CVPR #11145

Mitigating Object Dependencies: Improving Point Cloud Self-Supervised Learning through Object Exchange

Supplementary Material

A. The relative weight of the auxiliary task loss.

γ is the relative weight of the auxiliary task loss in Eq.6 in the main paper. To study the impact of it, we gradually increase the relative weight γ. As shown in Fig. 1, with the increase of γ, the performance first increase and then decrease.

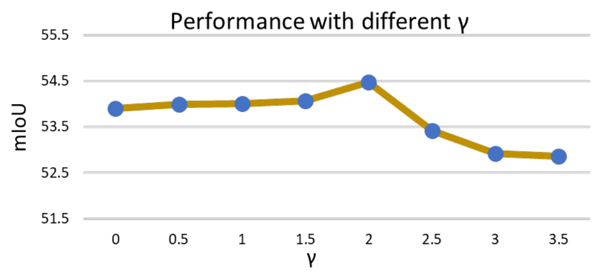

Figure 1. mIoU comparison under pre-training models with different γ. All the models are pre-trained and fine-tuned on ScanNet

B. Detailed ScanNet-C.

In Section 4.3 of the main paper, to evaluate the performance of models in changing contexts, we create a new dataset, ScanNet-C, by replacing a proportion δ of the objects in ScanNet.

Specifically, for each point cloud P^m with N_m objects in ScanNet, we randomly select a point cloud P^n with N_n from the entire dataset. And then δN_m objects in P^m are replaced with objects sharing comparable size from P^n using the object-exchanging strategy mentioned in the main paper. We replace objects in each point cloud in ScanNet and range δ from 0.1 to 0.9 in the experiments. In Fig. 2, we visualize the scenes in ScanNet and the corresponding scenes in ScanNet-C. As shown in the figure, the inter-object correlations are changed, for example, a bed is replaced with a chair on the left of Fig. 2. In Table. 2, we show each individual run on ScanNet-C semantic segmentation with varied proportions δ. As the table shows, our OESSL outperforms all other methods under all δ.

C. Detailed results and visualization.

The number of training epochs for every label regime can be found in Table 1. For completeness, we report in Table. 3 and Table. 4 the mIoU of each of the three individual runs performed to obtain the main results in the paper. As the table shows, our method performs better than other methods consistently.

Label regime	10%	20%	50%	100%
ScanNet [2]	250	250	100	75
S3DIS [1]	400	300	200	200
Label regime	0.1%	1%	10%	100%
Synthia4D [4]	250	200	25	20

Table 1. Number of training epochs used for different label regimes on different datasets.

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, 3

[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

[3] Siyuan Huang, Yichen Xie, Song-Chun Zhu, and Yixin Zhu. Spatio-temporal self-supervised representation learning for 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6535–6545, 2021. 3

[4] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 1, 3

[5] Xiaoyang Wu, Xin Wen, Xihui Liu, and Hengshuang Zhao. Masked scene contrast: A scalable framework for unsupervised 3d representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9415–9424, 2023. 2, 3

[6] Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3d features on any point-cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10252–10263, 2021. 2, 3

