A Appendix for AD-PT: Autonomous Driving Pre-training with Large-scale Point Cloud Dataset

In this supplementary material, we provide more details and experimental results not included in our main text.

Outlines:

B More Details about Large-scale Pre-training Dataset Preparation.

In this section, we give some preliminary experimental results and analysis on large-scale pre-training dataset preparation.

B.1 Preliminary Experiments on Class-aware Pseudo Label Generator

As mentioned in Sec. 3.2.1 in our submission, we explore how to improve the performance on ONCE. We first analyze the results in the ONCE benchmark and find that CenterPoint reaches the SOTA performance on pedestrian and cyclist while PV-RCNN achieves the best performance on vehicle. To use a stronger baseline to further improve the performance, we conduct experiments using PV-RCNN++ as the baseline detector. As shown in Tab. 10, PV-RCNN++ with center head can not obtain a satisfactory performance on ONCE while PV-RCNN++ with anchor head can achieve better accuracy on vehicle and pedestrian.

Further, to obtain more accurate pseudo labels, we use a semi-supervised learning method to further improve the performance as shown in Tab. 11. Finally, we individually train pedestrian using CenterPoint and other classes using PV-RCNN++.

Table 10: Effects of using different heads on PV-RCNN++. We report mAP using the ONCE evaluation metric.

Detector	Head Choice	Vehicle	Pedestrian	Cyclist
PV-RCNN++	Center Head	71.61	**45.27**	61.15
PV-RCNN++	Anchor Head	**81.72**	43.86	**66.17**

B.2 Analysis on Pseudo Label Threshold on Different Classes

Fig. 7 shows the precision under different IoU thresholds. The precision can be calculated by $\text{Precision} = \text{TP}/(\text{FP} + \text{TP})$, where FP and TP denote false positive and true positive, respectively. We can observe that when IoU thresholds are more than 0.8, 0.7, 0.7 for vehicle, pedestrian and cyclist, the number of TP instances is significantly more than that of FP instances.

Table 11: Effects of using MeanTeacher. We report mAP using the ONCE evaluation metric.

Detector	MeanTeacher	Vehicle	Pedestrian	Cyclist
CenterPoint	✗	-	46.22	-
CenterPoint	✓	-	**56.01**	-
PV-RCNN++	✗	81.72	-	66.17
PV-RCNN++	✓	**82.50**	-	**71.19**

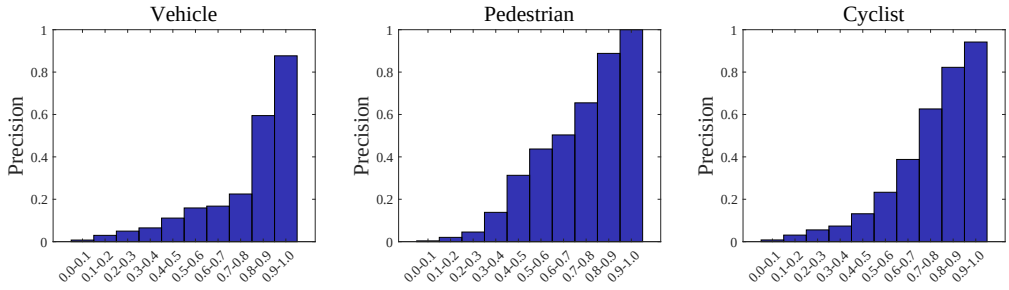

Figure 7: The Precision under different IoU thresholds.

The visualization of the pseudo label results under different thresholds in Fig. 8, we can see that some FP pseudo labels will be annotated when setting low thresholds, while some TP instances can not be annotated when the thresholds are relatively high. To more intuitively see the impact of the threshold on pseudo labeling, we use the model to annotate the samples of the ONCE validation set for comparison with ground-truths.

B.3 Visualization Results of Pseudo Labels

Fig. 9 shows the visualization results of our final pseudo label results.

