PS3: Part level instance segmentation in 3D

Supplementary Material

1. More Qualitative Results on Multiscan

Fig. 1 presents qualitative comparisons on the Multiscan dataset. We compare Search3D, our method, the ground truth (GT) segmentations, and the corresponding color mesh. Each color denotes a class-agnostic part-level 3D proposal. As shown in the results, Search3D often fails to separate parts with similar geometric attributes, particularly when they are located on the same plane. This limitation arises from its reliance on geometric over-segmentation, which tends to merge geometrically similar but semantically distinct parts into a single region. In contrast, our method produces more accurate and consistent part-level proposals. By leveraging 2D mask information rather than purely geometric cues, our approach better distinguishes adjacent parts with similar geometry, resulting in finer and more semantically meaningful segmentation.

2. Discussion and Limitation

Even though our method solves the limitation of Search3D[4], there are still some limitations in our method that need to be solved. The first one is how to decide the granularity of 2D masks. In our method, we pick masks of a specific granularity as the initial mask for tracking, and we use the tracking result as the standard for merging to

Figure 1. **Qualitative results on Multiscan.** From top to bottom, we show the results of Search3D[4] and our method, GT segmentations, and the color mesh. In (A) to (C), each color represents one class-agnostic part-level 3D proposal. Our approach achieves more accurate and consistent segmentation than Search3D[4].

generate a part proposal. However, different parts may not appear in the same granularity. This causes ambiguity in how to decide the most proper granularity. One possible solution is using 2D foundation models like [2][1] to help identifying the most proper granularity. The second limitation is the accuracy of the tracking model[3]. We found that SAM2[3] often fails to track correctly when multiple parts share similar textures. This limitation can be solved once there is a more powerful tracking model in the future, because each component in our method is plug-and-play.

References

[1] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 2

[2] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 2

[3] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2

[4] Ayca Takmaz, Alexandros Delitzas, Robert W. Sumner, Francis Engelmann, Johanna Wald, and Federico Tombari. Search3D: Hierarchical Open-Vocabulary 3D Segmentation. *IEEE Robotics and Automation Letters (RA-L)*, 2025. 1