Segment Anything Model (SAM) with ScanNet++ Point Cloud Data

Alejandro Torra I Benach Maximilian Summerer Matheus Ribeiro de Oliveira
Technical University of Munich

Abstract

In this work, we present an approach to bridge the gap between 2D image- and 3D point cloud instance segmentation, leveraging the state-of-the-art Segment Anything Model (SAM) [2] for mask generation. Our method entails the projection of point cloud data onto a spherical domain to generate 2D images, allowing us to employ SAM for robust instance segmentation. Subsequently, the segmented masks are employed to the 3D point cloud. Finally, k-nearest neighbors (k-NN) algorithm is applied to the segmented point cloud, providing a more comprehensive result. The proposed framework showcases the synergistic integration of advanced 2D segmentation techniques and 3D point cloud processing, facilitating improved scene understanding made possible with limited computing resources.

1. Introduction

The integration of 2D image- and 3D point cloud instance segmentation represents a pivotal challenge in computer vision and deep learning, with implications for diverse applications, such as robotics, autonomous driving, and augmented reality. In this work, we introduce a methodology that utilizes the capabilities of the Segment Anything Model (SAM) to generate precise and detailed masks for 2D images. Our approach involves the transformation of 3D point clouds from the ScanNet++ Dataset [5] onto a spherical projection, enabling the utilization of SAM for instance segmentation. By leveraging these segmented masks, we embark on a process of creating segmented 3D point clouds. This fusion of cutting-edge segmentation techniques with 3D point clouds provides a comprehensive solution for scene understanding, offering a high level of detail in complex real-world environments, with applications in several research fields.

Our main contirbutions are:

- Applying SAM after spherical projection of 3D point cloud data onto a 2D surface to generate 2D instance masks.
- Providing a solution for scene understanding by mapping 2D instance masks into 3D space to generate segmented point cloud data.
- Enabling the extraction of RGB information from 3D data and testing SAM's segmentation capabilities for distorted images from spherical projection with limited computing resources.
- Labeling unlabeled points due to occlusions caused by the spherical mapping.

2. Related Work

A groundbreaking addition to vision foundation models is the Segment Anything Model (SAM) [2] , introduced in 2023. Recognized as a milestone in the field, SAM possesses the remarkable capability to segment any object within an image, guided by diverse prompts that originate from user interactions. This model harnesses the power of a foundation model, extensively trained on the SA-1B dataset, composed by thousands of 2D images, which is currently the most extensive segmentation dataset available. SAM's versatility and proficiency mark it as a significant advancement in vision-based segmentation models. It can segment anything in an image using zero-shot learning, which means it can generalize across domains without additional training, and it can handle unfamiliar objects and images without requiring specific training for each case. An example of masks generated by SAM on 2D images can be seen on Figure 1.

In this project, however, we focus on the task of 3D instance segmentation, where the objective is to predict labels for individual points that form different instances within a 3D point cloud representing a scene. The work from Yang et al. [4], SAM3D, leverages the SAM model on RGB-D data, creating masks from 2D images and projecting the masks onto 3D using depth information. The adjacent point clouds are merged through "bidirectional-group-overlap-algorithm", adapted from the work of Hou et al. [1]. The process is iteratively repeated for point clouds of adjacent frames until the 3D masks of the whole scene is obtained. Finally, the result is merged with the over-segmentation masks to obtain an ensembled result.