B.4 Details of Object Re-scaling

In detail, given a bounding box $b = (c_x, c_y, c_z, l, w, h, \theta_h)$ and point clouds (p_i^x, p_i^y, p_i^z) within it, where (c_x, c_y, c_z), (l, w, h) and θ_h denote the center, size and heading angle of the bounding box. We first transform points into the local coordinate with the following formula:

$$
\begin{aligned}
(p_i^l, p_i^w, p_i^h) &= (p_i^x - c_x, p_i^y - c_y, p_i^z - c_z) \cdot R, \\
R &= \begin{bmatrix} \cos\theta_h & -\sin\theta_h & 0 \\ \sin\theta_h & \cos\theta_h & 0 \\ 0 & 0 & 1 \end{bmatrix},
\end{aligned} \tag{1}
$$

where $\cdot$ is matrix multiplication. Then, to derive the scaled object, the point coordinates inside the box and the bounding box size are scaled to be $\alpha(p_i^l, p_i^w, p_i^h)$ and $\alpha(l, w, h)$, where α is the scaling factor. Finally, the points inside the scaled box are transformed back to the ego-car coordinate system and shifted to the center (c_x, c_y, c_z) as

$$
\tilde{p_i} = \alpha(p_i^l, p_i^w, p_i^h) \cdot R^T + (c_x, c_y, c_z). \tag{2}
$$

B.5 Taxonomy difference between different datasets

As shown in Tab. 12, there exists a huge taxonomy difference between some fine-tuning datasets and the pre-training dataset. As a result, some foreground instances may be regarded as background if only using pseudo label as supervision.

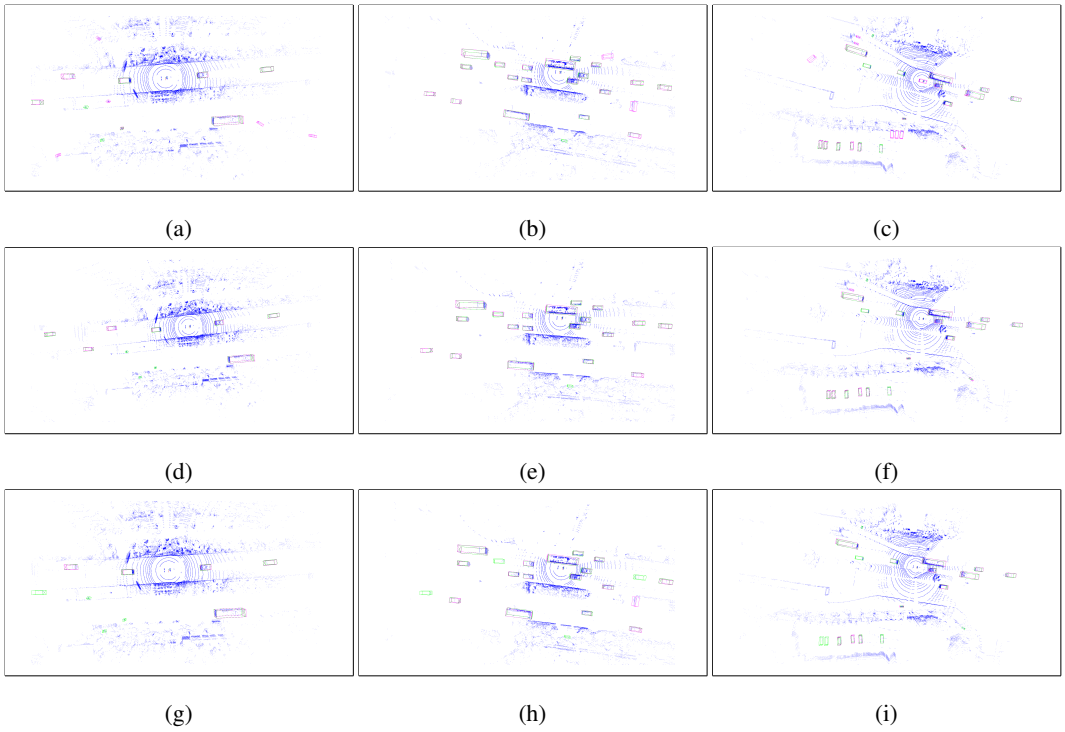

Figure 8: Visualization results under different pseudo label thresholds. (a-c): annotations with low thresholds (*i.e.*, 0.6, 0.5, 0.5 for vehicle, pedestrian and cyclist, respectively). (d-f): the thresholds used in our methods. (g-i): high thresholds (*i.e.*, 0.9, 0.8, 0.8 for vehicle, pedestrian and cyclist, respectively). The green and red bounding boxes represent ground-truths and detector predictions, respectively.

