Supplementary Material:
An MIL-Derived Transformer for Weakly Supervised Point Cloud Segmentation

Cheng-Kun Yang[1] Ji-Jia Wu[2] Kai-Syun Chen[2] Yung-Yu Chuang[1] Yen-Yu Lin[2,3]
[1]National Taiwan University [2]National Yang Ming Chiao Tung University [3]Academia Sinica

This document provides additional visual examples, including both the successful and less successful ones. In addition, we show more class-specific quantitative results.

A. Class-wise quantitative performance

In Table 1 and Table 2 of the submitted paper, we report our method's average performance for object segmentation on the ScanNet [1] and S3DIS [2] datasets. Here we provide the detailed class-wise results for each type of weak supervision, including scene-level, subcloud-level, and 20 labeled points (around 0.01%) annotations. Table 1 and Table 2 show the class-wise performance of our method on the ScanNet and S3DIS datasets under different types of supervisions, respectively.

Supervision	wall	floor	cabinet	bed	chair	sofa	table	door	window	B.S.	picture	cnt	desk	curtain	fridge	S.C.	toilet	sink	bathtub	other	mIOU
Scene	52.1	50.6	8.3	46.3	27.9	39.7	20.9	15.8	26.8	40.2	8.1	21.1	22.0	45.9	4.5	16.6	15.2	32.4	21.2	8.0	26.2
Subcloud	66.0	80.6	28.0	58.2	68.8	63.2	47.6	23.8	11.8	60.9	6.0	9.2	46.4	72.0	13.6	56.8	68.8	34.0	66.7	34.4	45.8
20pt (0.01%)	74.7	92.3	53.2	74.5	82.7	77.9	61.2	43.8	39.0	67.9	2.5	44.3	50.3	53.1	39.8	64.2	79.6	42.5	77.3	35.3	57.8

Table 1. Quantitative results (mIoU) of the proposed method with diverse supervision settings on the ScanNet dataset. "B.S." denotes bookshelf; "S.C." stands for shower curtain and "cnt" denotes counter.

Supervision	ceil	floor	wall	beam	column	window	door	chair	table	bookcase	sofa	board	clutter	mIOU
Scene	24.9	4.7	40.0	0.0	1.3	2.2	1.8	5.6	16.8	33.0	32.1	0.1	5.8	12.9
0.02%	86.6	93.2	75.0	0.0	29.3	45.3	46.7	60.5	62.3	56.5	47.5	33.7	32.2	51.4

Table 2. Quantitative results (mIoU) of the proposed method with diverse supervision settings on the S3DIS dataset.

B. Qualitative results

Figure 1 displays the point cloud segmentation results generated by our method on the S3DIS and ScanNet datasets, including both successful and less successful examples. In general, the proposed method generates clear segmentation contours for objects as shown in Figure 1. For example, most chairs in both datasets are well-segmented with fine chair legs. However, objects with similar colors and depths cause segmentation difficulties, such as the whiteboard on the wall in the example of the last column of the fifth row in Figure 1. Also, small objects may be left out, such as those in the last row of Figure 1.

Figure 2 demonstrates additional comparison with an existing method [3]. Our method using only 0.02% labeled points can generates much smoother segmentation results compared with Xu *et al.*'s method [3] under 0.2% labeled annotations.

References

[1] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *ICCV*, 2017. 1

[2] Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung. Scenenn: A scene meshes dataset with annotations. In *3DV*, 2016. 1

[3] Xun Xu and Gim Hee Lee. Weakly supervised semantic point cloud segmentation: Towards 10x fewer labels. In *CVPR*, 2020. 1, 2

