Supplementary Material: Hierarchical Intra-modal Correlation Learning for Label-free 3D Semantic Segmentation

1. Performance on ScanNet

We offer a comprehensive evaluation of semantic segmentation performance for each category, using our method and CLIP2Scene [2], on the ScanNet validation set, as presented in Table 1. Since the pre-trained model of CLIP2Scene is not yet available, we reproduced its semantic segmentation results using the released code. All models were trained for 120 epochs under the same training settings to ensure a fair comparison. In comparison to CLIP2Scene, our method achieves substantial improvements across multiple categories, including *Wall* (+12.50% IoU), *Bed* (+29.23% IoU), *Bookshelf* (+26.97% IoU), *Counter* (+11.10% IoU), *Sink* (+13.3% IoU), and *Bathtub* (+27.23% IoU). Moreover, we offer more visualization results of our method and CLIP2Scene [2] on the ScanNet validation set in Figure 3. Our method makes more accurate predictions compared to CLIP2Scene, as evidenced by the *Sink* in the first row, the *Window* in the third row, and the *Bookshelf* in the last row. Furthermore, our method generates more consistent predictions, as demonstrated by the *Sofa* in the second and fourth rows, and the *Chair* in the fifth row. These quantitative and qualitative results highlight the superiority of our approach.

2. Performance on nuScenes

In Tab. 2, we present semantic segmentation performance for each category of our method and CLIP2Scene [2] on the nuScenes validation set. The results of CLIP2Scene are reproduced using the published code. Compared with CLIP2Scene, our method demonstrates a total gain of 3.28% mIoU, with notable improvements across the majority of categories. For a more comprehensive understanding, we provide qualitative visualization results in Figure 4. Our method consistently outperforms CLIP2Scene in terms of accuracy, as illustrated by the *Manmade* in the first row, *Vegetation* in the second row, and *Sidewalk* in the fourth row. Moreover, our method exhibits a higher coherence in its predictions, as seen in the *Motorcycle* in the third row and *Bus* in the fifth row. These quantitative and qualitative results on the nuScenes dataset demonstrate both the effectiveness and the generalizability of our method.

3. Ablation of Geometric Consistency Set

Our hierarchical intra-modal correlation learning framework is constructed based on geometric consistency sets. In this work, we employ a normal-based over-segmentation algorithm [3, 5] to generate geometric consistency sets for 3D scenes. The over-segmentation results with varying parameters on the ScanNet dataset are displayed in Figure 1. As depicted in Figure 1, the 3D scene is divided into numerous segments, each representing a small part of an object. A larger clustering edge weight threshold results in larger segments. In addition, we tested our model using different over-segmentation parameters to assess its stability. The results are presented in Table 3. Each model was trained for 30 epochs under the same training settings. These results confirm that our method exhibits strong robustness to variations in clustering parameters.

To further demonstrate the adaptability of our approach with different geometry consistency sets, we conduct experiments utilizing a normal-based partitioning method [5] and a geometrically homogeneous partitioning algorithm [4]. The results presented in Table 4 demonstrate the robustness of our method with various geometry consistency sets.

Figure 1. Visualization of geometric consistency sets with different parameters on ScanNet dataset. Different colors indicate different geometric consistency sets.

4. Visualization of Attention Maps

We show the learned intra- and inter-scene feature correlations by visualizing the attention maps in Fig. 2. Given a query point labeled as *Chair*, the red-marked points indicate areas of higher correlation, while those marked in blue represent areas with lower correlation. The attention maps show that our intra- and inter-scene correlation learning module can effectively capture point-wise semantic cor-

