SAI3D: Segment Any Instance in 3D Scenes

Yingda Yin* [1,2] Yuzheng Liu* [2,3] Yang Xiao* [4]
Daniel Cohen-Or [5] Jingwei Huang [6] Baoquan Chen [2,3]

[1]School of Computer Science, Peking University [2]National Key Lab of General AI, China
[3]School of Intelligence Science and Technology, Peking University
[4]Ecole des Ponts ParisTech [5]Tel-Aviv University [6]Tencent

1. Implementation Details

Multi-level merging criteria. To compute the affinity score A_{R,Q_i} between a region R and a superpoint Q_i, we consider all the affinity scores $A_{i,k}$ between Q_i and the superpoints inside the region $\{Q_k\} \in R$. Specifically, it is computed as the weighted average

$$A_{R,Q_i} = \frac{1}{\sum_{Q_k \in R} \beta_{i,k}} \sum_{Q_k \in R} \beta_{i,k} \, A_{i,k} \tag{1}$$

where $A_{i,k}$ is the affinity between Q_i and Q_k, and $\beta_{i,k}$ is the weight factor indicating how much $A_{i,k}$ contribute to A_{R,Q_i}. $\beta_{i,k}$ is determined by the distance $d_{i,k}$ between Q_i and Q_k, as well as the number of points N_k inside Q_k.

$$\beta_{i,k} = \begin{cases} \gamma^d N_k & d \leq 2 \\ 0 & d > 2 \end{cases} \tag{2}$$

For superpoints that can directly reach Q_i, we define the distance as $d = 1$. Similarly, the set of superpoints that reach Q_i through one bridging superpoint is with $d = 2$, and so forth. We eliminate the contributions when $d > 2$ for computational efficiency. γ is set as 0.5.

Progressive region growing. The region-growing algorithm is illustrated in Algo. 1. For progressive growing, we build a multi-stage region-growing framework with the affinity threshold varying from high to low. Our algorithm benefits from the dynamic strategy and leads to robustness to the choice of the thresholds. We set [0.9, 0.8, 0.7] for ScanNet++ dataset and [0.9, 0.8, 0.7, 0.6, 0.5] for ScanNet dataset. We did not tune the hyper-parameters much.

Datasets. To balance the performance and the efficiency, we sample proportions of images for different datasets. We use 5% views for ScanNet++ dataset, 20% views for both ScanNetV2 and ScanNet200 dataset, and all images in Matterport3D since it is a sparsely-scanned dataset.

Following the common practice [3, 6, 7], we ignore the instances with semantics of “wall” and “floor” for ScanNetV2 and ScanNet200 datasets. For Matterport3D dataset, we consider the most frequent 160 classes provided in [4] and ignore “wall”, “floor” and “ceiling”. For ScanNet++ dataset, we use all GT labels provided for the instance segmentation task (in “instance_classes.txt”). IPhone images are served as our input to better reflect everyday cases.

Evaluation. Following the baselines [4–7], we evaluate the numerical results on the validation set for ScanNetV2, ScanNet200 and ScanNet++ datasets and on the test set for Matterport3D dataset. We follow UnScene3D [5] to implement *class-agnostic* instance segmentation, where all object categories are treated equally and only the mask AP values are considered. We set all the confidence scores as 1.0, the same as [7]. For 2D foundation models, we choose SAM-HQ [1] for ScanNet++ dataset, and Semantic-SAM [2] for ScanNetV2, ScanNet200 and Matterport3D datasets for better granularity control.

2. Additional Qualitative Results

We show more visual comparisons on ScanNet++ dataset in Fig. 1.

