Supplementary Material: Unsupervised Object Co-segmentation by Co-contrastive Learning and Mutual Attention Sampling

Cheng-Kun Yang[1] 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 experiments to validate the robustness and effectiveness of our proposed method. In the following, we first present the results with different pre-training methods for the feature extractor. Then, we study the sensitivity of the proposed co-contrastive learning with different batch sizes. Finally, we report the size and running time of the proposed method. More visualization examples, including both the successful and less successful cases, are added for qualitative evaluation.

1. Feature extractor pre-training

In Section 4.2 of the submitted paper, we experiment with different datasets for pre-training. Here we provide the detailed results. Moreover, we experiment with the unsupervised pre-training method, PointContrast [8], for being the feature extractor. This method requires only the point cloud data as input for learning features, which improves generalization on downstream tasks. Table 1 shows the result and reveals that the feature extractor pre-trained on the ScanObjectNN dataset can yield the best performance. However, it requires enormous effort for the point-level annotations. Compared to the ScanObjectNN dataset, ModelNet40 (a synthetic CAD dataset) shows comparable performance for feature extractor pre-training, with a much lower annotation cost. Finally, the unsupervised method pretrained on both datasets shows inferior performance since it can not capture the object features well.

2. Sensitivity of the batch size

The batch size is a crucial hyperparameter to contrastive learning. Previous work [2,3] reports that contrastive learning benefits from a large batch size. In Figure 1, we run the proposed method with different batch sizes on each variant of ScanObjectNN. In this work, we set the batch size to 48 because of the limited computing resource, though a larger batch size can further increase the performance.

Table 1: Segmentation results in mIoU on the OBJ_BG test set of ScanObjectNN by using different datasets for pretraining. PointContrast denotes the unsupervised pretraining on ModelNet40, while PointContrast† denotes pretraining on the ScanObjectNN OBJ_BG dataset.

	OBJ_BG test set
ScanObjectNN [5]	0.645
ModelNet40 [1]	0.605
PointContrast [8]	0.441
PointContrast† [8]	0.419

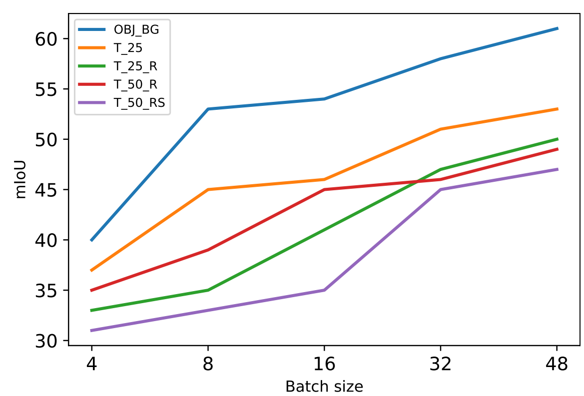

Figure 1: mIoU of different batch sizes on each variant of the ScanObjectNN dataset.

	Parameters	Time	FLOPs
Training	9.8M (frozen)+1.6M	9.2 ms	1950 M
Inference	0.8M	2.7 ms	220 M

Table 2: # of parameters, running time and FLOPs of our method.

3. Model size and running time

The proposed method is efficient in both training and inference. As mentioned in Section 3.5, DGCNN is pretrained and frozen during training and discarded during inference. The two samplers are with the PointNet-like structure combined with a lightweight mutual attention module. Only the object sampler is required for inference. Table 2 reports the numbers of model parameters and FLOPs, and the running time per point cloud of our model in an epoch. The number of epochs is set to 500 in the experiments.

4. Class-wise quantitative performance

In Table 3 of the submitted paper, we report our method's average performance for object segmentation on the four variants of the ScanObjectNN dataset. Here we provide the detailed class-wise results. Table 3, Table 4, Table 5, and Table 6 show the class-wise performance of our method on the PB_T25, PB_T25_R, PB_T50_R and PB_T50_RS datasets, respectively.

5. Qualitative results

Figure 2 and Figure 3 display the results generated by our method for object co-segmentation on point clouds of the S3DIS and ScanOjbectNN datasets, including both successful cases and less successful cases. It can be observed that the proposed method is robust to different background points in the point clouds. For example, in Figure 3, the background points present in the chair category belong to the box and table. Our approach can still segment the chair object accurately. Moreover, our method can segment the object with very different scales or views of the object. For example, in Figure 2, the chair and bookcase are diverse in their scales. Our method can still segment the object points correctly.