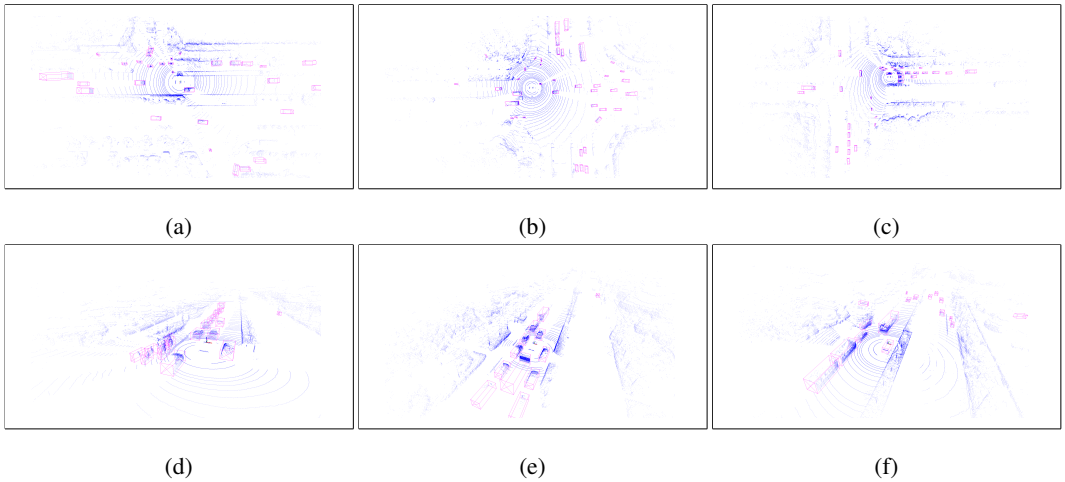

Figure 9: Pseudo-labeled annotation results on unlabeled set.

C Detailed Dataset Description and Evaluation Metrics

C.1 Dataset Description

ONCE dataset. ONCE dataset [4] is a large-scale dataset that is built to encourage the exploration of self-supervised and semi-supervised learning in the autonomous driving scenario. ONCE is collected by a 40-beam LiDAR in multiple cities in China and contains diverse weather conditions

Table 12: Taxonomy difference between different datasets.

Dataset	classes
ONCE (Pre-train)	Car, Truck, Bus, Pedestrian, Cyclist
Waymo (Fine-tune)	Vehicle, Pedestrian, Cyclist
nuScenes (Fine-tune)	Car, Truck, Construction vehicle, Bus, Trailer, Barrier, Motorcycle, Bicycle, Pedestrian, Traffic cone
KITTI (Fine-tune)	Car, Pedestrian, Cyclist

(*e.g.*, sunny, cloudy, rainy), traffic conditions, time periods (*e.g.*, morning, noon, afternoon, night) and areas (*e.g.*, downtown, suburbs, highway, tunnel, bridge).

Waymo Open Dataset. Waymo Open Dataset [5] is a widely-used large-scale autonomous driving dataset that is composed of 1000 sequences and divided into a train set with 798 sequences ($\sim$150k samples) and a validation set with 202 sequences ($\sim$40k samples). The Waymo dataset is gathered in the USA by a 64-beam LiDAR and 4 200-beam short-range LiDAR with annotations in full 360°. We use the 1.0 version of Waymo Open Dataset.

nuScenes Dataset. NuScnenes dataset [1] provides point cloud data from 32-beam LiDAR collected from Singapore and Boston, USA. It consists of 28130 training samples and 6019 validation samples. The data is obtained during different times in the day, different weather conditions and a diverse set of locations (*e.g.*, urban, residential, nature and industrial).

