

Supplementary Materials: Point Cloud Reconstruction Is Insufficient to Learn 3D Representations

Anonymous Author(s)

A PRE-TRAINING SETTINGS

We pre-train on Waymo and report the results of 3D object detection on Waymo val and test set. We pre-train on nuScenes and report the results of 3D object detection, 3D semantic segmentation, and occupancy prediction on nuScenes val set. Tab. A.1 and Tab. A.2 provide the pre-training settings on Waymo and nuScenes.

Table A.1: Pre-training settings on Waymo.

Parameter	Value
Point cloud range	[-74.88, -74.88, -2, 74.88, 74.88, 4.0]
Voxel size	[0.32, 0.32, 0.1875]
Voxel grid shape	[468, 468, 32]
Augmentor	×
Sparse point encoder	MinkUNet (Res16UNet34C)
Feature dimension of point encoder	64
Feature dimension of voxel encoder	192
Number of DSVT encoder	8
Number of DSVT decoder	4
Seal feature loss	SmoothL1Loss
Optimizer	AdamW
Weight decay	0.05
Epochs	30

Table A.2: Pre-training settings on nuScenes.

Parameter	Value
Point cloud range	[-51.2, -51.2, -5, 51.2, 51.2, 3]
Voxel size	[0.2, 0.2, 0.2]
Voxel grid shape	[512, 512, 40]
Augmentor	×
Sparse point encoder	MinkUNet (Res16UNet34C)
Feature dimension of point encoder	64
Feature dimension of voxel encoder	256
Number of DSVT/SST encoder	8
Number of DSVT decoder	4
Seal feature loss	SmoothL1Loss
Optimizer	AdamW
Epochs	72

B FINE-TUNING SETTINGS

B.1 3D Object Detection

In Tab. B.1, we present the fine-tuning settings for DSVT on Waymo. We conduct fine-tuning for only 12 epochs. For detailed settings, please refer to DSVT [7].

B.2 3D Semantic Segmentation

Tab. B.2 provides the fine-tuning settings for Cylinder3D in 3D semantic segmentation. We replace Asymm3DSconv with SST to extract voxel features. Subsequently, we use Cylinder3DHead to

Table B.1: Fine-tuning settings for 3D object detection on Waymo.

Parameter	Value
Point cloud range	[-74.88, -74.88, -2, 74.88, 74.88, 4.0]
Voxel size	[0.32, 0.32, 0.1875]
Voxel grid shape	[468, 468, 32]
Augmentor	gt sampling, flip, rotation, scaling, translation
Disable augmentation epoch	1
Number of frame	1
Feature dimension of voxel encoder	192
Number of DSVT encoder	8
Detection head	CenterHead
Optimizer	AdamW
Epochs	12
Grad norm clip	10
Max obj per sample	500

Table B.2: Fine-tuning settings for 3D semantic segmentation on nuScenes.

Parameter	Value
Point cloud range	[0, -3.14159265359, -4, 50, 3.14159265359, 2]
Voxel grid shape	[480, 360, 32]
With voxel center	True
Augmentor	flip, rotation, scaling, translation
Feature dimension of voxel encoder	128
Number of SST encoder	8
Segmentation head	Cylinder3DHead
Optimizer	AdamW
Epochs	24

predict the semantic categories of point clouds. For more detailed settings, please refer to Cylinder3D [10].

B.3 3D Occupancy Prediction

Table B.3: Fine-tuning settings for occupancy prediction on nuScenes (OpenOccupancy).

Parameter	Value
Point cloud range	[-51.2, -51.2, -5, 51.2, 51.2, 3]
Voxel size	[0.2, 0.2, 0.2]
Voxel grid shape	[512, 512, 40]
Augmentor	flip, rotation, scaling, translation
Number of frame	10
Feature dimension of voxel encoder	256
Number of DSVT encoder	8
Encoder neck	FPN3D
Num of level	4
Optimizer	AdamW
Epochs	24
Grad norm clip	35

In Tab. B.3, we provide the fine-tuning settings for occupancy prediction when using LiDAR only. We utilize DSVT as the encoder

Table C.1: Correspondence between the eight semantic classes and the categories in the downstream datasets SemanticKITTI, nuScenes Lidarseg, and WOD Semantic Segmentation.