Method	**mIoU**	Wall	Floor	Cabinet	Bed	Chair	Sofa	Table	Door	Window
CLIP2Scene⋆ [2]	28.75	52.80	89.37	3.13	27.97	65.20	52.80	47.00	23.87	23.67
Ours	**36.60**	**65.30**	**92.63**	**6.20**	**57.20**	**71.77**	**58.03**	**50.37**	**30.83**	**24.03**
Bookshelf	Picture	Counter	Desk	Curtain	Refrig	S curtain	Toilet	Sink	Bathtub	Other
24.27	10.30	9.33	20.47	4.80	2.90	0.00	50.97	26.23	39.83	0.03
51.23	**13.63**	**20.43**	**25.27**	**14.40**	**5.40**	**0.20**	38.53	**39.53**	**67.07**	0.00

Table 1. Per-class IoU(%) and mIoU(%) on the ScanNetV2 validation set. ⋆ indicates results are reproduced by us. *Refrig* and *S curtain* are the abbreviations for *refrigerator* and *shower curtain*, respectively.

Method	**mIoU**	Barrier	Bicycle	Bus	Car	Construction	Motorcycle	Pedestrian
CLIP2Scene⋆ [2]	19.73	8.60	10.70	67.50	27.70	28.30	11.60	0.50
Ours	**23.01**	**9.10**	**11.10**	**69.60**	**29.80**	**28.40**	**47.50**	0.30
TrafficCone	Trailer	Truck	DriveableSurface	OtherFlat	Sidewalk	Terrain	Manmade	Vegetation
28.80	0.00	54.70	0.00	0.10	20.10	0.00	2.10	55.00
23.10	0.00	**57.20**	0.00	**0.20**	**31.10**	0.00	**2.90**	**57.90**

Table 2. Per-class IoU(%) and mIoU(%) on the nuScenes validation set. ⋆ indicates results are reproduced by us.

Threshold	0.01	0.02	0.03	0.04	0.05
mIoU	30.06	30.14	30.10	30.04	29.94

Table 3. Semantic segmentation results on ScanNet dataset with different over-segmentation parameters. All models are trained for 30 epochs using the same training setting.

Methods	ScanNet mIoU	nuScenes mIoU
Papon et al. [5]	36.6	23.0
Landrieu et al. [4]	35.5	22.4

Table 4. Semantic segmentation results on ScanNet dataset with alternative geometric consistency sets.

respondences. Utilizing these correspondences, the cross-entropy loss function encourages features of high-correlated points to be compact and consistent during training, leading to a consistent feature distribution both intra- and inter-scene.

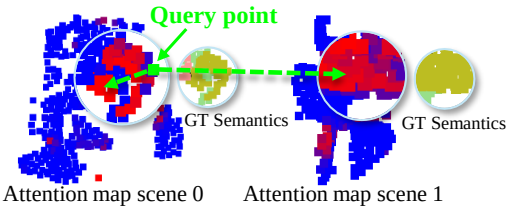

Figure 2. Visualization of attention maps on ScanNet dataset.

5. Comparison of Data Processing Cost

In the pseudo-label refinement stage, we adopt a normal-based method [5] to generate over-segmentation masks, which is different from Chen et al. [1] who use SAM to generate segmentation masks. To compare the computational efficiency, we test the date processing time on 100 scenes from ScanNet using an A100 GPU, finding that SAM processes each scene in an average of 22 seconds, while the normal-based method averages only 3 seconds per scene. The result underscores the substantial computational overhead incurred when utilizing SAM for data processing.

