alignment and registration of the visual and depth data. The RTAB framework plays a crucial role in this pipeline by seamlessly integrating and synchronizing the various data streams in real-time. It employs advanced computer vision and robotic mapping algorithms to process the incoming visual and depth information, enabling the creation of a coherent and accurate 3D representation of the scanned environment.

B. Image Segmentation

We employ the OneFormer model, a state-of-the-art Transformer-based architecture for universal image segmentation tasks. OneFormer is pre-trained on the COCO dataset and can generate instance, semantic, and panoptic segmentation outputs for a given input image. In our pipeline, we primarily utilize the semantic and panoptic segmentation masks as inputs for the subsequent 3D point cloud segmentation stage.

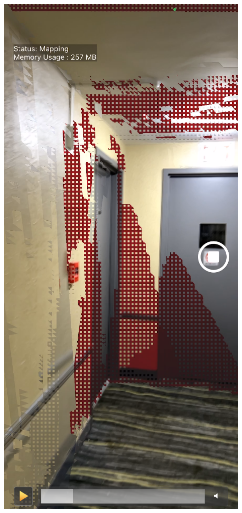
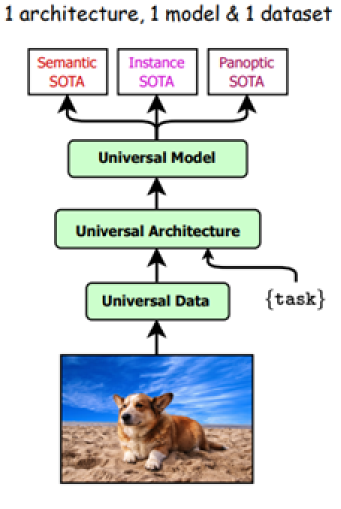

Status: Mapping
Memory Usage : 257 MB

Fig. 3. RTAB iOS Application



(c) Universal Architecture, Model and Dataset

Fig. 4. Oneformer Model

OneFormer, built upon the powerful Transformer architecture, represents a significant advancement in the field of image segmentation. Its pre-training on the large and diverse COCO dataset enables it to learn rich representations and semantic understanding of various objects and scenes. When an input image is fed into the OneFormer model, it generates three types of segmentation outputs: instance, semantic, and panoptic segmentation. In our pipeline, we leverage the semantic and panoptic segmentation outputs as inputs for the 3D point cloud segmentation stage. These masks provide valuable semantic context to guide and enhance the 3D segmentation, enabling accurate scene understanding and applications like object detection and robotic perception.

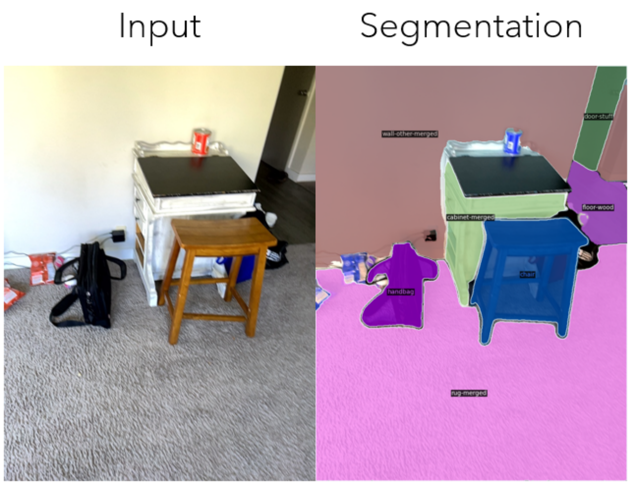


Fig. 5. Segmentation

The grayscale image undergoes thresholding to generate masks as shown in Fig. 6 capable of accommodating up to 256 classes within the 8-bit grayscale spectrum. These masks are then utilized for voting within 3D space, facilitating the analysis and processing of spatial data.

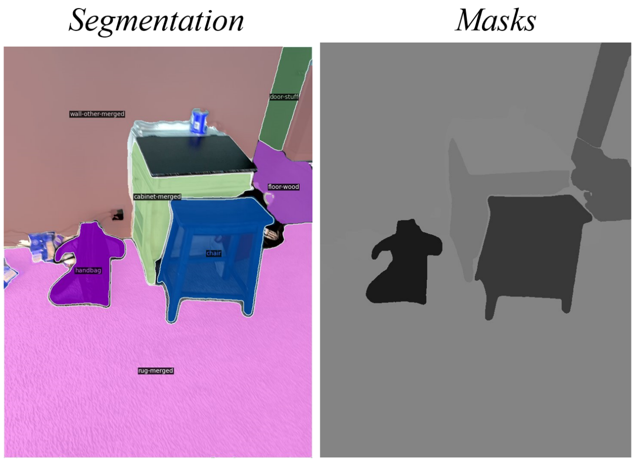


Fig. 6. Generated Masks

C. Point Cloud Fusion

To construct a comprehensive 3D point cloud from the captured data, we implement a fusion algorithm inspired by the COLMAP structure-from-motion pipeline [**b1**]. This involves merging individual point clouds from consecutive frames using geometric computations to align and combine the data. The fusion process also incorporates camera pose information and performs patch-based down sampling to ensure efficient processing.

