A. Appendix

This appendix contains the following sections: Sec. A.1 reports more experimental results; Sec. A.2 presents visualization results on the Waymo dataset.

A.1. More experiments

Results on the KITTI Test Set. GD-MAE is compared with previous approaches using different 3D backbones, e.g., sparse convolutions (SpCNN) and Transformer, on the KITTI *test* set. As illustrated in Table 1, GD-MAE achieves competitive results.

Table 1. Performance comparisons on the KITTI *test* set with AP calculated by 40 recall positions for the car class.

Methods	Backbone	3D Easy	3D Moderate	3D Hard
PointPillars [1]	CNN	82.58	74.31	68.99
SECOND [3]	SpCNN	84.65	75.96	68.71
VoTr-SSD [2]	Transformer	86.73	78.25	72.99
IA-SSD [4]	PointNet++	88.34	80.13	75.04
GD-MAE (Ours)	**Transformer**	**88.14**	**79.03**	**73.55**

The Number of Points. As different tokens contain a varying number of points, we randomly sample at most K points as the target for reconstruction. Table 2 shows the performance when different K is adopted.

Table 2. Ablation study of the number of the sampled points for the reconstruction target.

K	32	64	128
Vehicle	**66.57**	66.54	66.49
Pedestrian	64.64	**64.93**	64.67

Pre-training Epochs. Table 3 shows the effect of the pre-training epochs. We find that using more epochs can further improve performance, which demonstrates the learning capability of our model. In the main text, all models are only pre-trained for 30 epochs to save training time.

Table 3. Ablation study of the epoch for pre-training.

Epoch	10	30	60	120
Vehicle	66.23	66.54	66.82	**66.89**
Pedestrian	64.61	64.93	64.95	**65.20**

Different Pre-training Datasets. To prevent overfitting on the same dataset, we pre-train the model on the ONCE dataset and then fine-tune it on the Waymo dataset. As shown in the third row of Table 4, GD-MAE consistently boosts the accuracy of all categories.

Table 4. Ablation study of different pre-training datasets.

w/ GD-MAE	Data.	Vehicle	Pedestrian	Cyclist
	-	65.55	63.76	66.75
✓	Waymo	**66.54**$^{\uparrow 0.99}$	**64.93**$^{\uparrow 1.17}$	**67.75**$^{\uparrow 1.00}$
✓	ONCE	**67.18**$^{\uparrow 1.63}$	**64.82**$^{\uparrow 1.06}$	**67.83**$^{\uparrow 1.08}$

Masking Ratio. The impact of various masking ratios is displayed in Table 5. We discover that a ratio of 75% works best for creating a task that is adequately difficult for self-supervised pre-training. If the masking ratio is too high, performance suffers dramatically. The accuracy also degrades slightly with low making ratios.

Table 5. Ablation study of different masking ratios. Using 5% labeled data for fine-tuning.

Ratio	0.55	0.65	0.75	0.85	0.95
Vehicle	62.32	62.43	**62.65**	62.55	61.56
Pedestrian	60.91	61.19	**61.44**	60.97	60.19

The effect in multi-scale scenes. The vehicle category in the Waymo and ONCE datasets usually includes cars, buses, and trucks, which range from 4 to 12 meters in length. The overall vehicle gain demonstrates the effectiveness of GD-MAE in multi-scale scenes. In Table 6, we provide the results of subclasses of vehicle on the ONCE dataset.

Table 6. Ablation study of different classes on the ONCE dataset.

w/ GD-MAE	Car	Bus	Truck	Pedestrian	Cyclist
	76.94	59.31	34.45	45.92	66.30
✓	**77.57**$^{\uparrow 0.63}$	**67.08**$^{\uparrow 7.77}$	**39.82**$^{\uparrow 5.37}$	**48.84**$^{\uparrow 2.92}$	**69.14**$^{\uparrow 2.84}$

A.2. Qualitative Results

Figure 1 shows several examples of the reconstructed point clouds on the Waymo validation set. The model catches the distinctive LiDAR scans along the ground plane and demonstrates a knowledge of the basic geometry. Figure 2 illustrates the detection results of our method on the Waymo validation set. Our model can predict accurate bounding boxes for distant and highly occluded objects, demonstrating the high-quality predictions of our model.

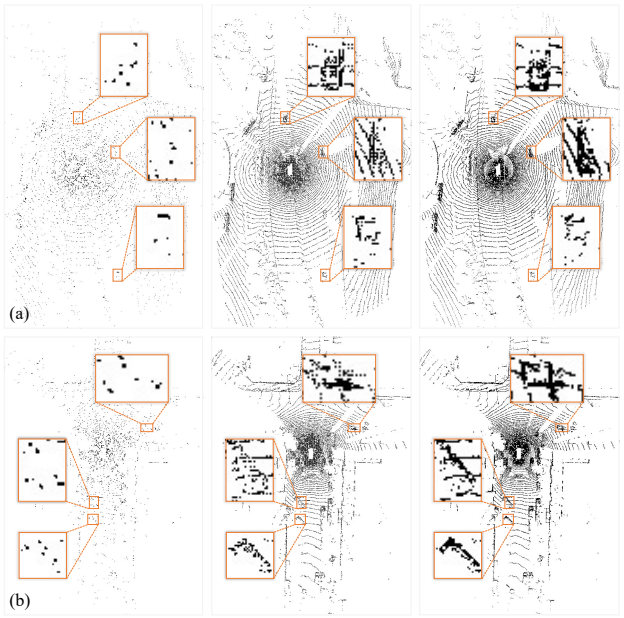

Figure 1. Reconstruction results on the Waymo validation set. On the left is the visible input, in the middle is the result of the reconstruction and on the right is the ground truth.

