PointCT: Point Central Transformer Network for Weakly-supervised Point Cloud Semantic Segmentation - Supplementary Material

Anh-Thuan Tran[1] Hoanh-Su Le[3,4] Suk-Hwan Lee[2] Ki-Ryong Kwon[1]
[1]Department of Artificial Intelligence Convergence, Pukyong National University, South Korea
[2]Department of Computer Engineering, Dong-A University, South Korea
[3]Faculty of Information Systems, University of Economics and Law, Ho Chi Minh City, Vietnam
[4]Vietnam National University, Ho Chi Minh City, Vietnam
thuantran@pukyong.ac.kr, skylee@dau.ac.kr, krkwon@pknu.ac.kr

In the appendix, we provide additional details to complement the main manuscript:

- Appendix 1: Qualitative experiment description and results in S3DIS 6-fold cross-validation, ScanNet-V2 and STPLS3D.
- Appendix 2: Complexity comparison on S3DIS Area-5.
- Appendix 3: Societal impact.
- Appendix 4: Limitations.
- Appendix 5: Visualization results on S3DIS Area-5, Scannet-V2 validation and STPLS3D.

1. Experiment details

Experiment environment. Software and hardware environment:

- CUDA version: 11.3
- PyTorch version: 1.10.1
- GPU: Nvidia RTX 2080 Ti $\times$ 2
- CPU: Intel(R) Core(TM) i9-9900K CPU @ 3.60GHz

Data license. The experiments are conducted with open-source datasets. S3DIS [1] has custom license that only allow for academic usage. ScanNet-V2 [4] is under MIT license, and STPLS3D [2] is under CC BY-NC-SA license (Creative Commons Attribution-NonCommercial-ShareAlike).

Data preprocessing. We adopt data processing and augmentation of Point Transformer [15] for S3DIS, Stratified Transformer [8] for ScanNet-V2, and STPLS3D from its original work [2]. Following previous studies [2, 12], we utilize data augmentation for these datasets.

Training details. Following SQN [6], which is designed to process purely 3D points in weak supervision, we assign unlabeled points with an appropriate unlabeled type during training. Cross-entropy loss is utilized across all experiments, with the unlabeled type being ignored. For evaluation, we use full point cloud scenes to test network performance.

Additional experimental results. More results on S3DIS Area-5, ScanNet-V2, and STPLS3D datasets are shown in Tables 1, 2, and 3, respectively. We add per-class experimental results in mIoU on all three datasets. By achieving significant performance on both indoor and outdoor point clouds, PointCT outperforms other weakly-supervised large-scale semantic segmentation methods purely based on 3D points by a large margin.

2. Complexity comparison

Table 4 describes computational costs compared to other works. We evaluate the complexity using two primary metrics, including number of parameters in millions (M) and floating-point operations (FLOPs) in gigabytes (G).

3. Societal impacts

While PointCT with central-based attention may require additional computational resources, we do not anticipate any immediate negative societal impact. Furthermore, our work in 3D weak supervision contributes to the community

Table 1. More results on S3DIS 6-fold cross-validation under 0.1% setting for point cloud semantic segmentation. Underline presents the best results under fully-supervised settings, and **Bold** shows the best results under weakly-supervised settings.

Settings	Method	mIoU	ceil.	floor	wall	beam	col.	wind.	door	chair	table	book.	sofa	board	clut.
100%	PointNet++ [10]	54.5	-	-	-	-	-	-	-	-	-	-	-	-	-
	RandLA-Net [7]	70.0	93.1	96.1	80.6	62.4	48.0	64.4	69.4	76.4	69.4	64.2	60.0	65.9	60.1
	PointTrans [15]	73.5	94.3	97.5	84.7	55.6	58.1	66.1	78.2	74.1	77.6	71.2	67.3	65.7	64.8
1%	Zhang et.al [13]	65.9	-	-	-	-	-	-	-	-	-	-	-	-	-
	PSD [14]	68.0	-	-	-	-	-	-	-	-	-	-	-	-	-
	HybridCR [9]	69.2	-	-	-	-	-	-	-	-	-	-	-	-	-
0.1%	SQN [6]	63.7	92.5	95.4	77.1	**50.8**	43.6	58.5	67.0	54.1	67.7	61.0	54.9	53.0	52.7
	PointCT	**71.2**	**94.6**	**97.1**	**83.8**	43.6	**51.9**	**59.6**	**79.0**	**83.2**	**71.3**	**65.4**	**68**	**62.8**	**65.5**

Table 2. More results on ScanNet-V2 test set under 0.1% setting for point cloud semantic segmentation. *Italic* presents the first row, and the other is the second row. Underline presents the best results under fully-supervised settings, and **Bold** shows the best results under weakly-supervised settings.