The COLMAP (COLorized Mapping) structure-from-motion pipeline is a widely-used and robust framework for reconstructing 3D models from unordered image collections. Our fusion algorithm takes inspiration from this pipeline, leveraging its proven techniques for accurately aligning and merging point cloud data from multiple viewpoints. The fusion process begins by constructing individual point clouds from each frame of the captured data. These point clouds are then registered and aligned using geometric computations, which take into account the relative poses and positions of the camera during the data acquisition process. By accurately aligning the point clouds, the algorithm can seamlessly merge them, creating a unified and comprehensive 3D representation of the scanned environment. Additionally, the fusion algorithm incorporates camera pose information, which provides valuable insights into the camera's orientation and position during the data capture. This information helps to further refine the alignment and integration of the point clouds, ensuring a high degree of accuracy in the final 3D reconstruction. To ensure efficient processing and management of the potentially large point cloud data, our algorithm employs a patch-based down sampling technique.

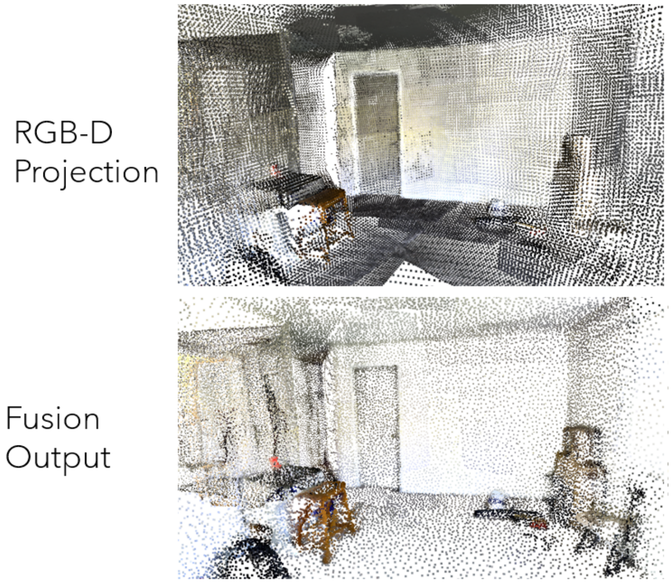


Fig. 7. Fusion Results

D. Voting-based 3D Segmentation

Our core contribution is a voting-based segmentation approach that utilizes the correspondence between 2D image pixels and 3D points in the point cloud. For each 3D point, we back-project its corresponding 2D image coordinates onto the point cloud using the camera intrinsic and extrinsic parameters. Each 3D point then accumulates votes from its corresponding 2D image pixels based on the segmentation labels provided by the 2D masks.

The accumulated votes represent the confidence of each point belonging to different semantic classes. We define a confidence threshold to determine the minimum required votes for a point to be assigned a specific semantic class. Points below this threshold are filtered out, and the remaining points are assigned semantic labels based on their highest accumulated votes. This voting-based approach effectively transfers the semantic information from the 2D image domain to the 3D point cloud, leveraging the rich contextual information provided by the 2D segmentation masks.

For panoptic segmentation, we combine the semantic and instance information to create a comprehensive representation that assigns unique identifiers to both stuff (e.g., road, sky) and thing (e.g., vehicles, pedestrians) classes. This unified representation provides a holistic understanding of the scene, capturing both the semantic categories and individual instances of objects. By fusing the semantic and instance information, our approach enables a more complete and nuanced interpretation of the 3D environment, paving the way for advanced applications in areas such as autonomous systems, augmented reality, and scene understanding.