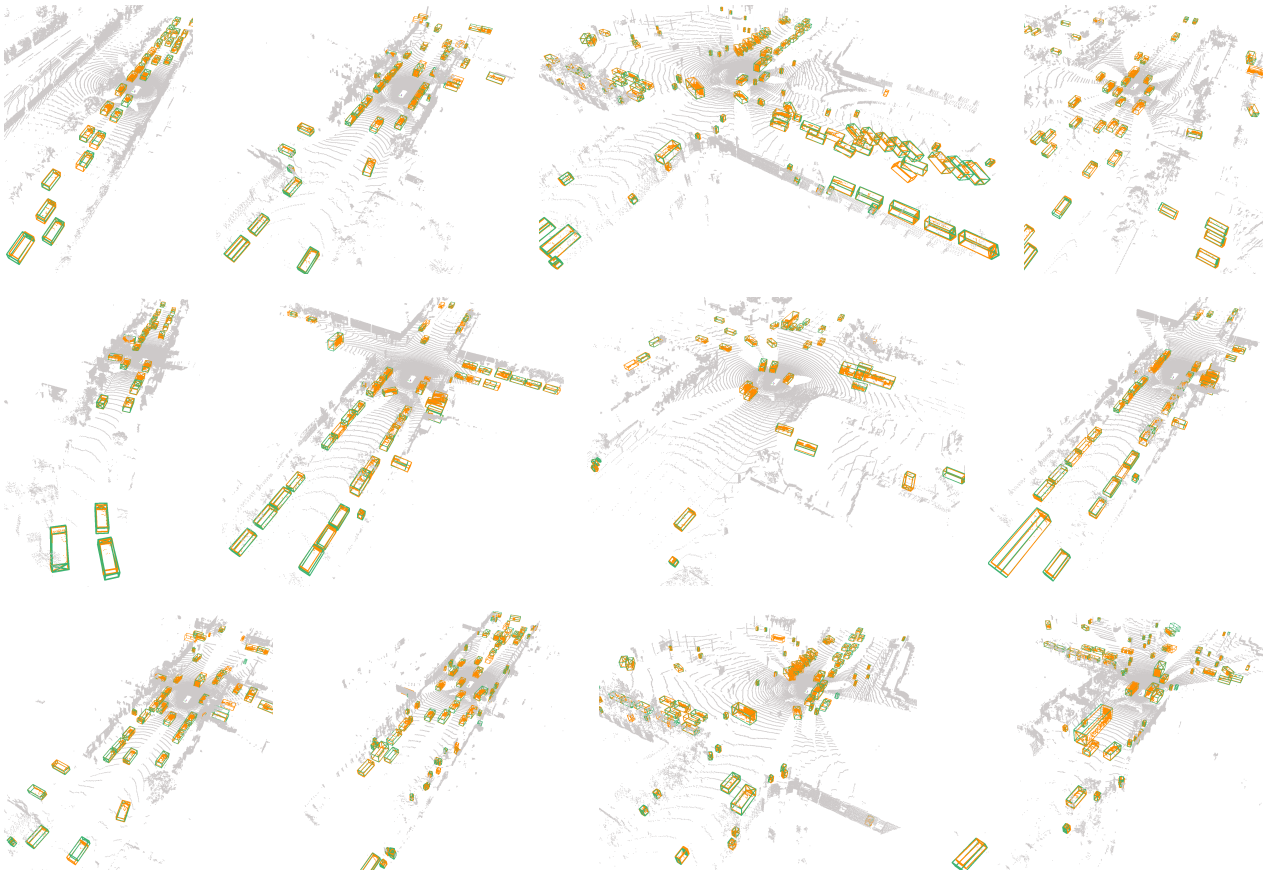

Figure 2. Qualitative results of 3D object detection on the Waymo validation set. We show the raw point cloud in gray, points inside our detected bounding boxes in orange, ground truth in green bounding boxes, and our detected objects in orange bounding boxes.

References

[1] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1

[2] Jiageng Mao, Yujing Xue, Minzhe Niu, Haoyue Bai, Jiashi Feng, Xiaodan Liang, Hang Xu, and Chunjing Xu. Voxel transformer for 3d object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 1

[3] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10), 2018. 1

[4] Yifan Zhang, Qingyong Hu, Guoquan Xu, Yanxin Ma, Jianwei Wan, and Yulan Guo. Not all points are equal: Learning highly efficient point-based detectors for 3d lidar point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1