SAS: Segment Any 3D Scene with Integrated 2D Priors —Supplementary Material—

Anonymous ICCV submission

Paper ID 681

In this **supplementary material**, we first provide implementation details in Sec. A. Then, we supply additional qualitative results in Sec. B.

A. Implementation Details

Training setting we provide full information of the training configuration as shown in Tab. A. Specifically, we adopt Adam [1] as the optimizer with a base learning rate of $1e{-}4$. The learning scheduler adjusts the learning rate linearly to $1e-5$ throughout the whole process. The weight decay is set to 0. We use a batch size of 12 and 8 for indoor scenes and outdoor scenes respectively to train for 100 epoches in total. Besides, the voxel size is set to 2cm and 5cm respectively for indoor scenes and outdoor scenes.

ScanNet v2 [2] / Matterport3D [3]		nuScenes [4]	
Config	Value	Config	Value
optimizer	Adam [1]	optimizer	Adam [1]
scheduler	Linear	scheduler	Linear
base lr	1e-4	base lr	1e-4
weight decay	0	weight decay	0
batch size	12	batch size	8
epochs	100	epochs	100
voxel size	2cm	voxel size	5cm

Table A. **Training settings.** Here we list the training settings for both indoor scenes and outdoor scenes.

Model architecture We adopt MinkowskiNet18A [5] to be the architecture of the 3D distilled model, which is consistent with OpenScene [6]. Besides, the input to the 3D distilled model is the pure point cloud without color or other attributes.

nuScenes inference As some category names in nuScenes [4] have ambiguous meanings, e.g., "drivable surface" and "other flat", we follow OpenScene [6] to pre-define some detailed category names that have clear meanings, and then map the predictions from these pre-defined categories back to the original categories. The original categories and the pre-defined categories are shown in Tab. B.

nuScenes 16 labels	OpenScene's pre-defined labels
barrier	barrier, barricade
bicycle	bicycle
bus	bus
car	car
construction vehicle	bulldozer, excavator, concrete mixer, crane, dump truck
motorcycle	motorcycle
pedestrian	pedestrian, person
traffic cone	traffic cone
trailer	trailer, semi trailer, cargo container, shipping container, freight container
truck	truck
driveable surface	road
other flat	curb, traffic island, traffic median
sidewalk	sidewalk
terrain	grass, grassland, lawn, meadow, turf, sod
manmade	building, wall, pole, awning
vegetation	tree, trunk, tree trunk, bush, shrub, plant, flower, woods

Table B. **Label Mappings for nuScenes 16 Classes.** Here we list the total 43 pre-defined non-ambiguous class names corresponding to the 16 nuScenes classes.

Multi-view feature fusion Multi-view feature fusion is to aggregate the 2D image features onto 3D points through pixel-point correspondence. Our multi-view feature fusion strategy is exactly the same with OpenScene's [6]. Specifically, for Matterport3D [3] and nuScenes [4], we aggregate features of all images from every scene onto the 3D point, while we only sample 1 image out of every 20 video frames and fuse them for ScanNet [2]. Besides, we con-

