Supplementary Materials for GeoFree-CoSeg: Unsupervised Point Cloud-Image Cross-Modal Co-Segmentation Without Geometric Alignment

Abstract

In this supplementary material, we give the omitted contents (for space reasons) indicated in the main paper. Specifically, i) more ablation studies and comparisons are given in Sec. 1; ii) more discussions of our model is given in Sec. 2; and iii) more details of the proposed two new indoor image co-segmentation datasets are given in Sec. 3.

1. More Ablation Study and Comparisons

Semantic Fusion and Enhancement Analysis. To get a deeper understanding of our 2D-3D common semantic fusion and enhancement method, we visualize the affinity correlation heatmaps in Fig. 2. For point clouds, the basic 3D branch (Sec. 3.1) captures coarse and inaccurate correlations, with attentions deviating from the true common object regions. After regularization by 2D common semantics, the fused 3D features effectively alleviate the influence of co-occurring noise and focus on common regions in the point cloud groups. For example, in the 'Bookcase' set, the basic 3D branch attends to background noises, whereas the fused 3D branch focuses on the common objects. For images, the basic 2D branch (Sec. 3.1) tends to focus on texture features, while the fused 2D branch exhibits more semantically aligned responses. For example, in the 'Sofa' set, although a texture-salient bicycle appears in front of the sofa, the fused 2D branch correctly attends to the common semantic region. These visualizations demonstrate that our 2D-3D common semantic fusion and enhancement strategy effectively strengthens cross-modal semantic consistency and improves feature discriminability.

Ablation of Hyperparameters. We analyze the hyperparameter λ in the 2D multi-granularity correlation module (MGCM) in Fig 1. By varying λ from 0.1 to 1.0, we evaluate the performance and identify the stable range. Both metrics peak around $\lambda = 0.2$, while $\lambda > 0.5$ leads to excessive emphasis on global semantics, limiting the ability of the model to capture fine details. The model achieves stable and optimal performance at $\lambda = 0.2$.

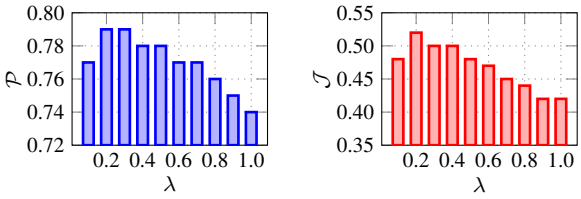

Figure 1. *Ablation study of λ in MGCM: (left) Precision $\mathcal{P}$, (right) Jaccard Index $\mathcal{J}$.*

Table 1. *Ablation of bidirectional graphs in GCSF (Graph-based Common Semantic Filtering).*

Method	mIoU	$\mathcal{P}$	$\mathcal{J}$
$\mathcal{G}^{I\to P}$-only	0.49	0.81	0.55
$\mathcal{G}^{P\to I}$-only	0.52	0.79	0.53
$\mathcal{G}^{P\to I}$ + $\mathcal{G}^{I\to P}$ (Ours)	**0.54**	**0.83**	**0.59**

Ablation of Bidirectional Graphs in GCSF. We further assess the contribution of bidirectional graphs in graph-based common semantic filtering (GCSF) by evaluating unidirectional variants in Tab. 1. For the $\mathcal{G}^{I\to P}$-only and $\mathcal{G}^{P\to I}$-only settings, the missing correlation matrix is replaced by the transpose of its counterpart to keep the pipeline functional. Using bidirectional graphs consistently yields superior performance: it improves 5% mIoU, 2% $\mathcal{P}$ and 4% $\mathcal{J}$ over the $\mathcal{G}^{I\to P}$-only variant, and improves 2% mIoU, 4% $\mathcal{P}$ and 6% $\mathcal{J}$ over the $\mathcal{G}^{P\to I}$-only variant. Experiments confirm that simply using the transpose cannot substitute the bidirectional formulation, as $\mathcal{Z}^{P\to I}$ and $\mathcal{Z}^{I\to P}$ are inherently not transpose-equivalent, reflecting that the two direction graphs encode asymmetric and complementary cross-modal semantics. Therefore, leveraging both directions leads to the best performance.