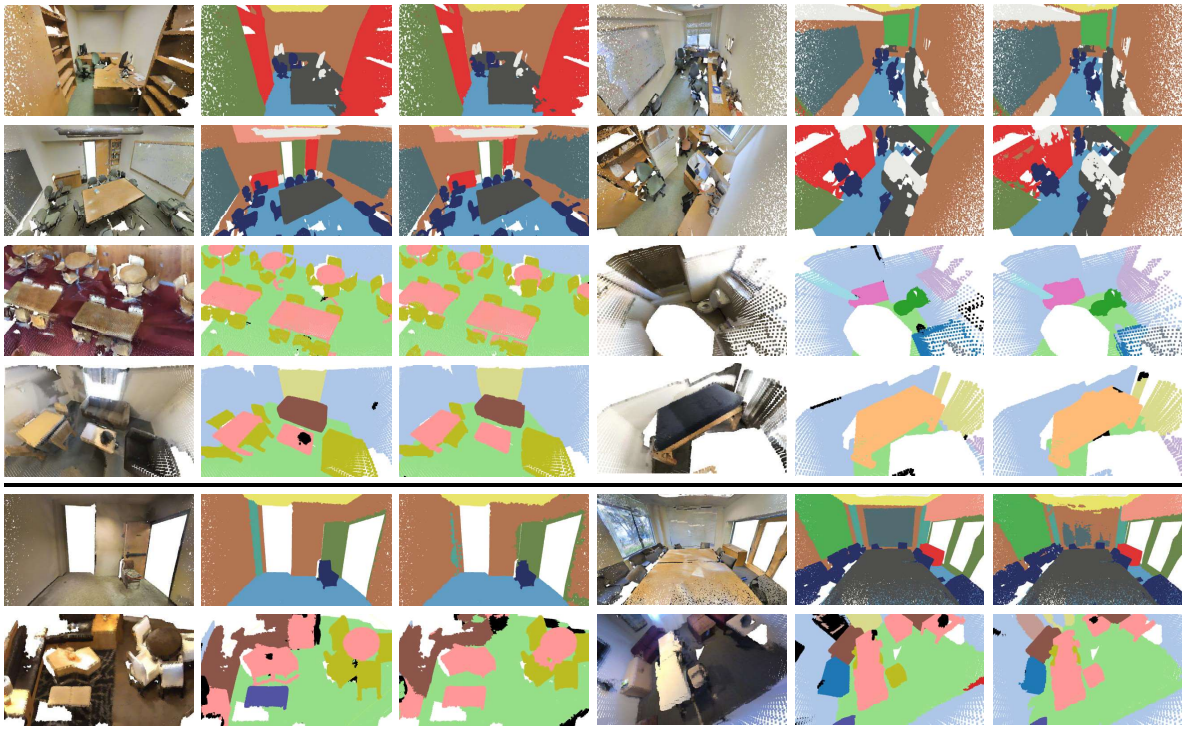

Figure 1. Examples of segmentation results on the ScanNet and S3DIS datasets. For each example, we show the input point cloud, the ground-truth segmentation, and our segmentation results in order. The first four rows show successful examples, while the last two rows show less successful ones.

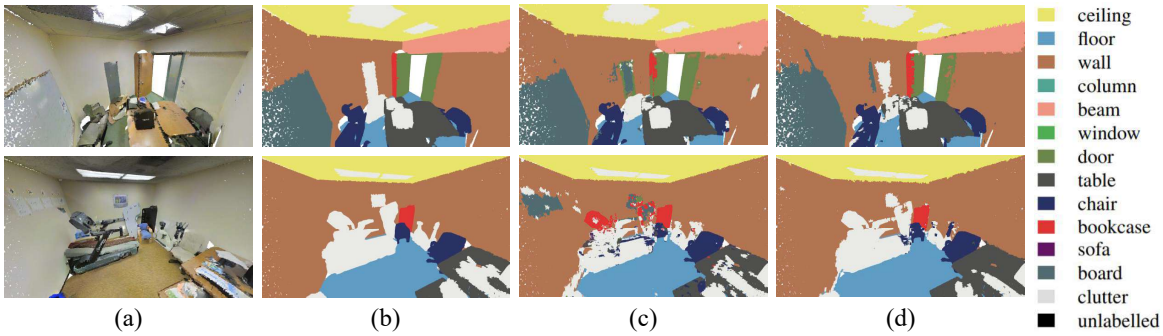

Figure 2. Examples of segmentation results on the S3DIS dataset. (a) Input point cloud, (b) Ground truth, (c) Xu *et al.*'s method [3], (d) Ours. Our method generates more accurate and smoother segmentation results than Xu *et al.*'s method.