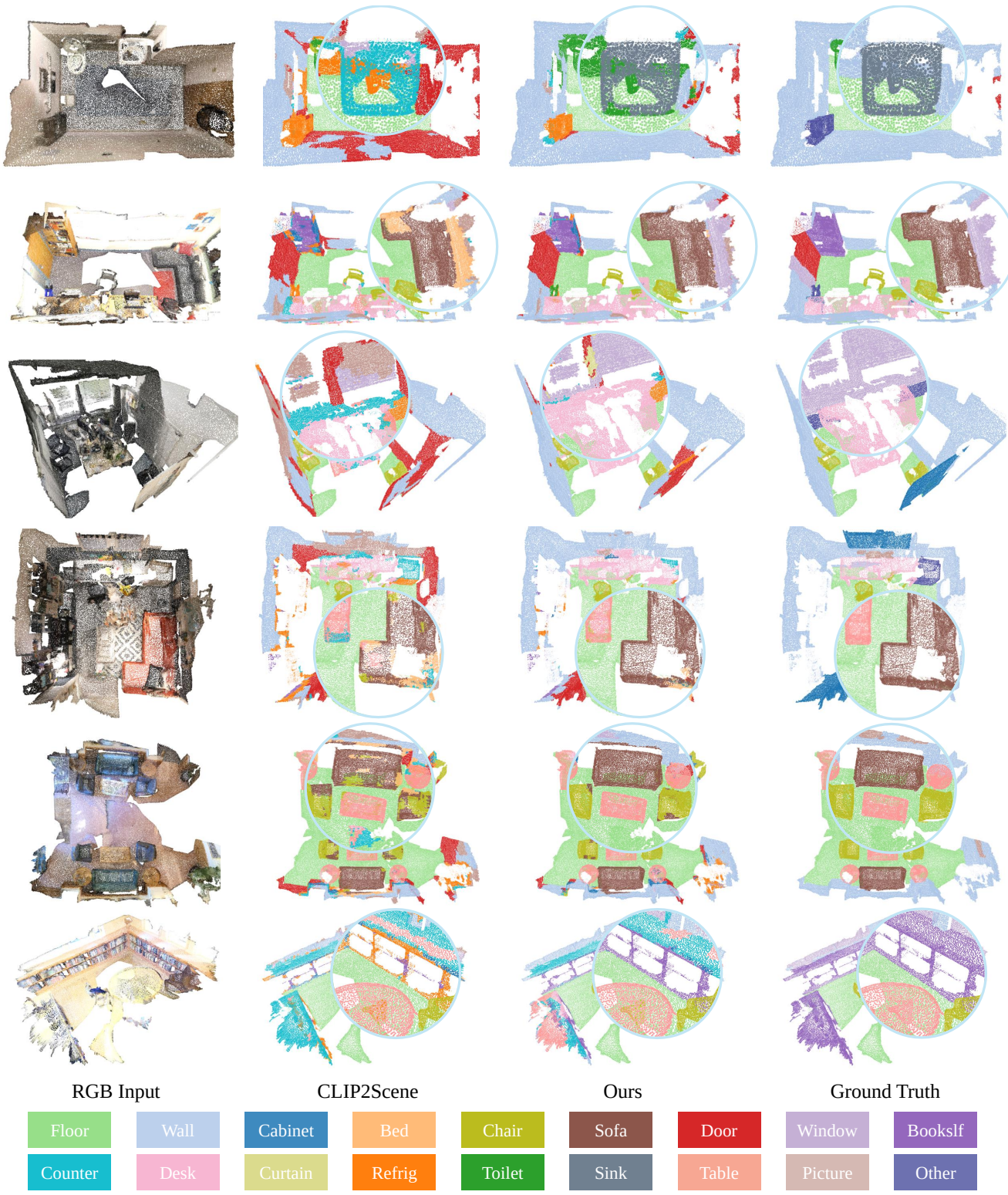

Figure 3. Qualitative comparison for semantic segmentation of our method and CLIP2Scene [2] on the ScanNet dataset.

Figure 4. Qualitative comparison for semantic segmentation of our method and CLIP2Scene [2] on the nuScenes dataset.

References

[1] Runnan Chen, Youquan Liu, Lingdong Kong, Nenglun Chen, ZHU Xinge, Yuexin Ma, Tongliang Liu, and Wenping Wang. Towards label-free scene understanding by vision foundation models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[2] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7030, 2023. 1, 2, 3, 4

[3] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59:167–181, 2004. 1

[4] Loic Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4558–4567, 2018. 1, 2

[5] Jeremie Papon, Alexey Abramov, Markus Schoeler, and Florentin Worgotter. Voxel cloud connectivity segmentation-supervoxels for point clouds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2027–2034, 2013. 1, 2