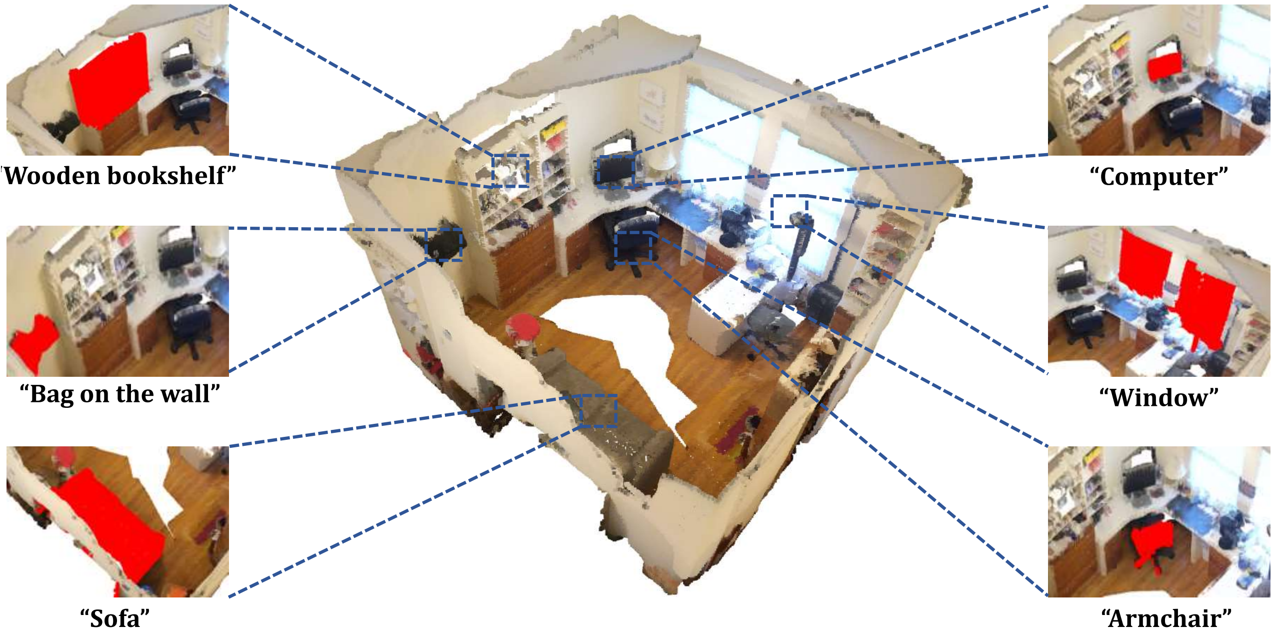

Figure A. **Querying about different objects in a scene.** The scene is collected from ScanNet v2. Red indicate the queried parts that match the text description.

duct occlusion tests for ScanNet v2 [2] and Matterport3D [3] as they provide depth information of each image, which guarantees that a pixel is only connected to a visible surface point. Specifically, for a single point, we calculate its distance between it and its corresponding pixel. If the difference between the distance and the pixel's depth value D is smaller than a threshold σ, we can connect the point to this pixel. Otherwise, we do not project the pixel's features onto the point cloud. We set $\sigma = 0.2D$ for ScanNet v2 [2] and $\sigma = 0.02D$ for Matterport3D [3], which is consistent with OpenScene [6].

Superpoint generation We compute superpoints only for indoor datasets ScanNet v2 [2] and Matterport3D [3]. Specifically, we use the mesh data provided by ScanNet v2 [2] and Matterport3D [3] as input. We extract superpoints from the mesh by performing a graph-based algorithm [7] on the computed mesh normals. For nuScenes [4], we do not compute any superpoint and treat every single point as a superpoint since ourdoor point clouds are normally dominated by "road", making it hard to extract superpoints.

Prompt engineering When extracting text features during inference, we apply a simple prompt engineering that modifies the class name "XX" to "a XX in a scene" to generate a better performance, which is proven by OpenScene [6]. Besides, when synthesizing images in Sec 3.2 in main paper, we apply another prompt engineering that modifies the class name "XX" to "a good photo of XX" to obtain high quality images.

Pre-built vocabulary We construct two pre-built vocabulary (Sec 3.2) for indoor scenes and outdoor scenes respectively, as shown in Tab. C.

Indoor scenes	Outdoor scenes
bookshelf, table, wall, bathtub, sofa, ceiling, door, bed, toilet, picture, desk, floor, counter, shower curtain, sink curtain, window, chair, cabinet, refrigerator,	person, bus, wall, grass, car, bicycle, crane, sky, tree, excavator, barricade, trailer, pavement, building, road, motorcycle, plant, truck, awning, container, lawn, traffi cone, bulldozer

Table C. **Pre-built vocabulary.** Here we give the detail of constructed pre-built vocabulary for indoor scenes and outdoor scenes respectiely.

B. Additional Qualitative Results

Querying objects in a scene We display a visualization of querying about different objects in a scene as shown in Fig. A. First, we adopt the 3D distilled model to output per-point features. Then we use different query texts and encode them with CLIP to obtain text features. By computing the similarity between point features and text features,