KITTI Dataset. KITTI dataset [2] is a common-used autonomous driving dataset that contains 7481 training samples and is divided into a train set with 3712 samples and a validation set with 3769 samples. The point cloud data is collected by a 64-beam LiDAR in Germany. KITTI dataset only provides the annotations for the objects within the field of view of the front RGB camera.

C.2 Evaluation Metrics

ONCE evaluation metric. Following ONCE official evaluation metric, we merge the car, bus and truck class into a super-class (*i.e.*, vehicle). AP_{3D}^{Ori} is used to evaluate the performance of the ONCE dataset, which can be obtained by the following formula:

$$AP_{3D}^{Ori} = 100 \int_0^1 max\{p(r'|r' \geq r)\}dr, \tag{3}$$

where r is recall rates from 0.02 to 1.00 at step 0.02 and $p(r)$ denotes the precision-recall curve. Mean average precision (mAP) is the average of the scores of the three categories. The Intersection over Union (IoU) thresholds are set to 0.7, 0.3 and 0.5 for vehicle, pedestrian and cyclist, respectively.

Waymo evaluation metric. Two difficulty levels (*i.e.*, LEVEL 1 and LEVEL 2) are utilized to evaluate the detection accuracy of Waymo dataset and we mainly focus on more difficult L2 performance. Among each difficulty level, we report AP and APH which can be formulated as:

$$AP = 100 \int_0^1 max\{p(r')|r' \geq r\}dr, \;\; AP = 100 \int_0^1 max\{h(r')|r' \geq r\}dr, \tag{4}$$

where the different between $h(r)$ and $p(r)$ is $h(r)$ is weighted by the accuracy of heading accuracy.

nuScenes evaluation metric. Following the official NuScenes Evaluation Metric, we report mAP and nuScenes detection score (NDS). AP is defined as matches by thresholding the 2D center distance d on the ground plane and the mAP can be calculated by:

$$mAP = \frac{1}{\mathbb{C}} \frac{1}{\mathbb{D}} \sum_{c \in \mathbb{C}} \sum_{d \in \mathbb{D}} AP, \tag{5}$$

where $\mathbb{C}$ is the set of classes and $\mathbb{D}$ is the set of thresholds (*i.e.*, {0.5,1,2,4}). We mainly focus on 10 classes. NDS is the weighted of mAP and five true positive metrics, including Average Translation

Error (ATE), Average Scale Error (ASE), Average Orientation Error (AOE), Average Velocity Error (AVE) and Average Attribute Error (AAE). The NDS can be formulated as:

$$mTP = \frac{1}{\mathbb{C}} \sum_{c \in \mathbb{C}} TP_c, \quad NDS = \frac{1}{10}[5mAP + \sum_{mTP \in \mathbb{TP}} (1 - min(1, mTP))], \tag{6}$$

where $\mathbb{TP}$ is the set of true positive metrics.

KITTI evaluation metric. We report mAP with 40 recall positions to evaluate the detection performance and the 3D IoU thresholds is set to 0.7 for cars and 0.5 for pedestrians and cyclists.

D More Implementation Details

As shown in Tab. 13, we list some details about pre-training and fine-tuning datasets. Note that the voxel size of nuScenes is set to [0.1, 0.1, 0.2] following [3]. It can be seen that different datasets may have different dimensions of input features (*e.g.*, ONCE use 4 dimension features as input while Waymo and nuScenes use 5 dimension features) causing the input dimension of the first layer network to be different. We simply do not load the parameters of the first layer when this happens while fine-tuning. In the pre-training phase, we merge the pseudo-labeled data and a small amount of labeled data (*i.e.*, ONCE train set) as the pre-training dataset. In the fine-tuning phase, we fine-tune 30 epochs for Waymo, 20 epochs for nuScenes and 80 epochs for KITTI.

Table 13: Some implementation details about pre-training and fine-tuning datasets.