Num.	Class name	Downstream dataset categories
1	Ground-related	SemanticKITTI: road, sidewalk, parking, other-ground nuScenes Lidarseg: driveable_surface, sidewalk, other_flat WOD Semantic Segmentation: curb, road, lane maker, other ground, walkable, sidewalk
2	Structures	SemanticKITTI: building, other-structure nuScenes Lidarseg: construction, mannade WOD Semantic Segmentation: building
3	Vehicle	SemanticKITTI: car, truck, other-vehicle nuScenes Lidarseg: bus, car, trailer, truck WOD Semantic Segmentation: car, truck, bus, other vehicle
4	Two-wheeled vehicle	SemanticKITTI: bicycle, motorcycle nuScenes Lidarseg: bicycle, motorcycle WOD Semantic Segmentation: bicycle, motor cycle
5	Nature	SemanticKITTI: vegetation, trunk, terrain nuScenes Lidarseg: terrain, vegetation WOD Semantic Segmentation: vegetation, treetrunk
6	Human	SemanticKITTI: person, bicyclist, motorcyclist nuScenes Lidarseg: pedestrian WOD Semantic Segmentation: motor cyclist, bicyclist, pedestrian
7	Object	SemanticKITTI: fence, pole, traffic sign, other-object nuScenes Lidarseg: barrier, trafficcone WOD Semantic Segmentation: sign, traffic light, pole, cone
8	Outlier	SemanticKITTI: outlier nuScenes Lidarseg: noise WOD Semantic Segmentation: —

with pre-trained initialization. FPN3D serves as the neck for incorporating multi-scale information from the encoder. The features are ultimately input into OccHead for occupancy prediction.

C METHODOLOGY FOR SEMANTIC CLASSIFICATION

To determine the number of superclasses for Seal features, we need to identify semantic categories that can represent autonomous driving scenes. Tab. C.1 lists the correspondence between 8 semantic categories and categories on downstream datasets SemanticKITTI [1], nuScenes Lidarseg [3], and WOD Semantic Segmentation [5]. Based on this, the number of superclasses is set to 8.

D MORE RESULTS

D.1 Ablation Study of Epochs for Pre-training

We conduct pre-training on the Waymo training set for 10, 30, and 60 epochs, followed by fine-tuning for 30 epochs on the 3D object detection. Tab. D.1 reports an ablation study of epochs for pre-training. It can be observed that as the number of pre-training epochs increases, the L2 mAP/mAPH gradually rises, eventually reaching 75.44% and 72.93%. This demonstrates the scalability of our method. For fair comparison and to reduce computational cost, we report the results of pre-training for 30 epochs in all results.

Table D.1: Ablation study of epochs for pre-training.

Epochs	L2 (AP/APH)↑			
	Overall	Vehicle	Pedestrian	Cyclist
0	73.20 / 71.00	70.90 / 70.50	75.20 / 69.80	73.60 / 72.70
10	74.40 / 72.22	72.01 / 71.86	76.53 / 71.06	74.68 / 73.74
30	75.13 / 72.69	72.93 / 72.45	77.18 / 71.66	75.27 / 73.96
60	75.44 / 72.93	73.13 / 72.77	77.44 / 71.85	75.74 / 74.18

D.2 Ablation Study of the Base Mask Ratio

Tab. D.2 provides an ablation study of (r_b^1, r_b^2, r_b^3) in inter-class discrimination-guided masking. We set the base mask ratio to be positively correlated with the average inter-class distance. Through the ablation study, (0.9, 0.45, 0) has been determined as a suitable base mask ratio without loss of generality.

D.3 Ablation Study of Expected Number of Superclass Partition

Tab. D.3 investigates the expected number (n_1, n_2, n_3) in Algorithm Fastest Class Sampling. This hyperparameter divides the eight superclasses into three sets and sets the base mask ratio accordingly. Experimental investigations have shown that among the two divisions, (3, 3, 2) yields superior results compared to (4, 2, 2).

Table D.2: Ablation study of the base mask ratio for superclass partition $\mathcal{K}_1, \mathcal{K}_2, \mathcal{K}_3$ in inter-class discrimination-guided masking.

(r_b^1, r_b^2, r_b^3)	L2 (AP/APH) \uparrow			
	Overall	Vehicle	Pedestrian	Cyclist
(1.0, 0.5, 0)	74.34 / 72.02	72.04 / 71.54	76.12 / 71.00	74.87 / 73.52
(0.9, 0.45, 0)	75.13 / 72.69	72.93 / 72.45	77.18 / 71.66	75.27 / 73.96
(0.8, 0.4, 0)	74.78 / 72.45	72.52 / 72.16	76.61 / 71.39	75.22 / 73.81
(0.6, 0.3, 0)	74.58 / 72.22	72.23 / 71.90	76.37 / 71.18	75.13 / 73.57

Table D.3: Ablation study of expected number of superclass partition n_1, n_2, n_3 in inter-class discrimination-guided masking.