References

[1] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 1

[3] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1

[4] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NIPS*, 2017. 3, 4

[5] 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 *ICCV*, 2019. 1

[6] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *TOG*, 2019. 3, 4

[7] Jiacheng Wei, Guosheng Lin, Kim-Hui Yap, Tzu-Yi Hung, and Lihua Xie. Multi-path region mining for weakly supervised 3d semantic segmentation on point clouds. In *CVPR*, 2020. 3, 4

[8] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pretraining for 3d point cloud understanding. In *ECCV*, 2020. 1

[9] Xun Xu and Gim Hee Lee. Weakly supervised semantic point cloud segmentation: Towards 10x fewer labels. In *CVPR*, 2020. 3, 4

[10] Chenyang Zhu, Kai Xu, Siddhartha Chaudhuri, Li Yi, Leonidas J Guibas, and Hao Zhang. Adacoseg: Adaptive shape co-segmentation with group consistency loss. In *CVPR*, 2020. 3, 4

Table 3: Segmentation results (mIoU) on ScanObjectNN PB_T_25 test set with diverse supervision levels and settings. 100%, 10%, and 1pt denotes the methods trained with 100%, 10%, and single labeled points per object category, respectively. Cloud indicates the methods trained with cloud-level labels.

Setting	Model	Label	CatAvg	Bag	Bin	Box	Cabinet	Chair	Desk	Display	Door	Shelf	Table	Bed	Pillow	Sink	Sofa	Toilet
Full Sup.	BGA-DGC [6]	100%	0.754	0.75	0.80	0.75	0.73	0.83	0.72	0.76	0.81	0.59	0.76	0.81	0.74	0.65	0.79	0.80
	BGA-PN++ [4]		0.774	0.71	0.80	0.75	0.73	0.83	0.76	0.79	0.79	0.63	0.79	0.84	0.75	0.70	0.82	0.86
Weak Sup.	Xu *et al*. [9]	10%	0.554	0.68	0.36	0.50	0.72	0.33	0.61	0.75	0.34	0.38	0.58	0.49	0.46	0.63	0.52	0.52
	Xu *et al*. [9]	1pt	0.482	0.45	0.55	0.33	0.48	0.64	0.34	0.56	0.69	0.36	0.40	0.52	0.53	0.40	0.54	0.38
	Xu *et al*. [9]	Cloud	0.207	0.22	0.33	0.10	0.12	0.01	0.21	0.50	0.24	0.11	0.30	0.17	0.34	0.25	0.07	0.14
	MPRM [7]	Cloud	0.431	0.36	0.52	0.25	0.43	0.54	0.26	0.49	0.64	0.42	0.30	0.41	0.43	0.39	0.50	0.45
Unsup.	K-means	-	0.331	0.37	0.35	0.35	0.30	0.35	0.27	0.33	0.38	0.32	0.23	0.34	0.31	0.31	0.33	0.39
	AdaCoSeg [10]	-	0.313	0.45	0.28	0.18	0.33	0.42	0.26	0.29	0.32	0.42	0.27	0.26	0.23	0.37	0.31	0.32
	Ours	-	0.510	0.56	0.59	0.55	0.42	0.54	0.39	0.53	0.65	0.42	0.36	0.51	0.49	0.47	0.51	0.57

Table 4: Segmentation results (mIoU) on ScanObjectNN PB_T_25_R test set with diverse supervision levels and settings. 100%, 10%, and 1pt denotes the methods trained with 100%, 10%, and single labeled points per object category, respectively. Cloud indicates the methods trained with cloud-level labels.

