Any3DIS: Class-Agnostic 3D Instance Segmentation by 2D Mask Tracking

Supplementary Material

In this supplementary, we first show one more application of class-agnostic 3DIS in interactive segmentation in Sec. 1. Next, we provide more analysis in Sec. 2. Finally, we present additional qualitative results that showcase Any3DIS's robustness and superior segmentation quality across diverse scenarios in Sec. 3.

1. Application: Interactive 3DIS

Thanks to the class-agnostic nature of our approach, it is applicable to a wide range of 3DIS tasks, including class-agnostic 3DIS, Open-vocabulary 3DIS, and Open-ended 3DIS, as demonstrated in Sec. 4.2 of the main paper. Here, we showcase an additional application: interactive 3DIS. In this scenario, given a 3D point cloud and H user-provided clicks or selected points specifying the object of interest, the goal is to segment the 3D mask of the object associated with these points. This application is particularly valuable for 3D instance segmentation annotation, enabling users to simply click on the object of interest to segment it.

To address this problem, we first apply our approach (and the baseline, Open3DIS [2]) to the given 3D point cloud to generate 3D mask proposals. Given H selected points, we identify H 3D proposals that contain these points and merge them to form the final 3D mask prediction. For complex objects, such as L-shaped sofas, more selected points are required compared to simpler objects like a remote control.

To evaluate performance, we introduce the IoU@H metric, which compares the predicted 3D mask with the ground truth (GT) mask containing these H selected points on the ScanNet++ dataset [4]. The results, presented in Tab. 1, demonstrate the effectiveness of our approach. Open3DIS utilizes SAM-HQ [1] as its 2D segmenter, whereas our approach employs SAM2-L [3]. Our method, Any3DIS, significantly outperforms Open3DIS across all metrics, achieving an improvement of approximately 20 IoU.

Method	IoU@1	IoU@3	IoU@5	IoU@8
Open3DIS	32.12	38.53	41.25	42.14
Any3DIS	47.61	59.36	61.40	62.44

Table 1. **Results of Interactive 3DIS on ScanNet++**, benchmarking on the masks of 1,554 classes.

2. Analysis

Time complexity of Algorithm 1. The time complexity of Algorithm 1 of the main paper is $O(T \times L)$, where T is the number of views and L is the number of superpoints lifted from the given 2D mask track. The time required to compute $\mathcal{L}$ is linear with respect to the number L of superpoints involved. This is substantially more efficient than the optimal solution, which involves evaluating all possible combinations of L superpoints and has a time complexity of $O(2^L \times L)$. Here, $2^L \gg T$.

Methods	S & T†	S‡	Lift	Alg.1 (Refine)	Merge	Total
Any3DIS	267	-	12	209	-	488
Open3DIS	-	3,841	18	-	367	4,226

Table 2. Runtime analysis of Any3DIS and Open3DIS for each component on ScanNet++ (in seconds). † indicates segmentation for the pivot frame and tracking for the remaining frames, while ‡ represents segmentation for all frames. A '-' denotes components that are not involved. Refer to Fig. 2 (main paper) for the details.

# neighboring superpoints κ	AP	AP$_{50}$	AP$_{25}$
0†	19.8	31.2	43.8
64	22.0	35.6	**47.3**
128	**22.2**	**35.8**	47.0
256	21.8	35.0	46.7

Table 3. Component analysis of the number of neighboring superpoints used to select pivot view for tracking. † denotes we do not consider any neighbors, or $s_t^l = 1$ in Eq. (2) of the main paper.

Inference time. We provide a detailed runtime analysis for each step of Open3DIS and Any3DIS (ours) on ScanNet++ in Tab. 2. The two methods are fundamentally different as illustrated in Fig. 2 of the main paper. The key difference lies in Open3DIS requiring high-quality segmentation for all frames, whereas Any3DIS only requires segmentation for the pivot frame, with tracking applied to the remaining frames. On ScanNet200, the runtime can be lower due to smaller scenes.

Study on the number of neighboring superpoints κ. Tab. 3 presents the results of varying the number of neighboring superpoints κ in Eq. (2) from 0 to 256. The best performance is achieved with $\kappa = 128$, yielding an AP of 22.2, compared to 19.8 when no neighbors are used. This highlights the importance of incorporating neighboring superpoints when selecting the pivot frame.

3. Qualitative Results

In this section, we present additional comparative results between our approach, Any3DIS, and Open3DIS, as shown in Figs. 1 and 2. These results cover diverse scenarios, including cluttered and sparse scenes, various object

categories, and complex spatial arrangements. The comparisons demonstrate that Any3DIS consistently outperforms Open3DIS by minimizing over-segmentation, preserving object boundaries, and achieving closer alignment with ground truth segmentations. These qualitative examples further highlight the robustness and generalizability of Any3DIS across challenging 3D scene segmentation tasks, showcasing its effectiveness in diverse indoor environments.