Dataset	Point cloud range	voxel size	input features
ONCE (Pre-train)	[-75.2, -75.2, -5.0, 75.2, 75.2, 3.0]	[0.1, 0.1, 0.2]	[x, y, z, intensity]
Waymo (Fine-tune)	[-75.2, -75.2, -2.0, 75.2, 75.2, 4.0]	[0.1, 0.1, 0.15]	[x, y, z, intensity, elongation]
nuScenes (Fine-tune)	[-51.2, -51.2, -5.0, 51.2, 51.2, 3.0]	[0.1, 0.1, 0.2]	[x, y, z, intensity, timestamp]
KITTI (Fine-tune)	[0.0, -40.0, -3.0, 70.4, 40.0, 1.0]	[0.05, 0.05, 0.1]	[x, y, z, intensity]

E More Experimental Results

E.1 Ablation Studies on Unknown-aware Instance Learning Head

In this part, we conduct experiments to ablate the hyper-parameters in unknown-aware instance learning head (*i.e.*, the number M of selected features and the distance threshold τ).

Tab. 14 shows the results using different numbers of selected features in unknown-aware instance learning head when pre-training. When M is small, some foreground instances with relatively low scores are ignored, while when M is large, the matched background regions are increased. Considering these factors, we choose M to be 256.

Tab. 15 shows the performance under different distance thresholds in Eq. 4 in the main submission. The number of matched features is relatively small when using a lower τ, thus can not fully exploit the unknown foreground instances. When using a larger threshold, some mismatches may occur. Finally, we set τ to 0.3 as mentioned in our main submission.

Table 14: Ablation studies of the number M of selected features.

M	Waymo L2 AP / APH			
	Overall	Vehicle	Pedestrian	Cyclist
128	67.71 / 64.98	67.91 / 67.45	68.54 / 61.87	66.67 / 65.63
256	**68.33 / 65.69**	**68.17 / 67.70**	**68.82 / 62.39**	**68.00 / 67.00**
512	67.93 / 65.24	68.04 / 67.36	68.63 / 62.12	67.14 / 66.23

E.2 More Results of Pre-training Scalability

In this section, we show more results to verify the pre-training scalability. We pre-train the model on the small, middle and large splits of the ONCE dataset and then fine-tune the model on 3% Waymo

Table 15: Ablation studies of the distance threshold τ.

τ	Waymo L2 AP/APH			
	Overall	Vehicle	Pedestrian	Cyclist
0.1	67.90 / 65.22	68.01 / 67.54	68.52 / 62.01	67.17 / 66.12
0.3	**68.33 / 65.69**	**68.17 / 67.70**	**68.82 / 62.39**	**68.00 / 67.00**
0.5	67.82 / 65.15	67.73 / 67.26	68.26 / 61.73	67.49 / 66.46

and 20% KITTI train data. As shown in Tab. 16, as the scale of the pre-training dataset and the diversity of scenarios increases, the performance of fine-tuning on the downstream dataset will also improve.

Table 16: The pre-training scalability. We use ONCE to pre-train and Waymo and KITTI to fine-tune.

Pre-training dataset	Waymo L2 AP/APH				KITTI Moderate mAP			
	Overall	Vehicle	Pedestrian	Cyclist	Overall	Car	Pedestrian	Cyclist
ONCE (∼100k)	68.33 / 65.69	68.17 / 67.70	68.82 / 62.39	68.00 / 67.00	69.43	82.75	57.59	67.96
ONCE (∼500k)	69.04 / 66.52	68.69 / 68.23	69.81 / 63.74	68.61 / 67.60	71.36	83.17	58.14	72.78
ONCE (∼1M)	**69.63 / 67.08**	**69.03 / 68.57**	**70.54 / 64.34**	**69.33 / 68.33**	**72.37**	**83.47**	**59.84**	**73.81**

E.3 Results of fine-tuning on ONCE.