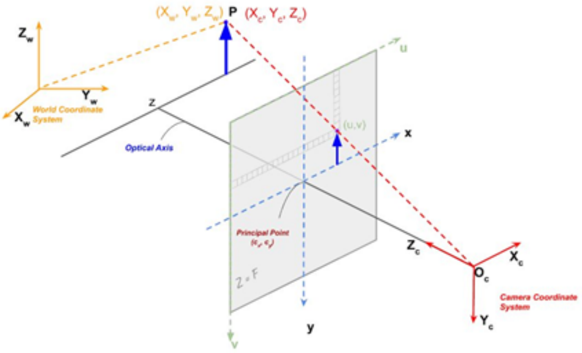


Fig. 8. 3D Space Projection

Through this voting-based segmentation approach, our pipeline effectively bridges the gap between the 2D image and 3D point cloud domains, leveraging the strengths of both modalities to achieve accurate and contextually rich segmentation results. The combination of robust 2D segmentation models and our novel 3D voting technique enables us to generate high-quality semantic and panoptic segmentation outputs for the reconstructed 3D point clouds, enabling a deeper understanding of the surrounding environment.

IV. Experiments

We evaluate our proposed method on multiple indoor and outdoor scenes captured using the iPhone 13 Pro LiDAR sensor. The experiments assess both the accuracy and computational efficiency of our approach compared to traditional point cloud segmentation methods like PointNet and PointFormer.

A. Datasets

We use a custom dataset collected from various indoor and outdoor environments, including buildings, streets, and parks. The dataset consists of synchronized RGB images, depth maps, and camera calibration parameters. Ground truth annotations for semantic and instance segmentation are manually labeled for a subset of the data to evaluate segmentation accuracy.

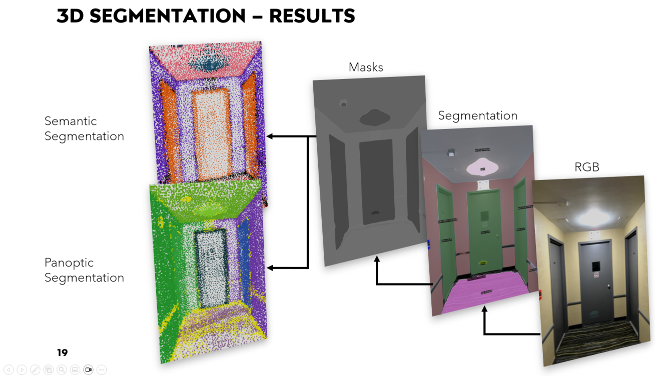

3D SEGMENTATION – RESULTS

Fig. 9. 3D Segmentation Results

B. Baseline Methods

We compare our approach against two baseline methods:

1) PointNet: A pioneering deep learning method for directly processing raw point clouds.

2) PointFormer: A more recent transformer-based approach that incorporates attention mechanisms for improved performance.

C. Results

Our experiments demonstrate that the proposed method achieves comparable segmentation accuracy to baseline methods like PointNet and PointFormer. Qualitative results showcase the effectiveness of our approach in accurately separating and labeling different semantic classes within the 3D environment.

Furthermore, we extensively analyze the computational efficiency of our pipeline, focusing on aspects like processing time, memory footprint, and scalability to large-scale point clouds. Our voting-based technique, which leverages rich semantic information from 2D image segmentation, offers significant computational advantages over traditional point cloud segmentation methods. The combination of qualitative analysis and computational efficiency highlights our method's practical value in the field of 3D point cloud segmentation and scene understanding for real-world applications.

V. Conclusion

We have presented a novel approach for 3D point cloud segmentation that leverages state-of-the-art 2D image segmentation techniques. By exploiting the spatial correspondence between 2D image pixels and 3D point coordinates, our method efficiently segments point clouds with reduced computational demands compared to traditional deep learning approaches. Experimental results across various scenes validate the accuracy and efficiency of our system, paving the way for real-time applications in domains such as autonomous driving and robotics.

References

[1] Colmap Fusion
https://github.com/colmap/colmap/blob/main/src/colmap/mvs/fusion.cc

[2] OneFormer: One Transformer to Rule Universal Image Segmentation

[3] https://github.com/SHI-Labs/OneFormer

[4] SEMANTIC ENRICHMENT OF 3D POINT CLOUDS USING 2D IMAGE SEGMENTATION A. Rai, N. Srivastava, K. Khoshelham, and K. Jain

[5] Determining the Proper Camera Position https://github.com/isl-org/Open3D/issues/2338

[6] Pixelwise View Selection for Unstructured Multi-View Stereo Johannes L. Sch¨onberger1 , Enliang Zheng2 , Marc Pollefeys1,3 , Jan-Michael Frahm2

[7] Intersection of a Ray and a Line Segment in 3D https://www.codefull.net/2015/06/intersection-of-a-ray-and-a-line-segment-in-3d/