Open Vocabulary 3D Scene Understanding via Geometry Guided Self-Distillation - Supplementary Material -

Pengfei Wang[1,2], Yuxi Wang[2], Shuai Li[1], Zhaoxiang Zhang[1,2,3,4(✉)], Zhen Lei[1,2,3,4], and Lei Zhang[1(✉)]

[1] The Hong Kong Polytechnic University
[2] Center for Artificial Intelligence and Robotics, HKISI, CAS
[3] State Key Laboratory of Multimodal Artificial Intelligence Systems, CASIA
[4] School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS)
pengfei.wang@connect.polyu.hk, zhaoxiang.zhang@ia.ac.cn, cslzhang@comp.polyu.edu.hk

The following contents are provided in this supplementary material:

- Section A: Detailed implementations of GGSD.
- Section B: More qualitative results.

A Implementation Details

More Details of Feature Fusion. OpenScene can be considered as our baseline model, and we strictly follow its approach for multi-view fusion of pixel embeddings. Specifically, on the nuScenes dataset, we perform fusion using all the images from each scene. However, on the ScanNet dataset, we sample one image out of every 20 video frames. For indoor datasets, we conduct occlusion tests. For each surface point, we first find its corresponding pixel in the image and calculate the distance between that pixel and the 3D point. Only when the difference between the distance and the depth value of that pixel is smaller than a threshold σ, do we pair the 3D point and the pixel. The threshold σ is proportional to the depth value D. Due to the high noise in depth values, we use $\sigma = 0.2D$ for ScanNet and $\sigma = 0.02D$ for Matterport. We do not project the features of pixels in the "invalid" regions of the depth map onto 3D points. For the nuScenes LiDAR point cloud, since no depth images are provided, we do not perform occlusion tests. We only use the synchronized images and the corresponding LiDAR points at the last timestamp of a 0.5-second segment.

More Details of Superpoint Generation. We employ VCCS [1] to generate superpoints for indoor datasets. VCCS starts with a set of seed points uniformly distributed on a voxel grid with a resolution of R_{seed}, and gradually expands the superpoints. Initially, we divide the input point cloud into voxel grids of size $2 \times 2 \times 2$ cm. Then, a set of seed points is uniformly distributed within the voxelized point cloud, with a spacing of 50 cm between seed points. For each seed point, within a sphere of radius 50 cm, we set the seed point as the initial center and search for its 27 neighboring points. The distance between

nuScenes 16 labels	Our 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 A1: Label Mappings for nuScenes 16 Classes. Here we list the total 43 pre-defined non-ambiguous class names corresponding to the 16 nuScenes classes.

each neighboring point and the center point is calculated using the following formula:

$$\boldsymbol{D} = \sqrt{wcD^2c + \frac{wsDs}{3R^2seed} + wnD_n}. \tag{A.1}$$

Here, Dc, Ds, Dn represent the Euclidean distances for color, spatial and normal attributes, respectively. For the hyperparameters, we directly follow the settings of GrowSP [2], setting the weights wc, ws, wn to 0.2, 0.4 and 1, respectively.

Considering that outdoor point clouds are typically dominated by "roads" and have significantly different point densities from indoor datasets, we adopt the random sample consensus (RANSAC) + Euclidean clustering as an alternative approach to VCCS for generating superpoints. Specifically, we utilize RANSAC to fit planes and considered points within a distance of 0.2m from the plane as a single large superpoint. After fitting the largest plane (usually corresponding to "roads"), we employ Euclidean clustering to create superpoints for the remaining points. Specifically, if the Euclidean distance between two points is less than 0.2m, they are assigned to the same superpoint; otherwise, they are not assigned.

Pre-defined Labels for nuScenes. For the indoor dataset, we directly utilize the predefined label names provided within the dataset. However, there are instances where certain class names may be ambiguous for the nuScenes benchmark. In such cases, we can take the approach employed by OpenScene and define unambiguous class names for each category in advance. To facilitate this mapping process, Table A1 presents the predefined class names and their corresponding nuScenes categories. By mapping the predictions back to these 16

categories, we can have a clearer understanding of our method's performance in the nuScenes benchmark.