References

[1] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In *NeurIPS*, 2023. 1

[2] Phuc D. A. Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 3, 4

[3] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1

[4] 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

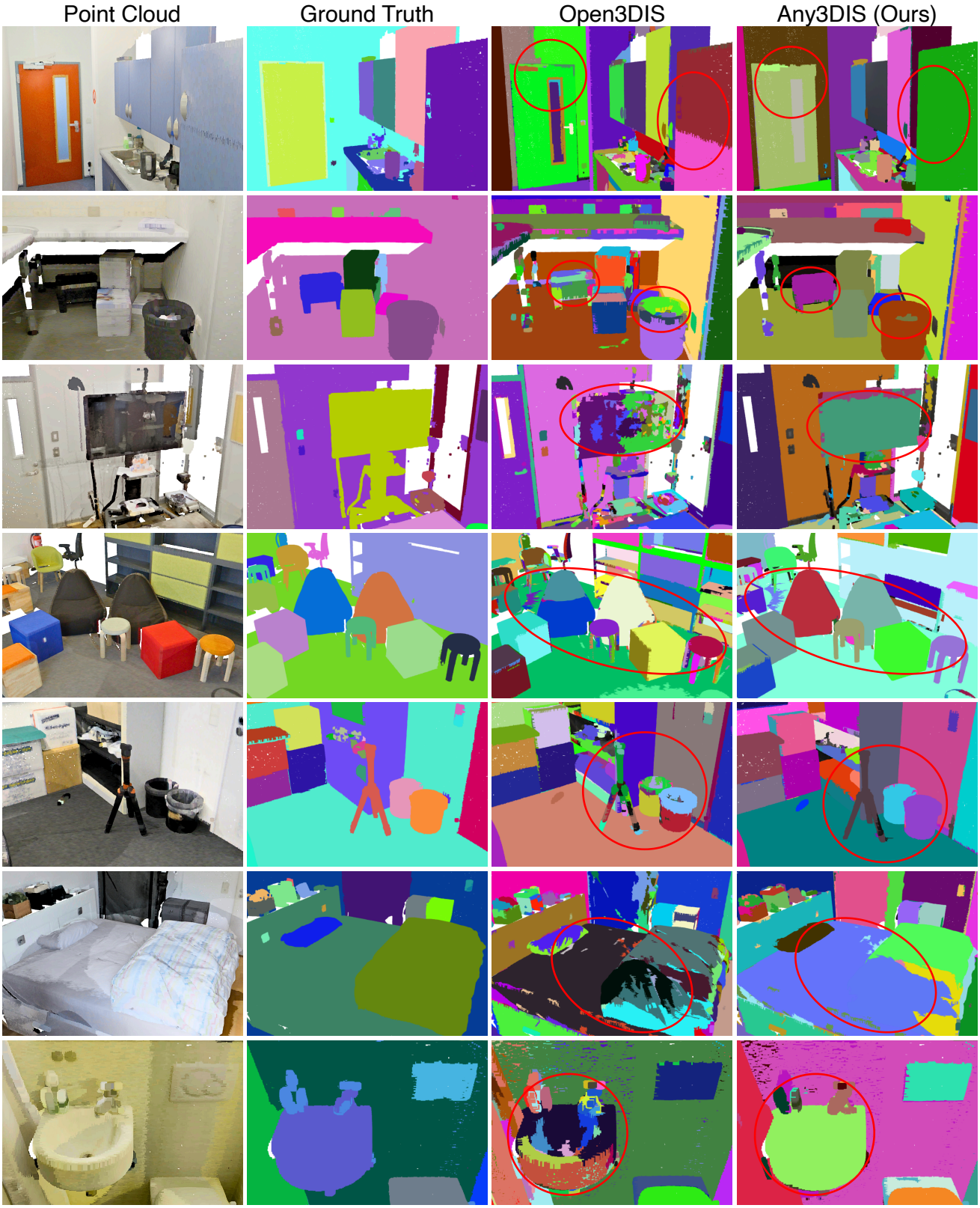

Figure 1. **Qualitative results on ScanNet++ Validation Set**: From left to right we show the input point cloud of 3D scenes, GT segmentation, Open3DIS [2], and Any3DIS (ours) results. Our approach achieves more accurate and consistent segmentation than Open3DIS.

Figure 2. **Qualitative results on ScanNet++ Validation Set**: From left to right we show the input point cloud of 3D scenes, GT segmentation, Open3DIS [2], and Any3DIS (ours) results. Our approach achieves more accurate and consistent segmentation than Open3DIS.