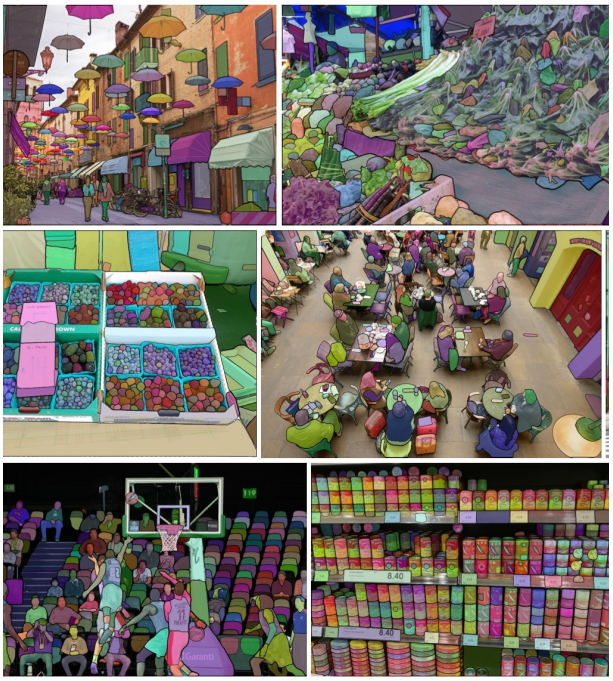

Figure 1. SAM masks on SA-1B dataset

3. Method

Our work takes inspiration from SAM3D [4], with the objective of utilizing SAM's capabilities for the 3D domain, and using only point cloud data to produce our images, in which the masks from SAM will be generated.

3.1. ScanNet++ Dataset

The ScanNet++ dataset [5] serves as the primary source of our point cloud data for this project. ScanNet++ is a large-scale indoor scene dataset, consisting of densely sampled point clouds, semantic segmentations, and instance annotations for various indoor scenes. The dataset is a crucial resource for researchers and practitioners in the field of computer vision, particularly for tasks related to indoor scene understanding and reconstruction. By utilizing this rich and diverse dataset, we aim to develop a robust and effective solution for segmenting point clouds and reconstructing segmented regions in 3D space. In the scope of our project, we used 10 selected scenes from ScanNet++. Due to the lack of sufficient computing resources, we sampled half of the points for each of the 10 point clouds.

3.2. Spherical Projection for 2D Image Generation

Our method begins by projecting the 3D point cloud onto a spherical surface. This process involves mapping each point in the cloud onto the surface of a sphere centered at the scene origin. The spherical projection serves as a transformative step, converting the intricate 3D information into a 2D representation.

The outcome of the spherical projection is an equirectangular image, as seen in Figure 2. This type of projection ensures that the resulting 2D image maintains equal spacing of longitude lines, providing a standardized representation of the spherical projection. Equirectangular output serves as a convenient and well-defined format for subsequent processing steps.

In summary, our spherical projection technique involving the mapping of 3D point clouds onto a sphere and the generation of equirectangular images provides a novel and effective method for creating our own 2D images with geometric fidelity.

Figure 2. Spherical projection [3]

An example of one of our generated 2D images from 3D point clouds from the ScanNet++ Dataset can be seen on Figure 3.

3.3. SAM Masks Generation

Initially, we employ the automatic mask generation approach implemented by the Segmentation Anything Model (SAM) [2] on individual 2D image frames to acquire pixel-level masks for each image. Figure 4 shows the result of the mask generation from SAM, applied to 3.

SAM produces hierarchical masks with varying granularity, encompassing entire objects, their constituent parts, and subparts. To ensure the creation of non-overlapping masks, when a pixel falls under the coverage of multiple masks, the mask ID associated with the highest predicted Intersection over Union (IoU) is assigned to that pixel.

In order to mitigate over-segmentation, fune-tuning of SAM was attempted. Unfortunately, due to limited resources, enough data could not be labeled to attain promising results. Therefore, overfitting significantly diminished performance, resulting in outcomes inferior to those of the original, unmodified model. Consequently, we opted to utilize the foundational model without engaging in fine-tuning.

SAM can be loaded with three different encoders: "ViT_B", "ViT_L" and "ViT_H" (where "B" stands for basic, "L" for large and "H" for huge), which have different parameter counts, and "H" being the one with the highest number of parameters. We opted for the "ViT_H" after noticing better results with it. Despite the distortion resulting from spherical projection, the instance segmentation achieves strong results, as can be seen in Figure 4.

3.4. Projection of Segmented Images to 3D

The next step is lifting the segmented 2D images back to 3D point clouds. In transitioning from the segmented 2D image to three-dimensional space, we associate each pixel with a corresponding 3D point. While this correlation is fundamental, it's important to note that not all pixels seamlessly translate into tangible 3D points. In such cases, denoted as black dots in Figure 3, we attribute a value of -1 within our mapping matrix. This meticulous approach ensures that our mapping matrix retains the dimensional congruence of the original 2D image. Finally, the points not labeled are addressed through k-NN.