Setting	Model	Label	CatAvg	Bag	Bin	Box	Cabinet	Chair	Desk	Display	Door	Shelf	Table	Bed	Pillow	Sink	Sofa	Toilet
Full Sup.	BGA-DGC [6]	100%	0.741	0.71	0.79	0.72	0.72	0.82	0.72	0.74	0.80	0.59	0.76	0.76	0.70	0.66	0.80	0.76
	BGA-PN++ [4]		0.773	0.71	0.80	0.74	0.74	0.82	0.79	0.79	0.80	0.63	0.78	0.84	0.76	0.70	0.82	0.85
Weak Sup.	Xu *et al*. [9]	10%	0.534	0.57	0.66	0.45	0.53	0.73	0.38	0.64	0.75	0.36	0.44	0.61	0.57	0.50	0.65	0.50
	Xu *et al*. [9]	1pt	0.503	0.43	0.56	0.35	0.47	0.64	0.36	0.60	0.70	0.34	0.41	0.56	0.59	0.41	0.55	0.46
	Xu *et al*. [9]	Cloud	0.192	0.26	0.32	0.11	0.11	0.01	0.20	0.34	0.22	0.12	0.27	0.18	0.36	0.21	0.06	0.12
	MPRM [7]	Cloud	0.429	0.38	0.55	0.23	0.42	0.52	0.29	0.47	0.64	0.42	0.30	0.41	0.42	0.42	0.50	0.40
Unsup.	K-means	-	0.326	0.34	0.36	0.35	0.29	0.34	0.26	0.34	0.40	0.30	0.23	0.32	0.30	0.30	0.33	0.35
	AdaCoSeg [10]	-	0.348	0.47	0.34	0.41	0.27	0.53	0.21	0.40	0.31	0.24	0.36	0.37	0.36	0.35	0.33	0.28
	Ours	-	0.484	0.54	0.58	0.51	0.40	0.48	0.38	0.49	0.62	0.41	0.37	0.50	0.44	0.46	0.47	0.53

Table 5: Segmentation results (mIoU) on ScanObjectNN PB_T50_R test set with diverse supervision levels and settings. 100%, 10%, and 1pt denotes the methods trained with 100%, 10%, and single labeled points per object category, respectively. Cloud indicates the methods trained with cloud-level labels.

Setting	Model	Label	CatAvg	Bag	Bin	Box	Cabinet	Chair	Desk	Display	Door	Shelf	Table	Bed	Pillow	Sink	Sofa	Toilet
Full Sup.	BGA-DGC [6]	100%	0.756	0.75	0.81	0.74	0.73	0.82	0.73	0.74	0.82	0.59	0.77	0.79	0.73	0.66	0.79	0.80
	BGA-PN++ [4]		0.764	0.69	0.80	0.73	0.73	0.81	0.78	0.77	0.77	0.63	0.78	0.84	0.74	0.71	0.82	0.83
Weak Sup.	Xu *et al*. [9]	10%	0.500	0.50	0.68	0.36	0.50	0.72	0.33	0.61	0.75	0.34	0.38	0.58	0.49	0.46	0.63	0.52
	Xu *et al*. [9]	1pt	0.471	0.41	0.56	0.39	0.47	0.61	0.31	0.56	0.67	0.31	0.35	0.54	0.40	0.43	0.56	0.41
	Xu *et al*. [9]	Cloud	0.170	0.20	0.21	0.10	0.11	0.01	0.20	0.34	0.16	0.09	0.20	0.16	0.34	0.22	0.06	0.14
	MPRM [7]	Cloud	0.407	0.39	0.52	0.22	0.41	0.50	0.27	0.42	0.62	0.39	0.29	0.39	0.37	0.39	0.47	0.39
Unsup.	K-means	-	0.317	0.32	0.34	0.32	0.31	0.34	0.28	0.32	0.37	0.29	0.25	0.28	0.30	0.29	0.33	0.33
	AdaCoSeg [10]	-	0.330	0.31	0.35	0.31	0.44	0.27	0.25	0.15	0.47	0.28	0.26	0.37	0.38	0.34	0.41	0.36
	Ours	-	0.459	0.48	0.55	0.50	0.38	0.49	0.35	0.46	0.59	0.39	0.36	0.46	0.43	0.43	0.46	0.48

Table 6: Segmentation results (mIoU) on ScanObjectNN PB_T_50_RS test set with diverse supervision levels and settings. 100%, 10%, and 1pt denotes the methods trained with 100%, 10%, and single labeled points per object category, respectively. Cloud indicates the methods trained with cloud-level labels.