Ablation of Soft-mask Strategy in GCSF. We analyze the effectiveness of the soft-mask strategy used in the cross-modal common semantic filtering of our GCSF module, as

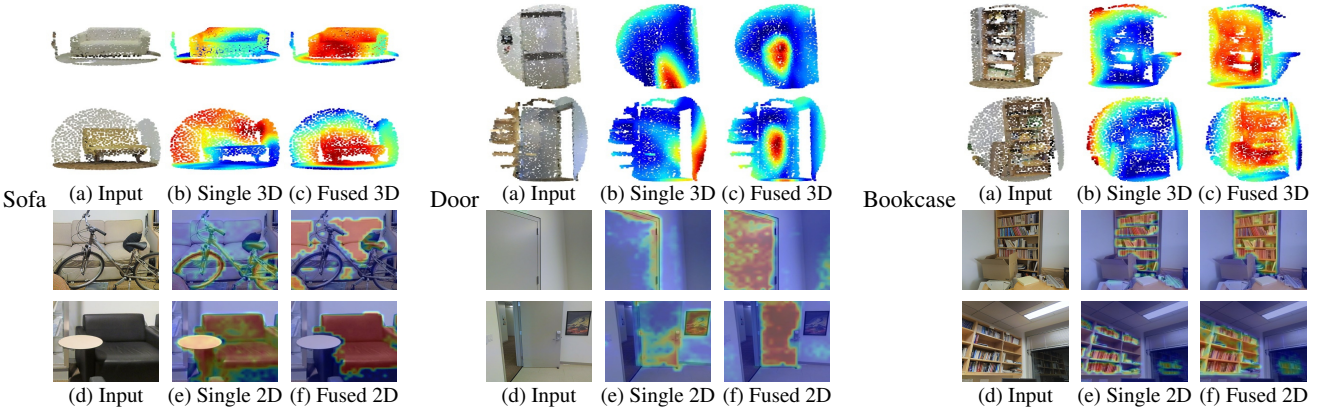

Figure 2. *Visualization of affinity correlation heatmaps for point clouds and images before and after the cross-modal common semantic fusion and enhancement, illustrating the effectiveness of incorporating cross-modal common semantics into both modalities. (a) Input point clouds. (b) Heatmaps from the basic 3D branch (Sec. 3.1.), using only single-modality point cloud features. (c) Heatmaps from the 3D branch after fusion with cross-modal common semantics. (d) Input images. (e) Heatmaps from our 2D branch (Sec. 3.1.), using only single-modality image features. (f) Heatmaps from our 2D branch after fusion with cross-modal common semantics.*

Table 2. *Ablation of soft-mask strategy in GCSF (Graph-based Common Semantic Filtering).*

Method	mIoU	$\mathcal{P}$	$\mathcal{J}$
Hard Mask	0.51	0.80	0.53
Soft Mask (Ours)	**0.54**	**0.83**	**0.59**

shown in Tab. 2. We replace the soft mask with a hard mask that either fully suppresses or fully retains each selected point and patch. Using the soft-mask strategy improves performance by 3% mIoU, 3% $\mathcal{P}$, and 6% $\mathcal{J}$ over the hard-mask variant, indicating that soft masks preserve weak yet informative semantic cues and thereby facilitate cross-modal co-segmentation.

2. More Discussions

We benchmark end-to-end inference time and FPS on point clouds with 2048 points and 224 $\times$ 224 images using a single NVIDIA RTX 4090 GPU. As shown in Tab. 3, our method is slightly slower than Yang [12] on point clouds and slightly slower than SCoSPARC [4] on images, due to the cross-modal common semantic filtering and fusion. LogoSP [14] and GrowSP [13] are slower because of point cloud voxelization, while DVFDVD [1] exhibits low FPS due to its complex iterative process of extracting image features and saliency maps and voting for cluster labels. Despite the slight increase in inference time, our method achieves substantially improved co-segmentation performance on both modalities, as reported in the main paper.

Table 3. *Inference time and FPS of 2D and 3D modalities.*

Modality	Method	Speed (s)	FPS (it/s)
3D	Yang [12]	0.010	105.263
	GrowSP [13]	0.037	27.327
	LogoSP [14]	0.036	27.861
	Ours	0.014	71.015
2D	DVFDVD [1]	0.354	2.826
	ReClip [11]	0.035	28.267
	SCoSPARC [4]	0.019	52.355
	Ours	0.026	38.523

3. New Image Co-Segmentation Datasets

Since cross-modal co-segmentation remains largely unexplored, no existing 2D co-segmentation datasets align with the categories of S3DIS [2] or ScanObjectNN [10]. To fill this gap, we construct two indoor image co-segmentation datasets, S3DIS-Coseg and ScanObjectNN-Coseg, for cross-modal co-segmentation: (i) S3DIS-Coseg is built from the 2D images in the 2D-3D-S dataset [3], which is the cross-modal extension of S3DIS [2]. We randomly sample images corresponding to S3DIS categories and apply random width and height to crop the images for the common objects in each category from the large and redundant image set. The dataset contains 7,642 training images and 1,102 evaluation images across 5 categories. (ii) ScanObjectNN-Coseg is constructed using images sourced from COCO [8], ScanNet [5], and ADE20K [15] datasets to match the rare indoor categories in ScanObjectNN [10]. Following the same procedure as S3DIS-Coseg, we randomly sample and crop images in the large and redundant categories. This dataset includes 6,545 training images and 1,337 evaluation images across 15 categories. Fig. 3 and