After segmenting the point clouds using our proposed algorithm, we employed a k-nearest neighbors (k-NN) approach using 10 neighbors to assign colors to the segmented regions with unlabeled points. Specifically, we utilized the mode color assignment strategy, where the most frequent color among the nearest neighbors was assigned to each point within a segment. This method ensures that the assigned color accurately represents the predominant color within the neighborhood, thus enhancing the visual coherence of the segmented regions. By prioritizing the most frequently occurring color, we mitigate the risk of introducing unrealistic color blends that may not accurately reflect the underlying characteristics of the scene, e.g., by using the mean color of the neighbors instead.

4. Results

In this section, we present the results of our proposed method for segmenting point clouds in 3D space. We provide visualizations of the projected point cloud image, the segmented image, and the instance segmentation of the point clouds to illustrate the effectiveness of our approach.

Figure 3. 2D Spherical projection of 3D point cloud from ScanNet++ dataset

The projected point cloud image as seen in Figure 3 serves as the initial input to SAM. It provides a comprehensive view of the indoor scene and serves as the foundation for subsequent segmentation steps.

The segmented image depicted in Figure 4 showcases the results of SAM applied to the projected point cloud data.

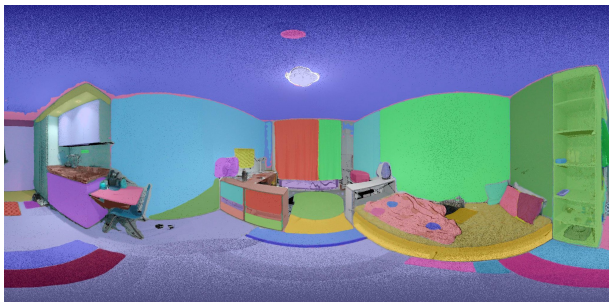

Figure 4. SAM masks applied on the spherical projection of Figure 3

Each different object is classified into one of several instances. The segmentation highlights distinct regions within the scene, enabling precise identification and classification of objects and structures.

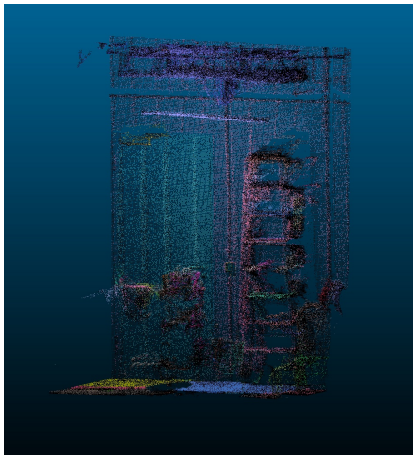

Figure 5. Segmented point cloud

Finally, we present the segmented point clouds observed in Figure 5, where each segmented region is lifted back into 3D space. They faithfully preserve the geometric details and instances of the original segmented regions, facilitating enhanced scene understanding and analysis, hereby enabling intuitive visualization and manipulation of the segmented regions, offering valuable insights into the structure and composition of the indoor scene.

The efficacy of our color assignment method using the k-NN mode color approach is demonstrated through visual inspection of the segmented point clouds. As shown in Figure 6, the segmented regions exhibit smooth transitions between points with coherent color representations. This is achieved by assigning each point the most frequent color among its 10 nearest neighbors, thereby preserving the local color distribution within the segmented regions. By com-

paring Figure 5 with the segmented point cloud colored using our mode color assignment method (Figure 6), it is evident that this process effectively captures the distinct features of the scene while maintaining visual consistency and achieving decent results with limited computing resources, in contrast to the heavy bidirectional merging process proposed by Yang et al. [4].

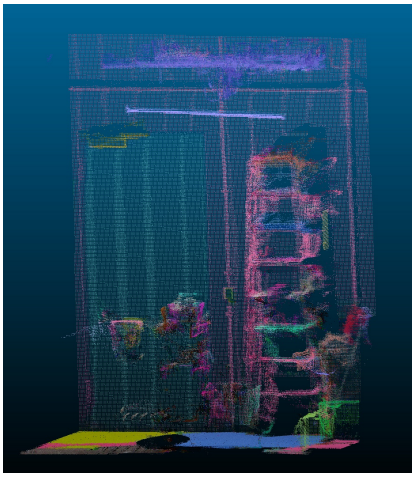

Figure 6. Segmented point cloud with mode color assignment

In Figures 7 and 8, we can observe that points within each segment share similar colors, indicating the successful application of the mode color assignment strategy. For instance, in the highlighted segment, the predominant color corresponds to the mode color of its neighboring points, resulting in a coherent and visually appealing representation of the segment.