Setting	Model	Label	CatAvg	Bag	Bin	Box	Cabinet	Chair	Desk	Display	Door	Shelf	Table	Bed	Pillow	Sink	Sofa	Toilet
Full Sup.	BGA-DGC [6]	100%	0.753	0.76	0.82	0.74	0.72	0.83	0.73	0.72	0.81	0.57	0.74	0.80	0.75	0.67	0.79	0.79
	BGA-PN++ [4]		0.761	0.72	0.79	0.75	0.75	0.78	0.76	0.78	0.80	0.61	0.76	0.84	0.75	0.71	0.81	0.85
Weak Sup.	Xu *et al.* [9]	10%	0.523	0.50	0.61	0.40	0.48	0.70	0.35	0.58	0.70	0.35	0.39	0.57	0.52	0.47	0.60	0.56
	Xu *et al.* [9]	1pt	0.463	0.40	0.56	0.30	0.45	0.60	0.29	0.56	0.68	0.33	0.35	0.54	0.43	0.42	0.54	0.42
	Xu *et al.* [9]	Cloud	0.130	0.15	0.23	0.07	0.08	0.01	0.12	0.22	0.14	0.06	0.19	0.14	0.20	0.19	0.05	0.09
	MPRM [7]	Cloud	0.398	0.36	0.51	0.21	0.39	0.49	0.25	0.41	0.55	0.39	0.28	0.39	0.36	0.49	0.39	0.40
Unsup.	K-means	-	0.340	0.37	0.36	0.38	0.32	0.37	0.29	0.36	0.40	0.32	0.26	0.30	0.30	0.34	0.35	0.33
	AdaCoSeg [10]	-	0.340	0.35	0.37	0.42	0.42	0.33	0.28	0.31	0.28	0.27	0.46	0.21	0.32	0.39	0.33	0.37
	Ours	-	0.484	0.55	0.56	0.54	0.39	0.43	0.39	0.51	0.63	0.41	0.38	0.52	0.44	0.47	0.47	0.52

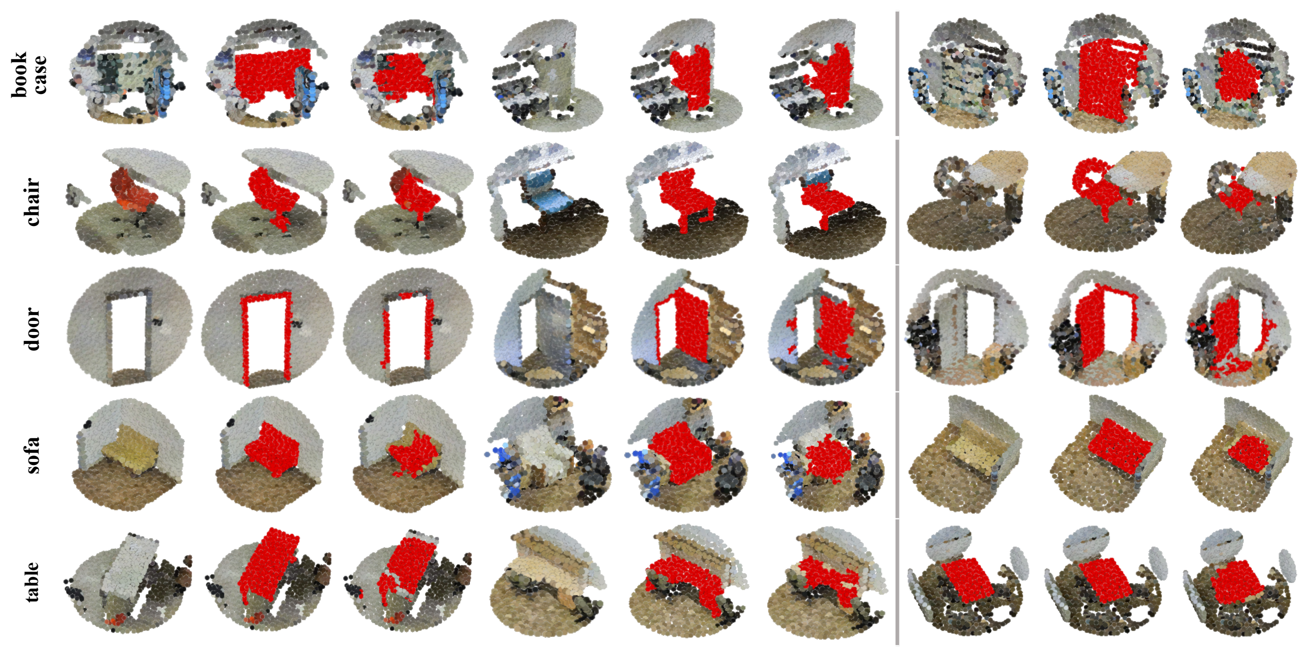

Figure 2: Object co-segmentation on the object categories of the S3DIS dataset. Each category is associated with the same row. For each example, we show the input cloud, the ground-truth label, and our segmentation result. The last example of each row (on the right of the gray line) shows a less successful case.