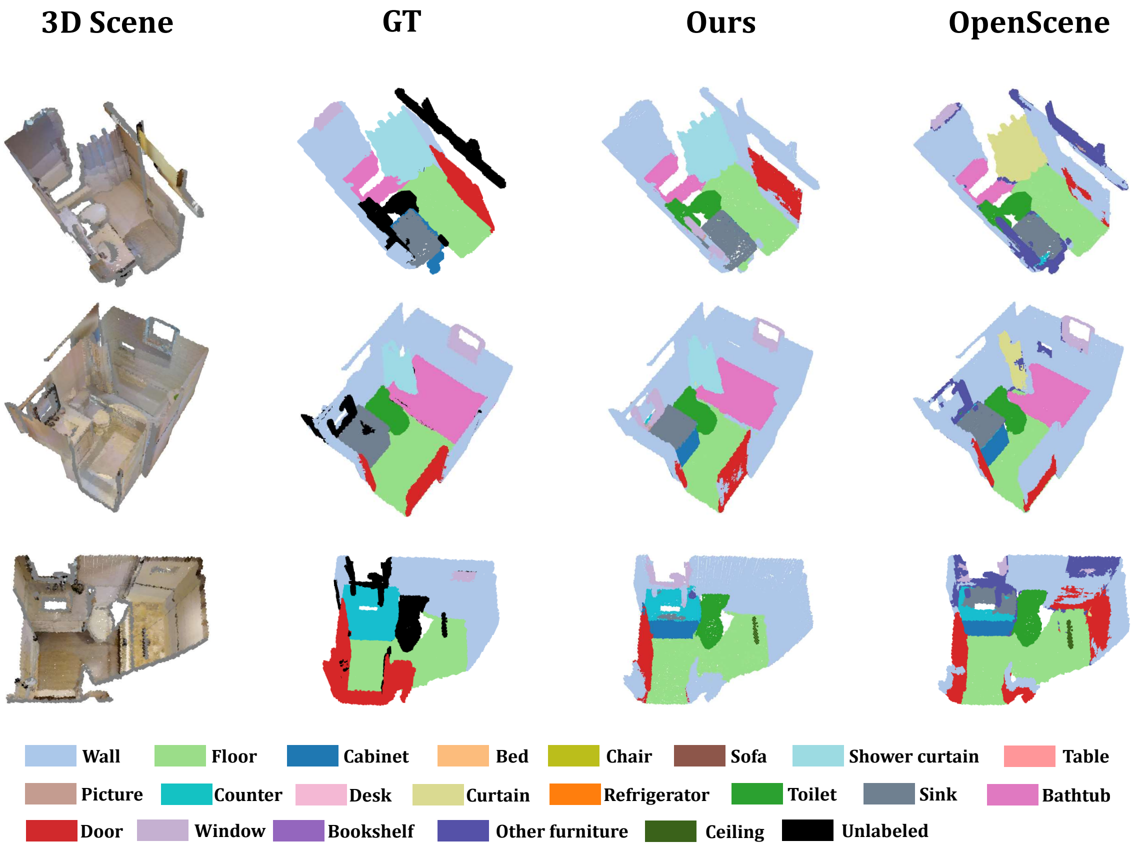

Figure B. **Visualization results.** Semantic segmentation results of SAS on Matterport3D [3].

we denote points with high similarity as red.

Visualization on Matterport3D Visual Comparisons with OpenScene [6] on semantic segmentation in Matterport3D [3] are shown in Fig. B, which our proposed SAS() effectively corrects some wrong predictions made by OpenScene [6]. For, example, OpenScene [6] misidentifies a shower curtain as a curtain, while SAS can easily fix it.

Visualization on nuScenes Visual Comparisons with OpenScene [6] on semantic segmentation in nuScenes [4] is also shown are Fig. C.

Visualization on gaussian segmentation results We also display the visualization of the gaussian segmentation on ScanNet v2 [2] in Fig. D.

References

[1] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[2] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5828–5839, 2017. 1, 2, 3, 4

