Segment3D: Learning Fine-Grained Class-Agnostic 3D Segmentation without Manual Labels

Rui Huang[1], Songyou Peng[2], Ayça Takmaz[2], Federico Tombari[3], Marc Pollefeys[2,4], Shiji Song[1], Gao Huang[1,*], and Francis Engelmann[2,3]

[1]Tsinghua University [2]ETH Zurich
[3]Google [4]Microsoft [*]Corr. Author
https://segment3d.github.io

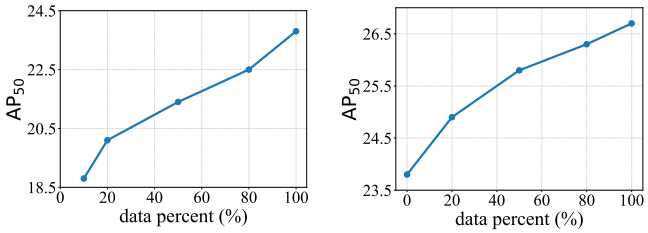
Appendix

A Implementation Details

Setup. ScanNet [4,7] comprises 1513 indoor scenes, encompassing $\sim 2 \cdot 10^6$ views, along with 3D camera poses and surface reconstruction. We train our model with the training set of ScanNet. For Stage 1, we sample every 25^{th} frame of the RGB-D sequences ($\sim$ 1 FPS) and obtain approximately $76 \cdot 10^3$ training frames. For Stage 2, we use the $\sim 1.2 \cdot 10^3$ reconstructed 3D scans of indoor spaces as full point clouds. Note that in both stages, we do not use any annotations of ScanNet. For the experiment with pre-training using more data, we select additional frames from the training set of ScanNet++ [12] (to increase the variety of training data). We sample frames at roughly 1 FPS resulting in 34k frames, in addition to the previous 76k frames of ScanNet.

For the class-agnostic segmentation, we evaluate on popular datasets including ScanNet++ [12], ScanNet200 [4, 7] and nuScenes [1], STPLS3D [2], Paris-Lille-3D [6] (PL3D) urban outdoor datasets. To apply the models trained on room-sized indoor scenes to the outdoor data, we split the large outdoor scenes into multiple smaller crops due to GPU memory limitations, and the point cloud coordinates are scaled down to be better aligned with the room-sized training data. Since the test set labels of PL3D are not publicly available, we use its training set to perform the zero-shot evaluation. For the open-set scene understanding, we adapt our model to OpenMask3D [10], and evaluate on the validation set of ScanNet++ [12]. We further test on the *office0, office1, office2, office3, office4, room0, room1, room2* scenes of Replica [9] following OpenMask3D [10].

Model Training. The backbone of Segment3D is a Minkowski Res16UNet34C [3]. We perform standard data augmentations, including horizontal flipping, random rotations, elastic distortion and random scaling. In addition, we use color augmentations including jittering, brightness and contrast augmentation. For Stage 1, we use AdamW optimizer and a one-cycle learning rate schedule with a peak learning rate of 2×10^{-4}. The model is trained for 20 epochs with a batch size of 16 partial RGB-D point clouds. Training on 2 cm voxelization takes



(a) Performance with increasing ScanNet data

(b) Performance with increasing additional ScanNet++ data

Fig. B: Performance with Increasing Training Data. Scaling up the training data can lead to consistent performance improvements.

approximately 60 hours with 2 RTX3090 GPUs. For Stage 2, the initial learning rate is set to 2×10^{-4}. We train the model for 50 epochs with a batch size of 8 full 3D point clouds. Training takes $\sim$ 10 hours with 2 cm voxels on 4 A100 GPUs. We set the number of queries as 100 for Stage 1 and 150 for Stage 2 during training. For Stage 1, following Mask3D [8], the values of λ_{obj}, λ_{dice}, and λ_{ce} are set to 2, 2, and 5, respectively. For Stage 2, the values of λ_{dice} and λ_{ce} are set to 2 and 5, respectively.

Masks Generation. In Stage 1, after adopting the automatic mask generation pipeline of SAM, we utilize the proposed Mask Generation Module (MGM) to select high-quality, complete masks. The masks generated by MGM are then projected into 3D to serve as the supervision signal. In Stage 2, we first apply DBSCAN [5] to split erroneously merged instances within the predicted masks and then select the masks with a confidence threshold τ_c. Fig. A shows segmentation performance for varying values of τ_c. In our experiments, we set $\tau_c = 0.6$.

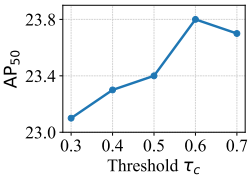


Fig. A: Performance for varying values of τ_c.

