Supplementary Material for "MV-JAR: Masked Voxel Jigsaw and Reconstruction for LiDAR-Based Self-Supervised Pre-Training"

Runsen Xu[1,2] Tai Wang[1,2] Wenwei Zhang[3,2] Runjian Chen[4] Jinkun Cao[5]
Jiangmiao Pang[2✉] Dahua Lin[1,2]
[1]The Chinese University of Hong Kong [2]Shanghai AI Laboratory [3]S-Lab, NTU
[4]The University of Hong Kong [5]Carnegie Mellon University
{runsenxu,wt019,dhlin}@ie.cuhk.edu.hk, wenwei001@ntu.edu.sg, rjchen@connect.hku.hk,
jinkunc@andrew.cmu.edu, pangjiangmiao@gmail.com

1. Necessity of Our New Benchmark

To demonstrate the importance of full convergence and our proposed benchmark, we follow previous works [3, 6] and fine-tune SST [1] on uniformly sampled 5% data. We conduct fine-tuning for 6 epochs and 84 epochs, respectively. The results in Tab. 1 reveal that ProposalContrast significantly improves the baseline when the model is trained with few iterations. However, it slightly diminishes performance when the model is fully converged.

The disappearance of pre-training benefits underscores the necessity of adequate fine-tuning. Therefore, fine-tuning the model on different uniformly sampled splits using the same number of *epochs*, which may result in incomplete convergence on smaller splits, cannot precisely assess pre-training effects. In our main paper, we present evidence that uniformly sampled splits are actually similar when the model reaches full convergence. Our proposed benchmark, which samples data by scene sequences to create diverse splits, can effectively and comprehensively reveal the pure improvements of pre-training.

Table 1. Fine-tuning on uniformly sampled 5% data.

Initialization	6 Epochs		84 Epochs	
	L2 mAP	L2 mAPH	L2 mAP	L2 mAPH
Random	40.90	34.12	63.00	58.86
ProposalContrast [6]	47.56	41.30	62.57	58.75
MV-JAR	**50.67**	**45.02**	**65.14**	**61.74**

2. More Results on Waymo Subsets

To reduce performance variance on our 5% and 10% splits of the Waymo [4] dataset, we randomly sample each split three times to form three subsets. In the main paper, we report the detection results of SST fine-tuned with Subset 0. In this supplementary material, we present the detection results of SST fine-tuned with the other two subsets in Tab. 4 and Tab. 5. Additionally, we report the average results across all three subsets in Tab. 6.

3. Experiments with Convolution-based Detectors

Implementation details. Our MVJ and MVR directly mask the raw inputs of LiDAR point clouds, making them suitable for most 3D detectors that downsample the point clouds into voxels and extract voxel features for perception. However, MVJ aims at predicting the voxel positions, which is necessary for convolutional operations. Directly applying MVJ to convolution-based backbones may result in information leakage and trivial pre-training. This is not an issue with Transformer-based backbones, as the attention mechanism does not require position information to perform, and we do not add positional embeddings during pre-training. On the other hand, MVR predicts voxel shapes while retaining position information, making it compatible with convolution-based detectors without modification.

To overcome the limitation of MVJ when applied to convolution-based detectors, we permutate the masked voxels by randomly placing them in the partitioned window before feeding them to the convolutional backbones. This permutation hides the original position information of the masked voxels, avoiding information leakage and making MVJ pre-training meaningful.

Experimental Results. To evaluate the performance of our proposed methods on convolution-based detectors, we pre-train PointPillar [2] and CenterPoint (Pillar) [7] with MVR and MVJ as SST. We fine-tune these models on our 5% split and report their overall L2 performances in Tab. 2. Our experimental results demonstrate that both MVJ and MVR can work effectively for convolution-based detectors,

✉ Corresponding author.

showcasing the generalization abilities of our methods.

Table 2. Performances with convolution-based detectors.

Initialization	PointPillar [2]		CenterPoint [7]	
	L2 mAP	L2 mAPH	L2 mAP	L2 mAPH
Random	41.27	35.10	36.21	32.26
MVR	43.82$^{+2.55}$	37.97$^{+2.87}$	38.02$^{+1.81}$	33.73$^{+1.47}$
MVJ	43.01$^{+1.74}$	37.11$^{+2.01}$	38.61$^{+2.40}$	34.48$^{+2.22}$

4. Effects of Pre-training on Varying Distances

In order to investigate the influence of pre-training across diverse distances, we present the L2 mAPH performance using 5% fine-tuning data for different distance intervals in Tab. 3. It can be observed that the performance enhancement of MVR declines as the distances increase, primarily boosting MVJ (i.e., MV-JAR) within the 0m-30m range. A plausible explanation for this phenomenon is the reduction in point density at greater distances, which makes the dense point clusters at closer ranges less ambiguous for the model to reconstruct voxel shapes. MVJ consistently outperforms MVR across various distance intervals, reinforcing our hypothesis that capturing voxel distributions plays a more crucial role in the model's representation learning. This is because LiDAR detectors downsample points into voxels to facilitate perception.

Table 3. Overall L2 mAPH across various distances.