Setting	Method	mIoU	*bath*	*bed*	*bkshf*	*cab*	*chair*	*cntr*	*curt*	*desk*	*door*	*floor*
			other	pic	fridg	show	sink	sofa	table	toil	wall	wind
100%	PointNet++ [10]	33.9	*58.4*	*47.8*	*45.8*	*25.6*	*36.0*	*25*	*24.7*	*27.8*	*26.1*	*67.7*
			18.3	11.7	21.2	14.5	36.4	34.6	23.2	54.8	52.3	25.2
	RandLA-Net [7]	64.5	*77.8*	*73.1*	*69.9*	*57.7*	*82.9*	*44.6*	*73.6*	*47.7*	*52.3*	*94.5*
			45.4	26.9	48.4	74.9	61.8	73.8	59.9	82.7	79.2	62.1
1%	Zhang et.al [13]	51.1	-	-	-	-	-	-	-	-	-	-
			-	-	-	-	-	-	-	-	-	-
	PSD [14]	54.7	*57.1*	*67.8*	*65.9*	*46.5*	*77.8*	***38.8***	*52.8*	*49.2*	*30.4*	*93.3*
			38.7	30.7	43.1	38.2	52.6	66.9	57.2	71.6	60.9	50.6
	HybridCR [9]	56.8	*58.9*	*65.8*	*66.8*	*42.3*	*80.2*	*36.7*	***61.2***	***58.1***	*45.5*	*90.1*
			47.5	**33.4**	41.0	37.5	51.1	70.5	**60.8**	71.0	60.1	57.9
	PointCT	**64.3**	***79.0***	***76.5***	***70.7***	***60.7***	***83.8***	*30.9*	*47.7*	*54.7*	***54.9***	***94.1***
			49.0	28.8	**55.5**	**73.9**	**62.1**	**75.0**	57.3	**91.1**	**81.2**	**59.4**
0.1%	SQN [6]	56.9	*67.6*	*69.6*	*65.7*	*49.7*	*77.9*	***42.4***	*54.8*	***51.5***	*37.6*	*90.2*
			42.2	**35.7**	37.9	45.6	**59.6**	65.9	54.4	68.5	66.5	55.6
	PointCT	**63.1**	***79.1***	***72.5***	***70.5***	***62.8***	***83.5***	*35.8*	***60.0***	*47.5*	***53.0***	***94.3***
			49.9	16.7	**53.4**	**73.4**	51.7	**77.7**	**56.2**	**80.6**	**81.3**	**61.3**

by reducing manual labeling efforts. Therefore, it allows researchers to focus on other vital aspects, leading to greater diversity and generality in computer vision research.

4. Limitations

As shown in Table 5, although PointCT outperforms Point Transformer [15] in 0.01% and 1 point per class (1pt) settings by 37.5% and 32.5% in mIoU, respectively, we can observe the performance drops dramatically when the annotation level decreases to these levels. The reason behind this situation can be attributed to the fact that the proposed network processes raw limited labeled points without any additional supervision. Furthermore, the network relies on central-based attention mechanism to extract features from these points and their relationships to the unlabeled ones. In extremely low annotation settings, the model is incapable of learning enough information, thereby lowering generalizability and overall performance. Therefore, addressing these cases remains an open challenge for future research.

