Weakly Supervised Point Cloud Semantic Segmentation via Artificial Oracle: Supplementary Materials

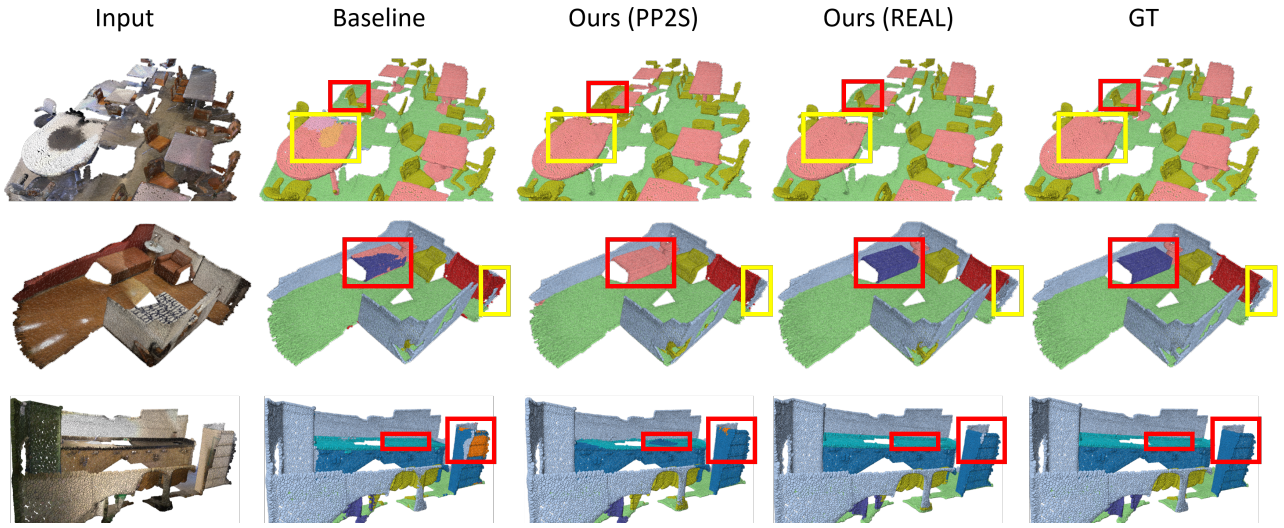

Figure 1. Qualitative comparisons between the semantic segmentation results of the baseline and ours on ScanNetV2 in 20 pts setting. The displayed boxes indicate the regions where significant improvements are achieved by the proposed method.

1. Implementation Details

For optimization, we utilize the SGD optimizer with a learning rate of $1e^{-2}$, weight decay of $1e^{-3}$, and a batch size of 16 for training with Closer [3]. In the case of PTv2 [4], we employ the AdamW optimizer with a learning rate of $5e^{-3}$, weight decay of $2e^{-2}$, and a batch size of 12 on both S3DIS [1] and ScanNetV2 [2]. The experiments are conducted using 4 Quadro RTX 8000(s) for PTv2 and 1 Quadro RTX 8000 for Closer.

2. Qualitative Results on ScanNetV2

We evaluate the performance of our method against SoTA methods on ScanNetV2 [2]. We provide qualitative results of experiments conducted in 20 pts setting illustrated in Fig. 1. Semantic confusion in single objects is observed in both the baseline and PP2S. While these two cases exhibit different types of confusion, both are notably improved in REAL. Specifically, the segmentation precision is significantly enhanced, as shown in the boxes. Consequently, we can conclude that REAL achieves a remarkable improvement in semantic segmentation capability.

3. More Ablation on Query Sampling Strategy

To further assess the effectiveness of the query sampling strategy, we quantitatively provide the variation of label Y, as shown in Fig. 2. Throughout the training process, we track the mean precision, mean recall, and mIoU of Y obtained by the Setting I and II mentioned in the method section of the main paper. Notably, there are no significant differences in precision between the two settings. However, for recall and mIoU, setting (I) consistently outperforms setting (II). Specifically, (I) surpasses (II) by approximately 8.5 in recall and 7.0 in mIoU once they reach saturation. Note that the results strongly support the design intention of our query sampling strategy, exploring the new region without harming the accuracy. These indicate that the proposed query sampling strategy consistently produces improved labels, facilitating enhanced segmentation learning by the model.

References

[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. 1

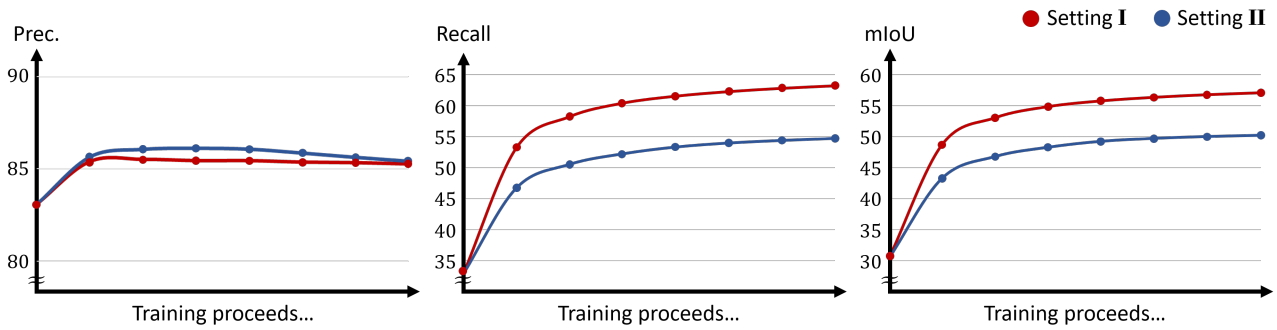

Figure 2. Verification of our query sampling strategy. Setting **I** (Red) employs our strategy, while Setting **II** (Blue) samples from the entire point set. The figure illustrates the variation in precision, recall, and mIoU respectively. Although precision does not exhibit significant differences, Setting **I** demonstrates superior performance in terms of recall and mIoU.

[2] 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

[3] Ze Liu, Han Hu, Yue Cao, Zheng Zhang, and Xin Tong. A closer look at local aggregation operators in point cloud analysis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 326–342. Springer, 2020. 1

[4] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. In *NeurIPS*, 2022. 1