(n_1, n_2, n_3)	L2 (AP/APH) \uparrow			
	Overall	Vehicle	Pedestrian	Cyclist
(3, 3, 2)	75.13 / 72.69	72.93 / 72.45	77.18 / 71.66	75.27 / 73.96
(4, 2, 2)	74.80 / 72.45	72.46 / 72.01	76.63 / 71.41	75.31 / 73.94

D.4 Ablation Study of Distance Threshold

Tab. D.4 investigates the distance threshold λ in intra-class discrimination guided masking. This threshold is used to determine whether the Seal feature of a voxel is too far away from its cluster center μ^{k_i} . It affects the intra-class consistency coefficient r_c and, consequently, influences the self-supervised learning process. Through experimentation, 0.6 is selected as the default value.

D.5 Ablation Study of the Number of Encoders

Tab. D.5 configures different numbers of encoders in DSVT. As the number of encoder layers deepens, the pre-training can extract more universal features from unlabeled data. Performance on downstream tasks improves as the number of encoder layers deepens. However, there is no significant improvement in results when going from 8 layers to 10. To enhance efficiency, we opt for 8 layers as the default value.

D.6 Ablation Study of the Type of Decoder

The decoder utilizes the features of unmasked voxels to reconstruct the low-level and high-level features of masked voxels. Tab. D.6 explores three types of decoder: Sparse Convolution [9], SST [2], and DSVT. Sparse Convolution extracts features from the neighborhood of non-empty voxels. SST and DSVT, on the other hand, perform reconstruction by computing attention between mask tokens and unmasked voxels. DSVT outperforms Sparse Convolution and SST by achieving improvements of 0.61% and 0.36% in L2 mAP, respectively. We select DSVT as the default decoder.

D.7 Study of Across Datasets and Joint Datasets

In Tab. D.7, transfer learning across datasets and on joint datasets can reflect the ability of self-supervised learning to acquire universal features from autonomous driving scenarios. This encompasses differences in street views, vehicle appearances, weather conditions, and architectural styles. When pre-trained on the Waymo and fine-tuned on nuScenes, despite a decrease of 0.38% in mAP compared

Table D.4: Ablation study of distance threshold λ in intra-class discrimination-guided masking.

λ	L2 (AP/APH) \uparrow			
	Overall	Vehicle	Pedestrian	Cyclist
0.8	74.59 / 72.22	72.25 / 72.00	76.54 / 71.05	74.98 / 73.61
0.7	74.77 / 72.46	72.53 / 72.22	76.68 / 71.39	75.11 / 73.77
0.6	75.13 / 72.69	72.93 / 72.45	77.18 / 71.66	75.27 / 73.96
0.5	74.56 / 72.22	72.40 / 71.98	76.48 / 71.15	74.79 / 73.54

Table D.5: Ablation study of the number of encoders.

Layer num.	L2 (AP/APH) \uparrow			
	Overall	Vehicle	Pedestrian	Cyclist
6	74.67 / 72.20	72.39 / 72.07	76.67 / 71.05	74.96 / 73.48
8	75.13 / 72.69	72.93 / 72.45	77.18 / 71.66	75.27 / 73.96
10	75.19 / 72.74	73.04 / 72.54	76.98 / 71.55	75.57 / 74.12

Table D.6: Ablation study of the type of decoder.

Decoder type	L2 (AP/APH) \uparrow			
	Overall	Vehicle	Pedestrian	Cyclist
sparse conv.	74.52 / 71.99	72.06 / 71.77	76.45 / 70.79	75.04 / 73.40
SST	74.77 / 72.25	72.42 / 72.03	76.76 / 71.17	75.12 / 73.54
DSVT	75.13 / 72.69	72.93 / 72.45	77.18 / 71.66	75.27 / 73.96

to pre-training on nuScenes, there is still an improvement of 1.32% compared to no pre-training. When pre-trained on nuScenes and fine-tuned on Waymo, there is an improvement of 0.87% compared to training from scratch. In particular, when pre-trained on the joint dataset, optimal fine-tuning performance is achieved both on Waymo and nuScenes. This highlights the value of leveraging abundant unlabeled point cloud data.