5. Visualization

In this section, we provide more visualization results in indoor S3DIS, Scannet-V2 and real-world STPLS3D. As seen from Figure 1 and 2, the segmentation performance achieves remarkable results at limited point annotations compared to ground truth (GT), which effectively captures primary features from limited labeled points. Furthermore, the proposed model can filter out noise points in outdoor scenes under weak supervision. Specifically, the resulting segmentation presented in Figure 3 is notably more explicit

Table 3. More results on STPLS3D for point cloud semantic segmentation. Underline presents the best results under fully-supervised settings, and **Bold** shows the best results under weakly-supervised settings.

Setting	Method	mIoU	ground	building	tree	car	light pole	fence
100%	KPConv [11]	<u>53.7</u>	<u>87.4</u>	<u>78.5</u>	<u>66.2</u>	<u>39.6</u>	41.3	9.3
	RandLA-Net [7]	50.5	82.9	66.6	63.8	33.9	41.8	14.2
	SCF-Net [5]	50.7	77.8	59.0	64.9	46.4	40.5	15.4
	MinkowskiNet [3]	51.4	80.9	74.0	59.2	31.7	<u>45.5</u>	<u>16.8</u>
	PointTrans [15]	47.6	80.2	76.4	57.1	36.4	23.7	12.1
0.1%	PointCT	49.2	**84.1**	**74.9**	**62.4**	30.4	28.1	**15.2**
0.01%	PointCT	**53.2**	80.3	72.5	57.2	**44.6**	**54.1**	10.2

Table 4. Computational cost.

Method	FLOPs (G)	Parameters (M)
PointNet++ [10]	7.2	1.0
RandLA-Net [7]	5.8	1.3
PointTrans [15]	5.6	7.8
PointCT	17.9	10.1

(yellow boxes), albeit different from the ground truth (GT).

References

[1] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1534–1543, 2016. 1

[2] Meida Chen, Qingyong Hu, Hugues Thomas, Andrew Feng, Yu Hou, Kyle McCullough, and Lucio Soibelman. Stpls3d: A large-scale synthetic and real aerial photogrammetry 3d point cloud dataset. In *British Machine Vision Conference*, 2022. 1

[3] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3075–3084, 2019. 3

[4] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 1

[5] Siqi Fan, Qiulei Dong, Fenghua Zhu, Yisheng Lv, Peijun Ye, and Fei-Yue Wang. Scf-net: Learning spatial contextual features for large-scale point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14504–14513, 2021. 3

[6] Qingyong Hu, Bo Yang, Guangchi Fang, Yulan Guo, Aleš Leonardis, Niki Trigoni, and Andrew Markham. Sqn: Weakly-supervised semantic segmentation of large-scale 3d point clouds. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, pages 600–619. Springer, 2022. 1, 2

[7] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11108–11117, 2020. 2, 3

[8] Xin Lai, Jianhui Liu, Li Jiang, Liwei Wang, Hengshuang Zhao, Shu Liu, Xiaojuan Qi, and Jiaya Jia. Stratified transformer for 3d point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8500–8509, 2022. 1

[9] Mengtian Li, Yuan Xie, Yunhang Shen, Bo Ke, Ruizhi Qiao, Bo Ren, Shaohui Lin, and Lizhuang Ma. Hybridcr: Weakly-supervised 3d point cloud semantic segmentation via hybrid contrastive regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14930–14939, 2022. 2

[10] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 2, 3

[11] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6411–6420, 2019. 3

[12] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. In *Advances in Neural Information Processing Systems*, 2022. 1

[13] Yachao Zhang, Zonghao Li, Yuan Xie, Yanyun Qu, Cuihua Li, and Tao Mei. Weakly supervised semantic segmentation for large-scale point cloud. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3421–3429, 2021. 2

[14] Yachao Zhang, Yanyun Qu, Yuan Xie, Zonghao Li, Shanshan Zheng, and Cuihua Li. Perturbed self-distillation: Weakly supervised large-scale point cloud semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15520–15528, 2021. 2

Table 5. More results on S3DIS Area-5 under extremely-low labeled point settings. **Bold** shows the best results under weakly-supervised settings.