B Performance with Increasing Training Data

We plot the change in performance as the training data from ScanNet or ScanNet++ increases in Fig. B. In both scenarios, a performance increase is observed, confirming that additional automatic labels can further enhance results.

C Additional Results on Outdoor Data

We report the zero-shot segmentation results on STPLS3D [2] and Paris-Lille-3D [6] datasets in Table A and Table B respectively. Segment3D exhibits greater generalization capabilities compared to the fully-supervised Mask3D when transferring from indoor to outdoor scenes.

Table A: Segmentation Scores on STPLS3D [2] dataset.

Model	AP_{50}	AP_{25}
Mask3D	29.4	40.0
Segment3D (**Ours**)	**37.7**	**47.6**

Table B: Segmentation Scores on Paris-Lille-3D [6] dataset.

Model	AP_{50}	AP_{25}
Mask3D	1.5	3.3
Segment3D (**Ours**)	**4.2**	**6.3**

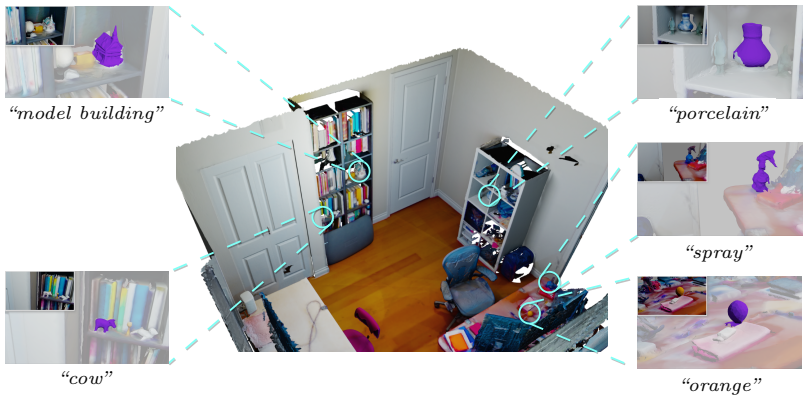


Fig. C: An example of open-set 3D object retrieval in a scene. Given a text prompt, OpenMask3D [10] based on Segment3D finds the corresponding object masks ■ in a given 3D scene. Since Segment3D can accurately generate fine-grained masks in a class-agnostic manner, we are able to retrieve small-scale objects of interest.

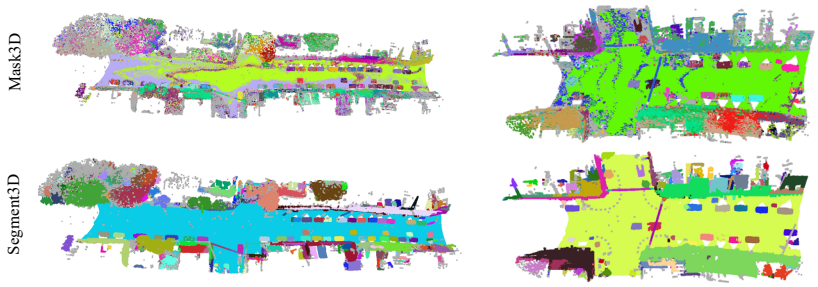


Fig. D: Qualitative Results on Paris-Lille-3D. We show segmentations predicted by Mask3D [8] (*top*) and our Segment3D (*bottom*). Segment3D performs impressively well even on the outdoor scenes it has never seen during training.

D Additional Qualitative Results

We show more qualitative results on ScanNet++ [12] dataset in Fig. E. Segment3D demonstrates superior results in segmenting fine-grained details and can even identify small object masks not annotated in the ground truth. Consequently, the full performance of Segment3D might not be accurately reflected in the scores. We also provide an example of open-set 3D object retrieval within a scene in Fig. C. Furthermore, we present the qualitative results on Paris-Lille-3D [6] dataset in Fig. D.