Initialization	Overall		
	0m-30m	30m-50m	50m-inf
Random	60.15	34.61	18.18
MVR	62.77 $^{+2.62}$	36.56$^{+1.95}$	19.94$^{+1.76}$
MVJ	65.66$^{+5.51}$	40.61$^{+6.00}$	23.07$^{+4.89}$
MV-JAR	66.95$^{+6.80}$	40.62$^{+6.01}$	22.88$^{+4.70}$

References

[1] Lue Fan, Ziqi Pang, Tianyuan Zhang, Yu-Xiong Wang, Hang Zhao, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. Embracing single stride 3d object detector with sparse transformer. In *CVPR*, 2022. 1

[2] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, 2019. 1, 2

[3] Hanxue Liang, Chenhan Jiang, Dapeng Feng, Xin Chen, Hang Xu, Xiaodan Liang, Wei Zhang, Zhenguo Li, and Luc Van Gool. Exploring geometry-aware contrast and clustering harmonization for self-supervised 3d object detection. In *CVPR*, 2021. 1

[4] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In *CVPR*, 2020. 1

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

[6] Junbo Yin, Dingfu Zhou, Liangjun Zhang, Jin Fang, Cheng-Zhong Xu, Jianbing Shen, and Wenguan Wang. Proposalcontrast: Unsupervised pre-training for lidar-based 3d object detection. In *ECCV*, 2022. 1, 3

[7] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *CVPR*, 2021. 1, 2

Table 4. Data-efficient 3D object detection results of SST on the Waymo validation set, fine-tuned with Subset 1.

Data amount	Initialization	Overall		Car		Pedestrian		Cyclist	
		L2 mAP	L2 mAPH	L2 mAP	L2 mAPH	L2 mAP	L2 mAPH	L2 mAP	L2 mAPH
5%	Random	47.74	43.69	50.24	49.75	51.68	41.26	41.29	40.07
	PointContrast [5]	48.97	44.91	52.35	51.85	52.49	41.95	42.07	40.91
	ProposalContrast [6]	49.87	45.83	52.79	52.31	53.30	43.00	43.51	42.18
	MV-JAR (Ours)	**52.73**$^{+4.99}$	**48.99**$^{+5.30}$	56.66	56.21	57.52	47.61	44.02	43.15
10%	Random	55.95	52.15	55.23	54.76	60.61	50.86	52.01	50.84
	PointContrast [5]	55.22	51.31	55.62	55.15	59.25	49.17	50.81	49.60
	ProposalContrast [6]	55.59	51.67	55.57	55.12	60.02	49.98	51.18	49.90
	MV-JAR (Ours)	**58.61**$^{+2.66}$	**55.12**$^{+2.97}$	58.92	58.49	63.44	54.40	53.48	52.47

Table 5. Data-efficient 3D object detection results of SST on the Waymo validation set, fine-tuned with Subset 2.

Data amount	Initialization	Overall		Car		Pedestrian		Cyclist	
		L2 mAP	L2 mAPH	L2 mAP	L2 mAPH	L2 mAP	L2 mAPH	L2 mAP	L2 mAPH
5%	Random	42.59	38.83	50.09	49.59	53.88	44.06	23.79	22.85
	PointContrast [5]	44.48	40.55	51.87	51.37	55.36	45.03	26.22	25.24
	ProposalContrast [6]	45.21	41.45	52.29	51.82	56.23	46.28	27.10	26.24
	MV-JAR (Ours)	**47.93**$^{+5.34}$	**44.50**$^{+5.67}$	56.22	55.78	58.80	49.77	28.75	27.95
10%	Random	54.85	51.22	54.95	54.51	62.11	52.76	47.49	46.40
	PointContrast [5]	54.80	51.02	55.41	54.95	60.56	50.86	48.44	47.24
	ProposalContrast [6]	54.77	51.09	55.64	55.20	60.54	51.16	48.14	46.92
	MV-JAR (Ours)	**58.29**$^{+3.44}$	**54.99**$^{+3.77}$	59.17	58.74	64.58	56.02	51.12	50.20

Table 6. Average results on the Waymo validation set, averaged across SST fine-tuned with Subset 0-2.

Data amount	Initialization	Overall		Car		Pedestrian		Cyclist	
		L2 mAP	L2 mAPH	L2 mAP	L2 mAPH	L2 mAP	L2 mAPH	L2 mAP	L2 mAPH
5%	Random	44.91	40.96	50.45	49.94	52.77	42.53	31.52	30.40
	PointContrast [5]	46.26	42.25	52.11	51.61	53.84	43.40	32.82	31.74
	ProposalContrast [6]	47.23	43.28	52.58	52.10	54.61	44.37	34.50	33.38
	MV-JAR (Ours)	**50.39**$^{+5.48}$	**46.72**$^{+5.76}$	56.45	56.00	57.99	48.36	36.74	35.81
10%	Random	55.04	51.28	55.01	54.55	61.09	51.44	49.02	47.84
	PointContrast [5]	54.57	50.75	55.26	54.80	59.85	50.05	48.61	47.41
	ProposalContrast [6]	54.75	50.96	55.47	55.01	60.19	50.51	48.60	47.37
	MV-JAR (Ours)	**58.12**$^{+3.08}$	**54.72**$^{+3.44}$	58.84	58.41	63.77	55.03	51.74	50.73