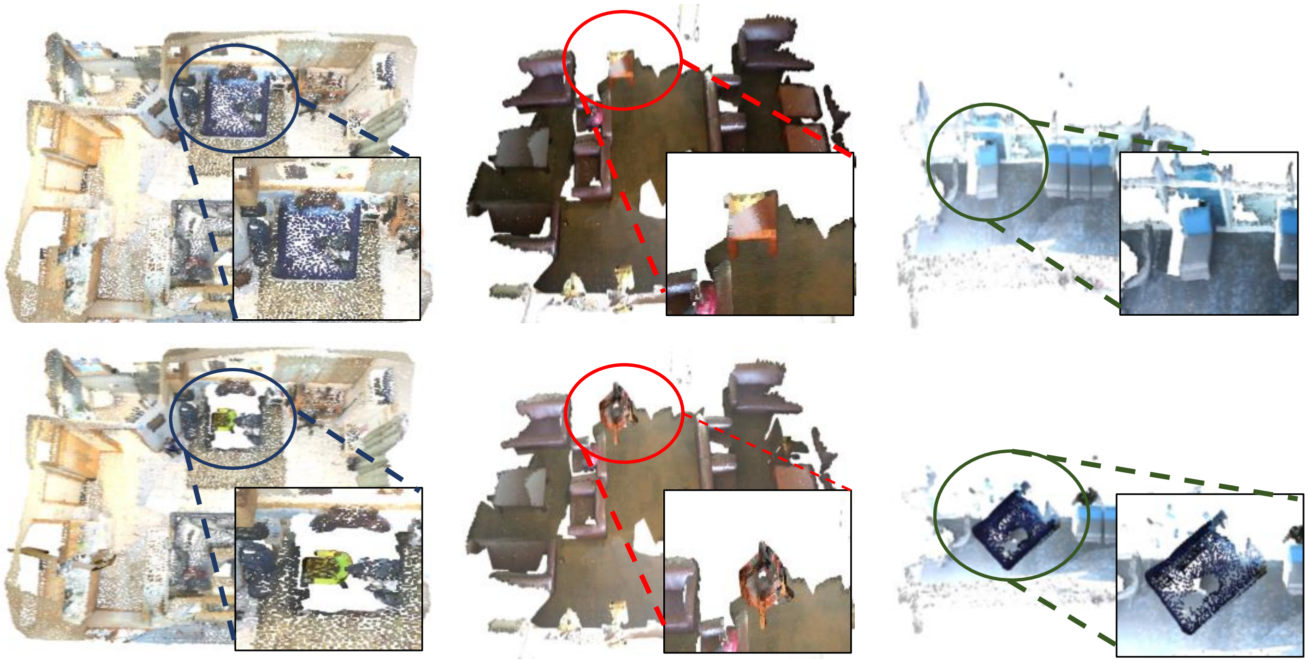

Figure 2. **Top**: Visualization of scenes in ScanNet. **Bottom**: Visualization of corresponding scenes in ScanNet-C

Method		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
From Scratch		51.73	46.51	40.66	37.82	34.09	30.79	30.43	27.60	26.38	26.29
	Runs	51.73	46.15	40.92	36.52	33.65	30.97	29.30	28.28	26.37	24.83
		51.73	46.22	42.21	35.81	33.46	30.64	30.01	29.21	26.39	25.51
	Average	51.73	46.29	41.26	36.72	33.73	30.80	29.91	28.36	26.38	25.55
DepthContrast [6]		51.36	45.59	39.58	37.65	33.27	30.55	30.15	27.47	26.77	25.63
	Runs	51.36	45.67	40.15	36.59	33.18	30.28	28.80	27.95	26.46	25.14
		51.36	45.15	41.84	34.90	33.02	30.71	29.61	28.76	26.71	25.48
	Average	51.36	45.47	40.52	36.38	33.15	30.51	29.52	28.06	26.65	25.42
MSC [5]		55.50	49.85	43.28	41.72	37.56	34.25	33.67	30.85	29.87	28.82
	Runs	55.50	49.68	43.95	40.74	36.86	33.60	32.44	31.19	29.20	27.98
		55.50	49.49	45.48	39.07	37.22	34.10	33.17	32.40	29.05	28.70
	Average	55.50	49.67	44.24	40.51	37.21	33.98	33.09	31.48	29.37	28.50
OESSL(Ours)		56.72	51.54	44.98	42.95	38.30	35.82	35.46	32.10	31.32	29.86
	Runs	56.72	50.77	45.49	41.87	38.41	35.10	33.48	32.79	30.32	29.52
		56.72	51.13	47.34	40.89	38.58	35.55	34.29	33.52	30.97	30.01
	Average	56.72	51.15	45.94	41.90	38.43	35.49	34.41	32.80	30.87	29.80

Table 2. Detailed of individual runs on **ScanNet-C** semantic segmentation with different proportions δ of replaced objects. We report mIoU% for each of the individual runs averaged in the main paper.