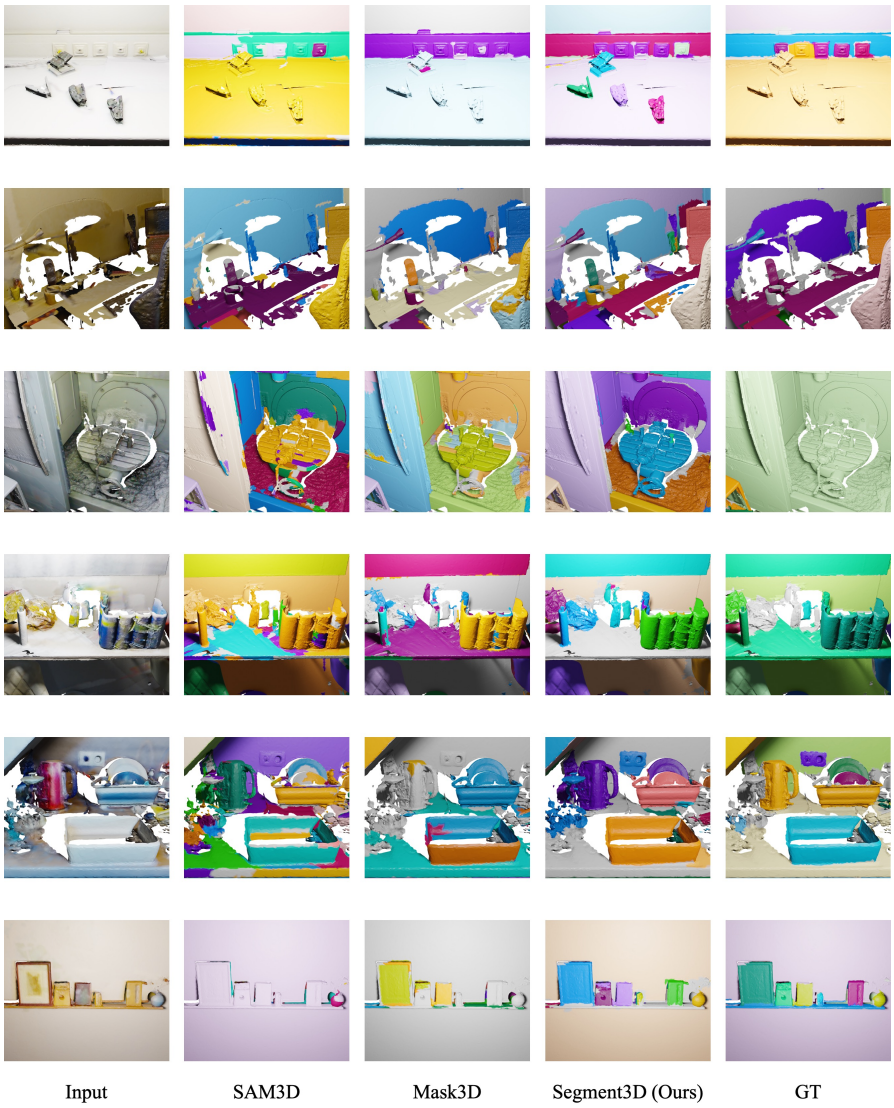


Fig. E: Additional qualitative results on ScanNet++ [12]. From left to right, we show the colored input 3D scenes, the segmentation masks predicted by SAM3D [11], Mask3D [8], our Segment3D and the ground truth 3D mask annotations. Segment3D shows superior results in segmenting fine-grained details and can even identify small object masks not annotated in the ground truth.

References

1. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuScenes: A Multimodal Dataset for Autonomous Driving. In: CVPR (2020) 1
2. Chen, M., Hu, Q., Yu, Z., Thomas, H., Feng, A., Hou, Y., McCullough, K., Ren, F., Soibelman, L.: STPLS3D: A Large-Scale Synthetic and Real Aerial Photogrammetry 3D Point Cloud Dataset. arXiv preprint arXiv:2203.09065 (2022) 1, 2, 3
3. Choy, C., Gwak, J., Savarese, S.: 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks. In: CVPR (2019) 1
4. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In: CVPR (2017) 1
5. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: KDD (1996) 2
6. Roynard, X., Deschaud, J.E., Goulette, F.: Paris-Lille-3D: A Large and High-Quality Ground-Truth Urban Point Cloud Dataset for Automatic Segmentation and Classification. IJRR (2018) 1, 2, 3
7. Rozenberszki, D., Litany, O., Dai, A.: Language-Grounded Indoor 3D Semantic Segmentation in the Wild. In: ECCV (2022) 1
8. Schult, J., Engelmann, F., Hermans, A., Litany, O., Tang, S., Leibe, B.: Mask3D: Mask Transformer for 3D Semantic Instance Segmentation. In: ICRA (2023) 2, 3, 4
9. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., et al.: The Replica Dataset: A Digital Replica of Indoor Spaces. arXiv preprint arXiv:1906.05797 (2019) 1
10. Takmaz, A., Fedele, E., Sumner, R.W., Pollefeys, M., Tombari, F., Engelmann, F.: OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In: NeurIPS (2023) 1, 3
11. Yang, Y., Wu, X., He, T., Zhao, H., Liu, X.: SAM3D: Segment Anything in 3D Scenes. In: ICCVW (2023) 4
12. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: ScanNet++: A High-Fidelity Dataset of 3D Indoor Scenes. In: ICCV (2023) 1, 3, 4