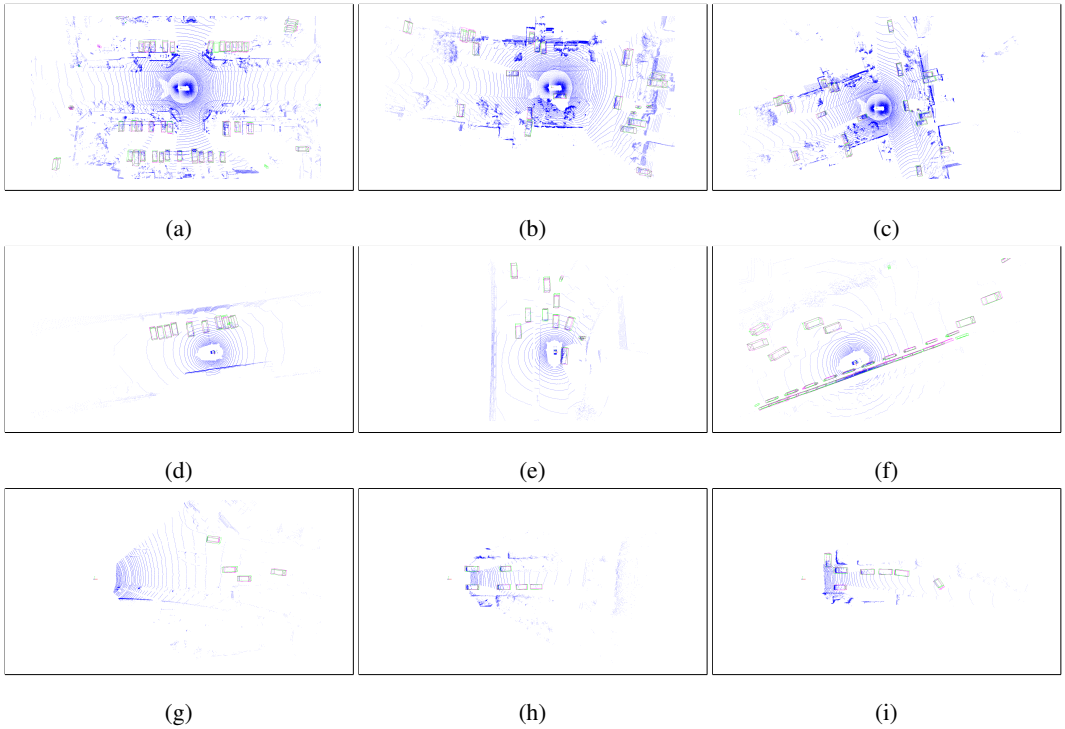

Figure 10: Visualization of fine-tuning results. We visualize the results of three downstream datasets. (a-c): results of Waymo. (d-f): results of nuScenes. (g-i): results of KITTI.
The green and red bounding boxes represent ground-truths and detector predictions, respectively.

In our main submission, we report the fine-tuning performance on multiple datasets which are different from the pre-training dataset. Here, we show some fine-tuning performance on ONCE. As shown in Tab. 17, the performance can be largely improved when the baseline detectors are initialized by AD-PT. For example, when using SECOND as the baseline detector, the overall performance can be improved from 56.47% to 64.10% (+7.63%). We use the ONCE train set to fine-tune the model.

Table 17: The fine-tuning performance on ONCE validation set.

Init.	SECOND				CenterPoint			
	Overall	0-30m	30-50m	>50m	Overall	0-30m	30-50m	>50m
Random Initialization	56.47	65.94	51.05	36.44	64.94	74.52	59.47	44.28
AD-PT Initialization	**64.10**	**74.34**	**57.69**	**41.23**	**67.73**	**76.48**	**61.85**	**46.29**

E.4 Visualization Results.

Fig. 10 shows the visualization results of three downstream datasets (*i.e.*, Waymo, nuScenes, KITTI).

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020. 4

[2] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 4

[3] Zhiwei Lin and Yongtao Wang. Bev-mae: Bird's eye view masked autoencoders for outdoor point cloud pre-training. *arXiv preprint arXiv:2212.05758*, 2022. 5

[4] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Hanxue Liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, et al. One million scenes for autonomous driving: Once dataset. *arXiv preprint arXiv:2106.11037*, 2021. 3

[5] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020. 4