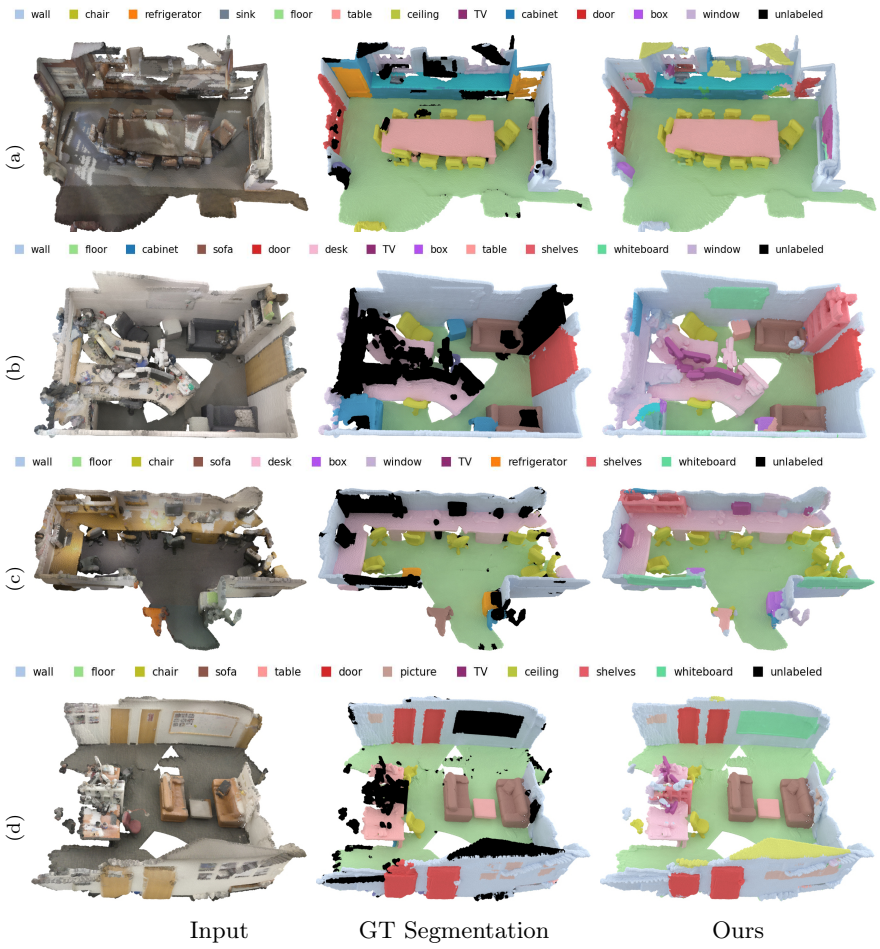

Fig. A1: More qualitative results. We present more qualitative results of 3D semantic segmentation on ScanNet benchmarks.

Dataset Partitioning for ScanNet with NYU-40 Label. To evaluate the open vocabulary capability of GGSD, we expand the original vocabulary size by using the NYU-40 label set. We remove the NYU-40 labels that do not have specific semantics (*e.g.*, "other structure", "other furniture", "other prop") and evenly divide all the rest categories into *Head*, *Common* and *Tail*. *Head* classes contain wall, floor, cabinet, bed, chair, bathtub, table, door, toilet, bookshelf, curtain, and ceiling. *Common* classes contain sofa, counter, desk, dresser, refrigerator, shelves, shower curtain, night stand, window, picture, sink and floor mat.

Tail classes contain blinds, mirror, clothes, pillow, book, box, whiteboard, lamp, towel, bag, person, and television.

B More Qualitative Results

We present additional qualitative results in Fig. A1. Our method successfully segments all categories that exist in the closed-set ScanNet benchmark, such as wall, desk, chair, and table. Additionally, our method successfully segments some categories that are not annotated in traditional benchmarks, as shown in Fig. A1 (a) with examples of ceiling and television, and in Fig. A1 (b), (c) and (d) with examples of whiteboard and shelves. These qualitative results obtained from GGSD demonstrate its excellent open vocabulary capability.

However, there are also some failure cases in the qualitative results. For example, in Fig. A1 (b), (c) and (d), computer screens are misclassified as televisions, and in Fig. A1 (c), the object picture is not segmented. These errors indicate that our model still needs further improvement in fine-grained segmentation and in segmenting objects with less prominent geometric structures, such as pictures. We will explore these areas in future work.

References

1. Papon, J., Abramov, A., Schoeler, M., Worgotter, F.: Voxel cloud connectivity segmentation-supervoxels for point clouds. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2027–2034 (2013) 1
2. Zhang, Z., Yang, B., Wang, B., Li, B.: Growsp: Unsupervised semantic segmentation of 3d point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17619–17629 (2023) 2