D.8 Comparisons of Different Mask Sampling Strategies

The comparison of different mask sampling strategies is presented in Tab. D.8. We compare our approach with random masking [6], range-aware random masking [4], and FPS-based masking [8]. The key distinction is that only our proposed l^2 Mask is based on feature attributes rather than voxel positions. Compared to random masking, our method achieves an improvement of 0.39% in L2 mAP. This indicates that considering the inter-class and intra-class discrimination of Seal voxel features can further unlock the potential of self-supervised learning.

D.9 Visualization of Feature Heatmap

Feature heatmaps for more scenes are displayed in Fig. D.1. By using the reconstruction of Seal voxel features as a pretext task, output features of the encoder become more aligned with Seal features. This is beneficial for downstream tasks that rely on semantics.



Figure D.1: Heatmaps for more scenes. Each row represents a scene.

Table D.7: Self-supervised learning across datasets and joint datasets in 3D object detection.

Pre-train	Downstream	Waymo		nuScenes	
		L2 mAP	L2 mAPH	mAP	NDS
×		73.20	71.00	66.40	71.10
Waymo		75.13 ^{+1.93}	72.69 ^{+1.69}	67.72 ^{+1.32}	72.43 ^{+1.33}
nuScenes		74.07 ^{+0.87}	72.08 ^{+1.08}	68.10 ^{+1.70}	72.60 ^{+1.50}
Waymo + nuScenes		75.33 ^{+2.13}	72.91 ^{+1.91}	68.43 ^{+2.03}	72.76 ^{+1.66}

Table D.8: Comparisons of different mask sampling strategies.

Pre-train	Reconstruction target	Masking	L2 mAP	L2 mAPH
None	-	-	73.20	71.00
PICTURE	Coord. + Seal	Random Masking	74.74	72.56
	Coord. + Seal	Range-aware Random Masking	74.51	72.40
	Coord. + Seal	FPS-based Masking	74.57	72.44
	Coord. + Seal	l ² Mask	75.13	72.69

E SOCIETAL IMPACTS

Firstly, autonomous driving is trapped in significant domain adaptation problems, where the transferability of model between different data distributions is inadequate, hindering its ability to generalize to new scenes. We propose using high-level features as reconstruction targets, which allows the network to learn more universal representations from autonomous driving scenes. These features can better generalize to data from new scenes. **Secondly**, we introduce novel insights into 3D generative self-supervised learning, thereby motivating the autonomous driving community to develop more robust point cloud encoders with semantic information.

REFERENCES

- [1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. 2019. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9297–9307.
- [2] Lue Fan, Ziqi Pang, Tianyuan Zhang, et al. 2022. Embracing single stride 3d object detector with sparse transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8458–8468.
- [3] Whye Kit Fong, Rohit Mohan, Juana Valeria Hurtado, Lubing Zhou, et al. 2022. Panoptic nusenes: A large-scale benchmark for lidar panoptic segmentation and tracking. *IEEE Robotics and Automation Letters* 7, 2 (2022), 3795–3802.
- [4] Chen Min, Liang Xiao, Dawei Zhao, et al. 2023. Occupancy-MAE: Self-Supervised Pre-Training Large-Scale LiDAR Point Clouds With Masked Occupancy Autoencoders. *IEEE Transactions on Intelligent Vehicles* (2023).
- [5] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, et al. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2446–2454.
- [6] Xiaoyu Tian, Haoxi Ran, et al. 2023. GeoMAE: Masked Geometric Target Prediction for Self-supervised Point Cloud Pre-Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13570–13580.
- [7] Haiyang Wang, Chen Shi, Shaoshuai Shi, Meng Lei, Sen Wang, Di He, Bernt Schiele, and Liwei Wang. 2023. Dsvt: Dynamic sparse voxel transformer with rotated sets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13520–13529.
- [8] Runsen Xu, Tai Wang, Wenwei Zhang, Runjian Chen, Jinkun Cao, Jiangmiao Pang, and Dahua Lin. 2023. MV-JAR: Masked Voxel Jigsaw and Reconstruction for LiDAR-Based Self-Supervised Pre-Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13445–13454.
- [9] Yan Yan, Yuxing Mao, and Bo Li. 2018. Second: Sparsely embedded convolutional detection. *Sensors* 18, 10 (2018), 3337.
- [10] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. 2021. Cylindrical and asymmetrical 3d convolution networks

for lidar segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9939–9948.