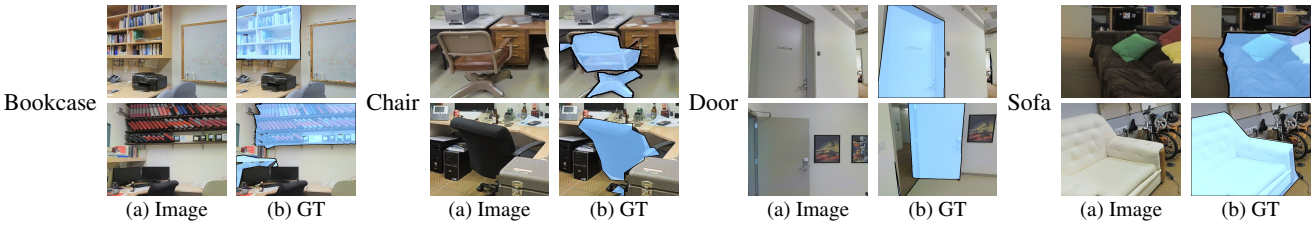

Figure 3. *Visualization of the S3DIS-Coseg dataset. (a) Images. (b) The ground-truth (GT) image co-segmentation masks.*

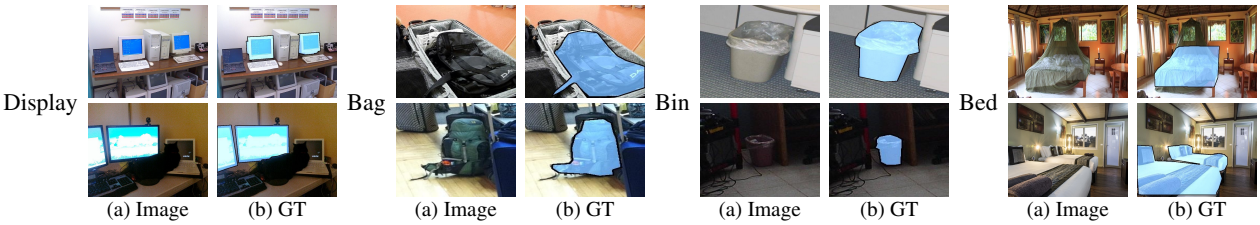

Figure 4. *Visualization of the ScanObjectNN-Coseg dataset. (a) Images. (b) The ground-truth (GT) image co-segmentation masks.*

Fig. 4 present examples of the images and their corresponding ground-truth co-segmentation masks in the datasets.

4. Future Work

Following prior work [12], we evaluate our method on the standard indoor benchmarks, S3DIS and ScanObjectNN. Extending the method beyond indoor datasets is an important direction for future work. Consistent with prior image co-segmentation methods [6, 7, 9] and point cloud co-segmentation methods [12], our framework focuses on inputs containing common semantic objects of an unknown category. Scaling our method to handle scenes with multiple co-occurring objects constitutes a significant direction for future exploration.

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep ViT features as dense visual descriptors. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, 2022. 2

[2] 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 (CVPR)*, pages 1534–1543, 2016. 2

[3] 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. 2

[4] Souradeep Chakraborty and Dimitris Samaras. Self-supervised co-salient object detection via feature correspondences at multiple scales. In *European Conference on Computer Vision (ECCV)*, pages 231–250. Springer, 2024. 2

[5] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 2

[6] Xin Duan, Yan Yang, Liyuan Pan, and Xiabi Liu. LCCo: Lending clip to co-segmentation. *Pattern Recognition*, 161: 111252, 2025. 3

[7] Xin Duan, Yan Yang, Liyuan Pan, Xiabi Liu, and Mingyang Gong. SZCo: Self-supervised zero-shot co-segmentation with region-text alignment learning. *Pattern Recognition*, 177:113308, 2026. 3

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014. 2

[9] Yukun Su, Jingliang Deng, Ruizhou Sun, Guosheng Lin, Hanjing Su, Qingyao Wu, et al. A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection. *IEEE Transactions on Multimedia*, pages 1–13, 2023. 3

[10] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1588–1597, 2019. 2

[11] Jingyun Wang and Guoliang Kang. Learn to rectify the bias of clip for unsupervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4102–4112, 2024. 2

[12] Cheng-Kun Yang, Yung-Yu Chuang, and Yen-Yu Lin. Unsupervised point cloud object co-segmentation by co-contrastive learning and mutual attention sampling. In

Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 7335–7344, 2021. 2, 3

[13] Zihui Zhang, Bo Yang, Bing Wang, and Bo Li. Growsp: Unsupervised semantic segmentation of 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17619–17629, 2023. 2

[14] Zihui Zhang, Weisheng Dai, Hongtao Wen, and Bo Yang. LogoSP: Local-global grouping of superpoints for unsupervised semantic segmentation of 3D point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1374–1384, 2025. 2

[15] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 2