Settings	Method	mIoU	ceil.	floor	wall	beam	col.	wind.	door	chair	table	book.	sofa	board	clut.
0.01%	PointCT	**39.7**	75.2	96.8	56.0	0.0	5.9	18.6	20.4	**54.1**	**81.1**	**11.9**	**42.5**	**21.8**	**31.6**
1pt	PointCT	34.7	**82.2**	**92.7**	**70.9**	0	**24.3**	**36.1**	**23.2**	35.1	52.3	0.0	0.0	0.1	0.2
	PointTrans [15]	2.2	0.0	0.0	29.2	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

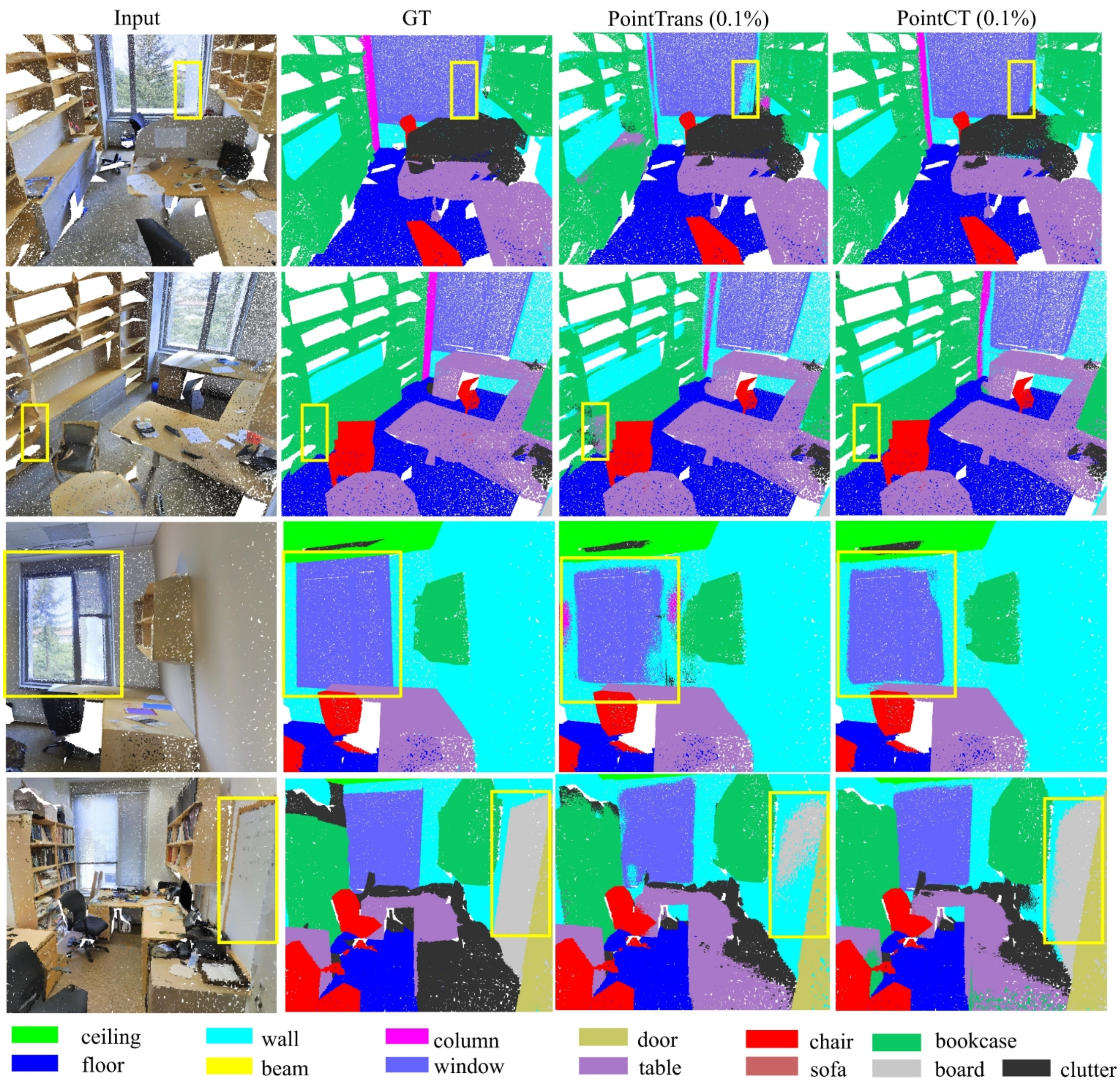

Figure 1. Visualization on indoor S3DIS Area-5.