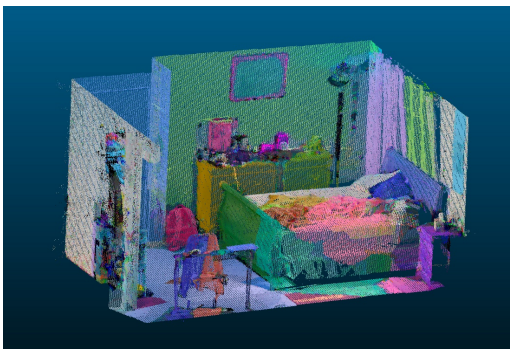

Figure 7. Segmented point cloud of a bedroom after k-NN

This visual assessment validates the effectiveness of our proposed method in accurately assigning colors to segmented point clouds, thereby enhancing the overall perceptual quality of the rendered scenes.

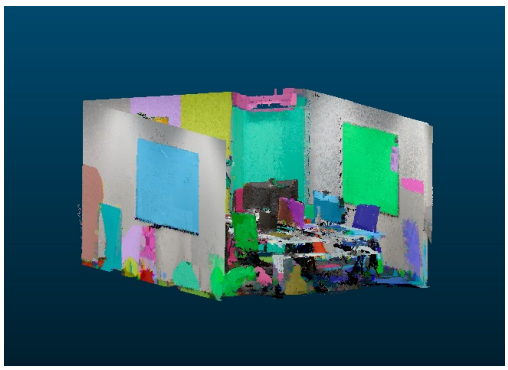

Figure 8. Segmented point cloud of an office space after k-NN

During the course of our work, the main challenges we faced were:

- Handling the heavy point cloud data, with limited computing resources available. Therefore, random sampling half of the points for each of the 10 point clouds was the solution adopted to overcome this obstacle.
- Achieving worse results with SAM fine-tuning due to limited training data.
- Bidirectional merging [4] for oversegmentation mitigation, since the iterative merging of local frames was deemed not feasible due to our available resources. Therefore, the k-NN approach was adopted.

5. Conclusion

In this work, we presented an approach for the instance segmentation of 3D point clouds, leveraging spherical projection and the Segmentat Anything Model (SAM). Our methodology aimed to address the challenges of processing point cloud data in a spherical domain while preserving geometric details. Through extensive experimentation and evaluation, we have demonstrated the effectiveness and applicability of our proposed method. In conclusion, our work contributes to the advancement of point cloud processing techniques, particularly in the context of indoor scene understanding. By combining spherical projection, instance segmentation, k-NN, and 3D point clouds, we have developed a robust framework for analyzing and manipulating point cloud data in a spherical domain. We believe that our approach holds promise for applications in robotics and computer vision, paving the way for further research in these exciting areas. Moving forward, it is recommended to incorporate Mean Average Precision (mAP) as a quantitative

measure for evaluating the efficacy of the proposed instance segmentation approach. This entails first computing the similarities between instances to establish correspondences, for example, by utilizing the Hungarian algorithm. Such an inclusion would provide a better understanding of performance and facilitate a comprehensive quantitative analysis of the results. Furthermore, it would be valuable to expand the number of labeled images, in attempting to enhance the efficacy of fine-tuning efforts, potentially leading to improved segmentation results. Additionally, a focused exploration into object detection by refining SAM to output single masks tailored to specific object classes holds considerable potential, particularly in applications such as autonomous driving where precise identification, such as pedestrian detection, is crucial. Leveraging SAM's flexibility to adjust the model output to a single mask offers an intriguing avenue for further investigation in this regard.

References

[1] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15587–15597, 2021. 1

[2] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 1, 2

[3] Bashar Alsadik Research Gate. Equirectangular panorama in a spherical projection, 2018. 2

[4] Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. Sam3d: Segment anything in 3d scenes, 2023. 1, 2, 4

[5] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 1, 2