		ScanNet [2] Validation				S3DIS [1] Area5			
%	Method	Split 1	Split 2	Split 3	Average	Split 1	Split 2	Split 3	Average
10%	From Scratch	51.73	46.12	49.12	48.99	35.32	41.86	44.27	40.48
	DepthContrast [6]	51.36	49.93	49.6	50.30	45.10	47.84	46.76	46.57
	STRL [3]	50.29	48.00	42.52	46.94	31.21	37.42	42.33	36.99
	MSC [5]	55.5	52.71	53.34	53.85	43.61	48.46	42.48	44.85
	OESSL(Ours)	**56.72**	**52.97**	**53.43**	**54.37**	**46.71**	**49.88**	**51.07**	**49.22**
20%	From Scratch	55.22	57.78	59.73	57.58	43.02	49.92	44.88	45.94
	DepthContrast [6]	55.81	57.59	57.83	57.08	46.55	48.52	47.95	47.67
	STRL [3]	57.85	59.01	59.97	58.94	44.48	49.6	44.44	46.13
	MSC [5]	59.67	59.85	61.88	60.47	46.17	52.4	51.8	50.12
	OESSL(Ours)	**60.33**	**60.58**	**62.91**	**61.27**	**49.75**	**55.53**	**52.72**	**52.67**
50%	From Scratch	62.38	61.51	61.22	61.70	51.27	53.51	54.97	53.25
	DepthContrast [6]	61.66	61.89	60.87	61.47	52.86	53.55	55.14	53.85
	STRL [3]	61.78	62.38	61.38	61.85	54.19	55.56	55.58	55.11
	MSC [5]	**63.92**	64.66	63.36	63.98	56.56	56.48	58.43	57.16
	OESSL(Ours)	63.67	**65.46**	**64.54**	**64.56**	**60.98**	**61.95**	**62.43**	**61.79**
100%	From Scratch	71.40	70.98	70.94	71.11	65.54	66.18	66.75	66.16
	DepthContrast [6]	70.78	71.00	70.98	70.92	63.68	61.18	65.41	63.42
	STRL [3]	70.38	**71.56**	71.15	71.03	66.13	65.92	62.08	64.71
	MSC [5]	**71.52**	70.84	70.64	71.00	65.83	63.55	**66.83**	65.40
	OESSL(Ours)	71.29	71.24	**71.32**	**71.28**	**67.55**	**67.49**	65.65	**66.90**

Table 3. Details of individual runs on **ScanNet** and **S3DIS** semantic segmentation. Each run corresponds to fine-tuning using a different regime. We report mIoU% for each of the individual runs averaged in the main paper

		Synthia4D [4] Test				Synthia4D [4] Validation			
%	Method	Split 1	Split 2	Split 3	Average	Split 1	Split 2	Split 3	Average
0.1%	From Scratch	16.81	21.92	20.79	19.84	17.66	21.57	21.28	20.17
	DepthContrast [6]	48.87	44.69	44.78	46.11	46.20	46.55	45.93	46.23
	STRL [3]	46.34	32.92	39.65	39.64	43.67	41.37	29.77	38.27
	MSC [5]	49.51	45.58	46.24	47.11	45.39	46.31	47.55	46.42
	OESSL(Ours)	**52.56**	**48.13**	**49.62**	**49.44**	**50.82**	**49.11**	**48.04**	**49.32**
1%	From Scratch	63.38	62.80	63.92	63.37	67.74	67.77	67.92	67.81
	DepthContrast [6]	66.60	67.17	64.97	66.25	71.14	71.57	72.27	71.66
	STRL [3]	67.67	64.88	64.23	65.59	71.63	71.26	68.59	70.49
	MSC [5]	67.08	65.23	66.95	66.42	72.93	71.83	69.98	71.58
	OESSL(Ours)	**68.26**	**70.83**	**67.16**	**68.75**	**73.88**	**74.66**	**73.98**	**74.17**
10%	From Scratch	71.84	68.75	70.76	70.45	75.22	73.17	74.66	74.35
	DepthContrast [6]	69.31	70.82	71.33	70.49	73.04	74.65	74.31	74.00
	STRL [3]	67.32	70.78	70.26	69.45	75.54	72.92	72.95	73.80
	MSC [5]	**72.64**	73.50	73.30	73.15	75.52	74.96	76.10	75.53
	OESSL(Ours)	71.40	**73.73**	**75.12**	**73.42**	**76.60**	**77.16**	**77.37**	**77.04**
100%	From Scratch	**77.57**	77.06	76.37	77.00	80.71	80.74	80.06	80.50
	DepthContrast [6]	76.72	75.34	73.56	75.21	76.88	79.44	79.36	78.56
	STRL [3]	77.34	76.53	78.11	77.33	81.28	**81.66**	79.92	80.95
	MSC [5]	76.80	77.75	77.11	77.25	80.84	80.78	**81.52**	81.05
	OESSL(Ours)	76.05	**78.10**	**78.29**	**77.48**	**81.41**	81.20	81.32	**81.31**

Table 4. Details of individual runs on **Synthia4D** semantic segmentation. Each run corresponds to fine-tuning using a different regime. We report mIoU% for each of the individual runs averaged in the main paper.