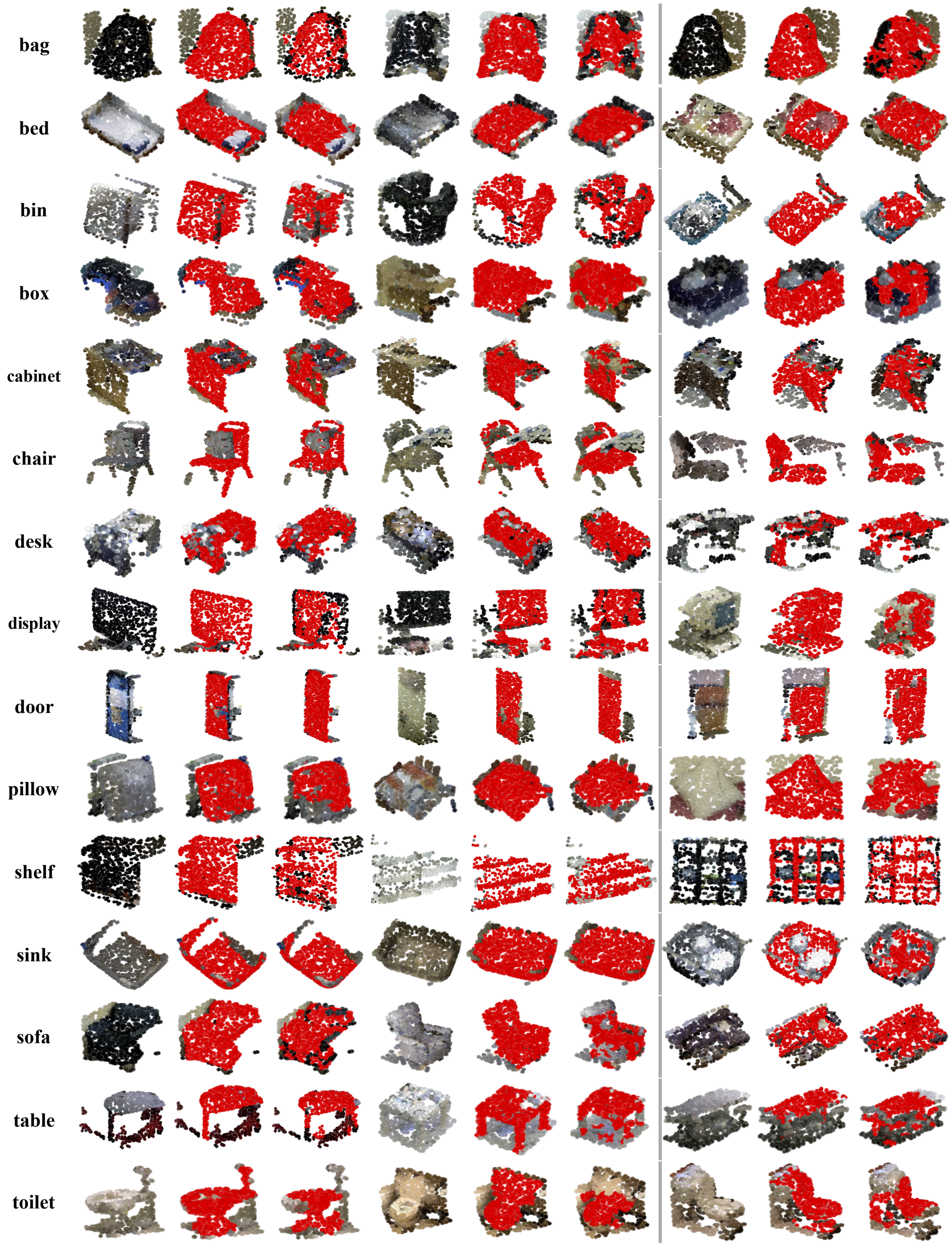

Figure 3: Object co-segmentation on the object categories of the ScanObjectNN dataset. Each category is associated with the same row. For each example, we show the input cloud, the ground-truth label, and our segmentation result. The last example of each row (on the right of the gray line) shows a less successful case.