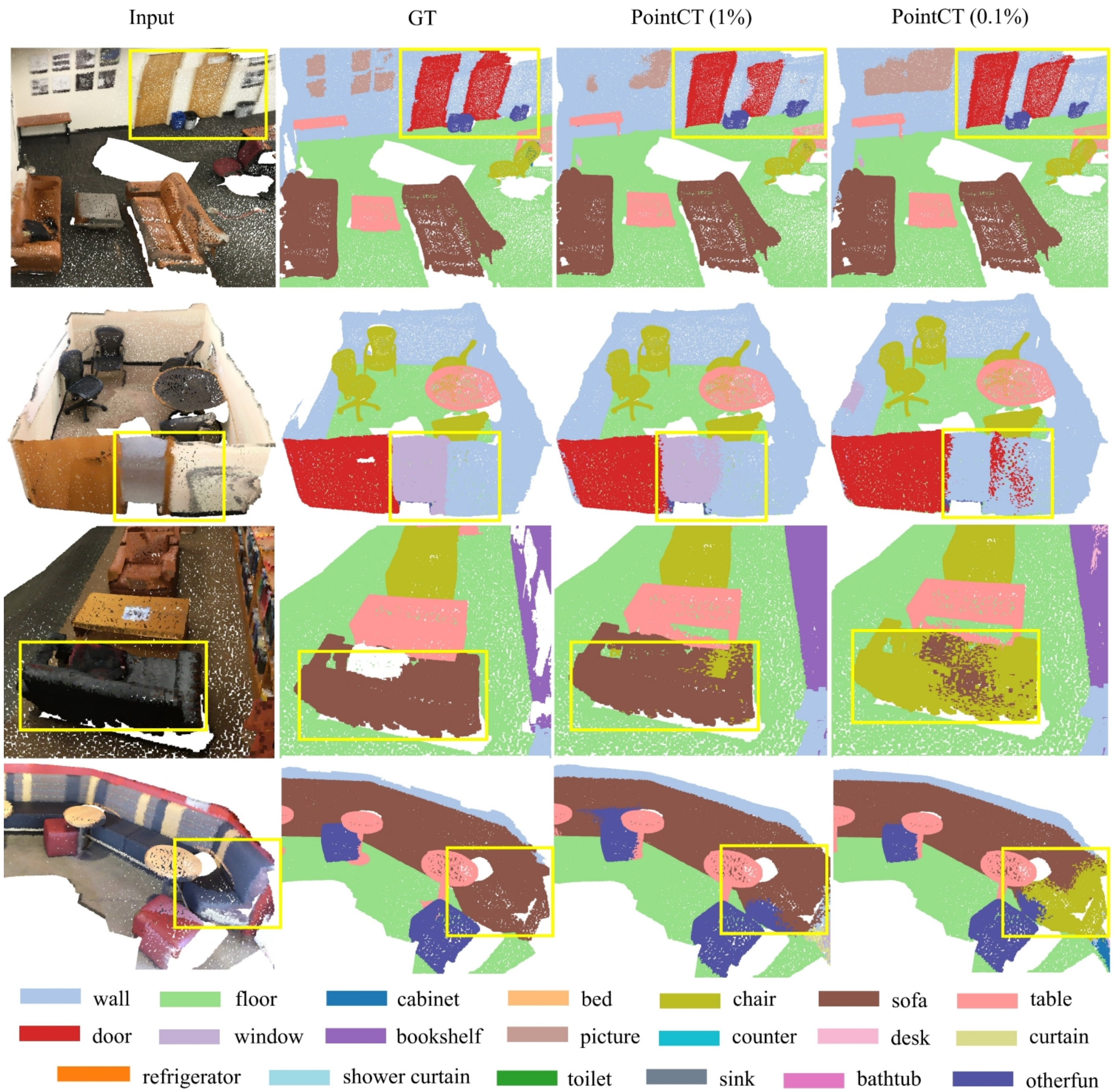

Figure 2. Visualization on Scannet-V2 validation.