[3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 1, 2, 3

[4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1, 2, 3, 4

[5] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3075–3084, 2019. 1

[6] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 815–824, 2023. 1, 2, 3

[7] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59:167–181, 2004. 2

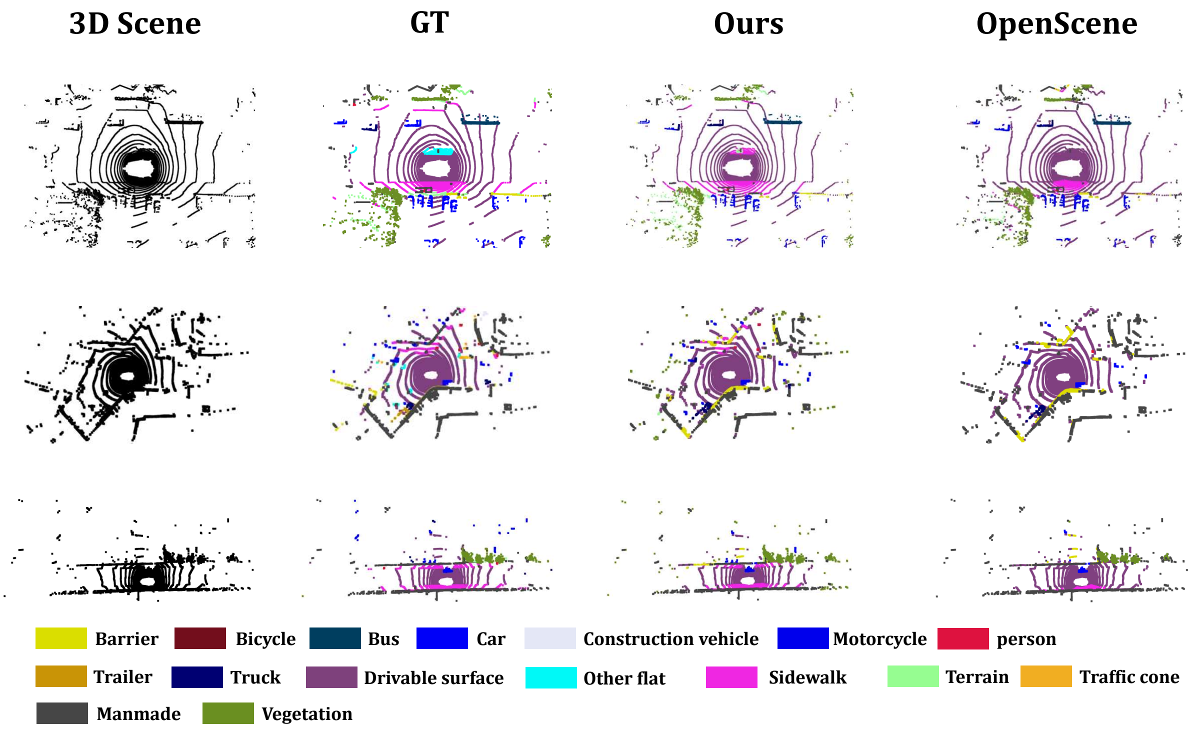

Figure C. **Visualization results.** Semantic segmentation results of SAS on nuScenes [4].

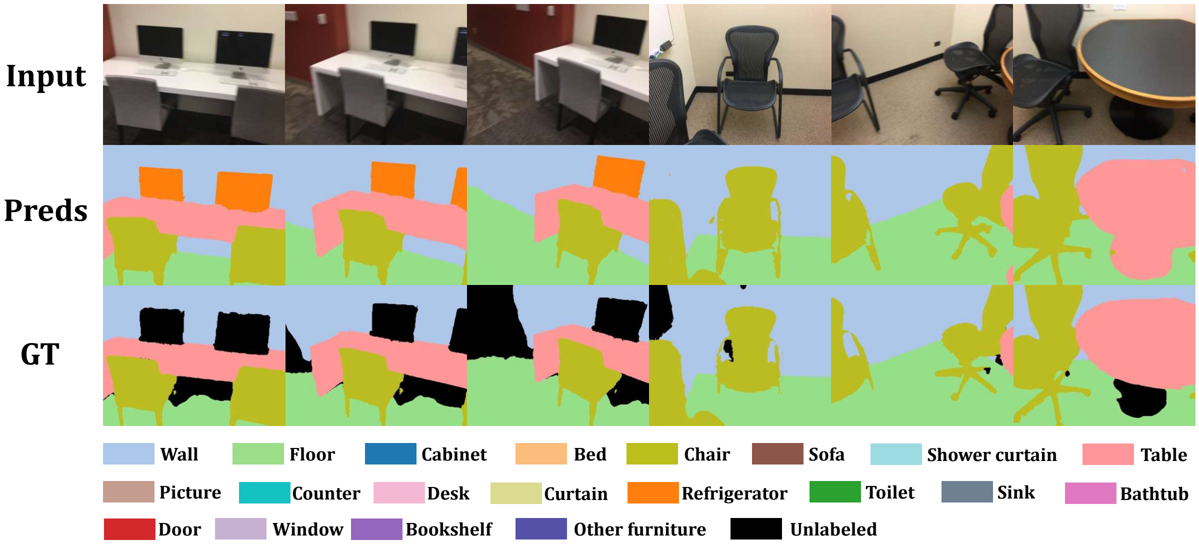

Figure D. **Visualization results.** Gaussian semantic segmentation results of SAS on ScanNet v2 [2].