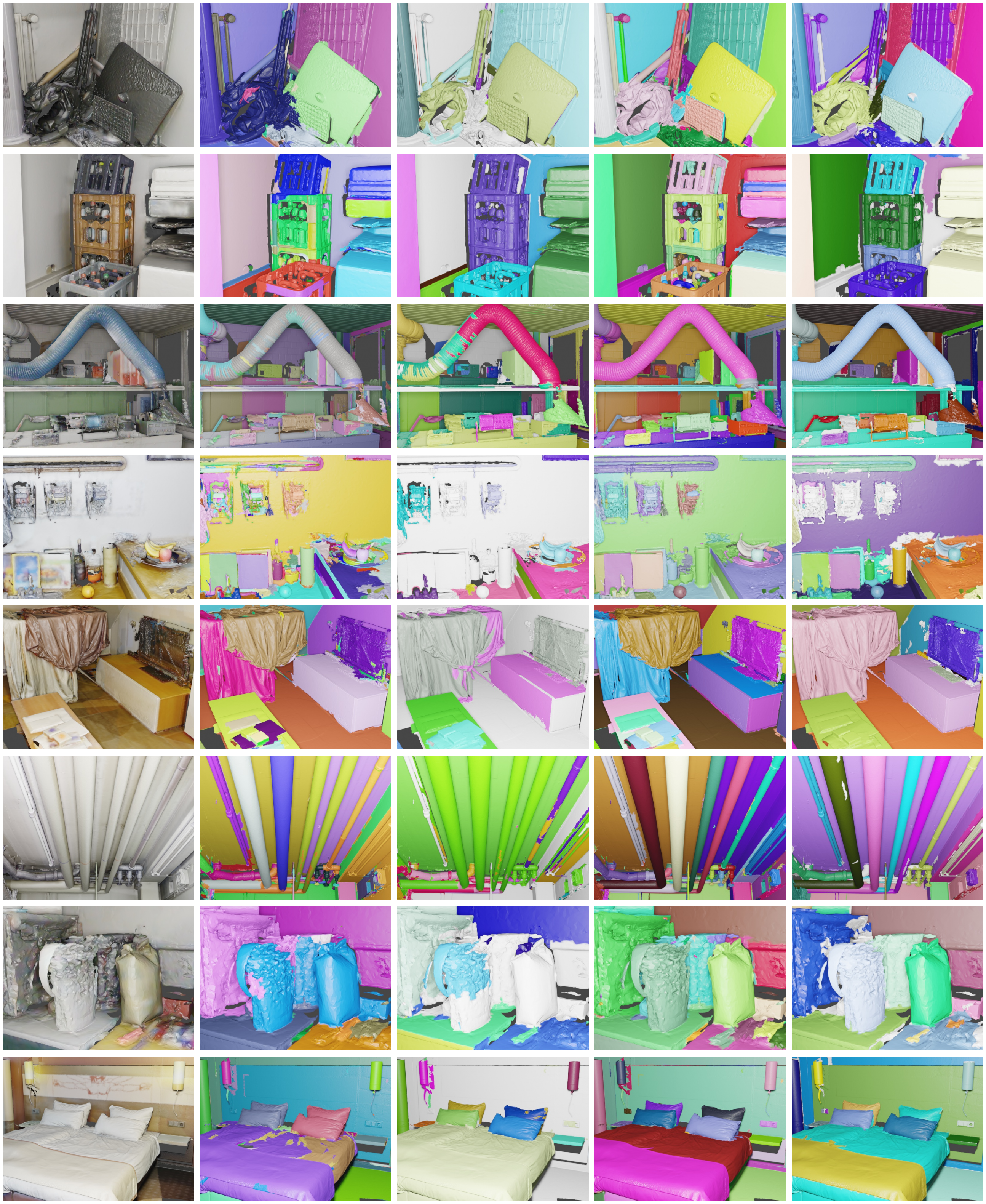
Visual results on ScanNet dataset are illustrated in Fig. 2. We find that sometimes our method even results in finer and more accurate segmentation masks than the ground truth annotations. See the clutter in the first four rows.

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] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao.

*Equal contribution.



Input SAM3D Mask3D Ours GT

Figure 1. **Additional visual results of 3D instance segmentation on ScanNet++ dataset.**

Figure 2. **Visual results of 3D instance segmentation on ScanNetV2 dataset.**

Algorithm 1 Region-growing algorithm

Input: Affinity matrix $\mathbf{A} \in \mathbb{R}^{N_Q \times N_Q}$ where $A_{i,j}$ indicates the affinity score between two super points Q_i and Q_j, and N_Q is the number of superpoints.
Output: Instance label $\mathbf{l} \in \mathbb{R}^{N_Q}$.

```
l ← 0
id ← 1
for i ← 1 to N_Q do
    if l_i = 0 then
        Queue B
        B.push(i)
        l_i ← id
        while B not empty do
            v ← B.pop()
            for j ← neighbors of v do
                if l_j ≠ 0 then
                    continue
                R ← {Q_k | l_k = id}
                A_{R,Q_j} ← Multi-level_criteria(R, Q_j, A)
                if A_{R,Q_j} > τ then
                    B.push(j)
                    l_j ← id
        id ← id + 1
```


Semantic-sam: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767*, 2023. 1

[3] Shiyang Lu, Haonan Chang, Eric Pu Jing, Abdeslam Boularias, and Kostas Bekris. Ovir-3d: Open-vocabulary 3d instance retrieval without training on 3d data. In *Conference on Robot Learning*, pages 1610–1620. PMLR, 2023. 1

[4] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[5] David Rozenberszki, Or Litany, and Angela Dai. Unscene3d: Unsupervised 3d instance segmentation for indoor scenes. *arXiv preprint arXiv:2303.14541*, 2023. 1

[6] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8216–8223. IEEE, 2023. 1

[7] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1