[15] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021. 1, 2, 3, 4

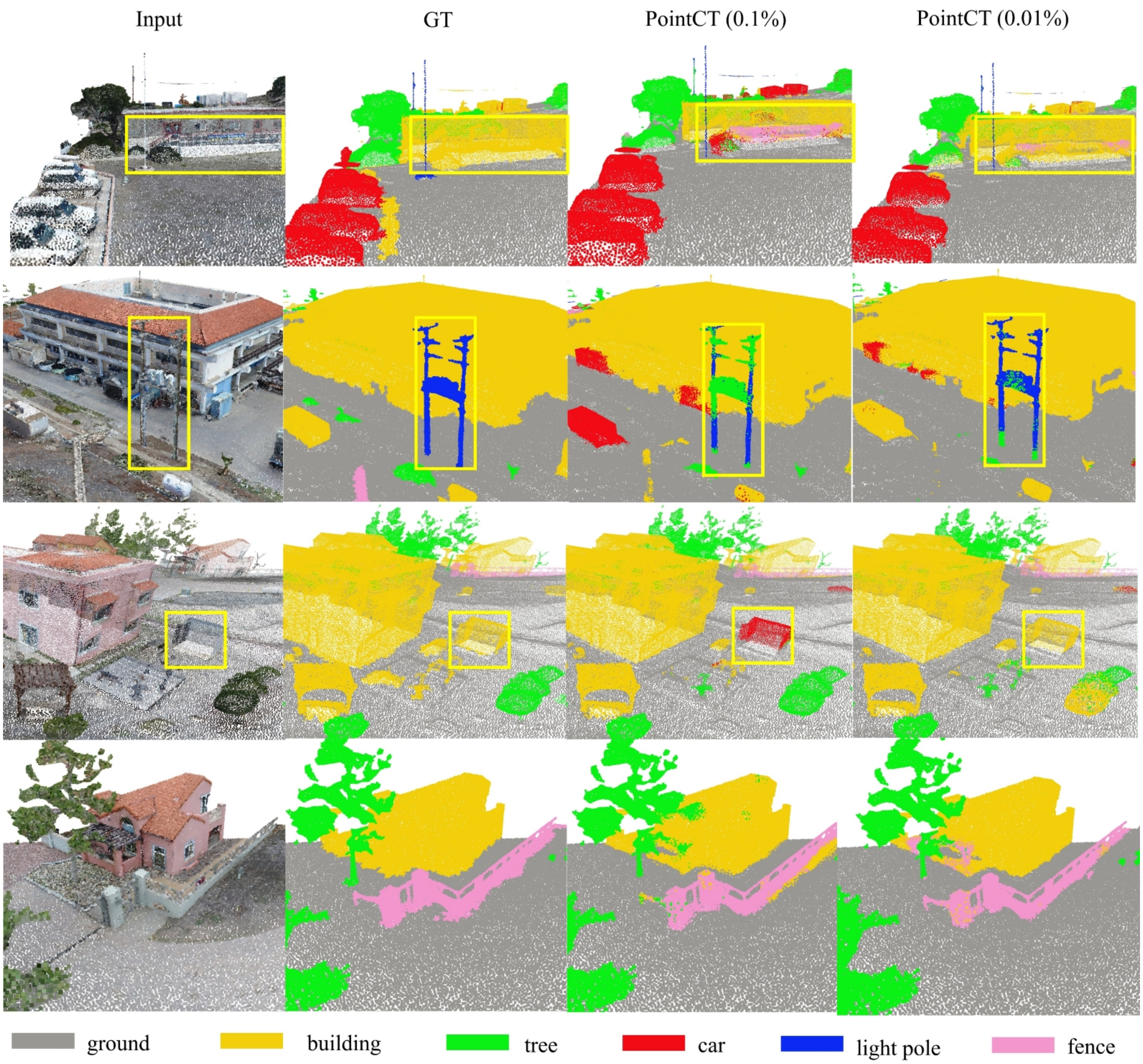

Figure 3. Visualization on real-world STPLS3D.