Supplementary Materials

A. Overview

This document presents additional technical details, and provides both quantitative and qualitative results to support the submitted paper. In Sec. B, we discuss the large-scale datasets used in the experiments, and analyze their intrinsic characteristics that cause severe domain shifts. In Sec. C, we elaborate on the network architectures of the 3D detectors employed for comparisons, and describe the implementation details of GPA-3D. In Sec. D, we offer more comprehensive quantitative results and visualizations of our approach.

B. Datasets

We conduct comprehensive experiments on the prevalent autonomous driving datasets, namely Waymo [6], nuScenes [1], and KITTI [2]. These datasets have diverse weather conditions, sensor configurations, foreground styles, and annotation quantities, thereby causing serious domain shifts when adapting a LiDAR-based 3D detector from one dataset to another. Fig. 1 presents randomly selected examples from the aforementioned datasets. Subsequently, we will introduce each dataset in detail.

Waymo. For recent 3D detection task, Waymo [6] is the most large-scale and challenge benchmark, which includes 798 sequences (more than 150,000 frames) for training and 202 sequences (approximately 40,000 frames) for validation. Waymo provides the point clouds captured by a 64-beam LiDAR and 4 200-beam blind LiDAR for each frame. In our experiments, we use the 1.2 version of Waymo and subsample only 50% of the training samples, consistent with ST3D [9] and ST3D++ [10].

nuScenes. The nuScenes [1] dataset comprises of 28,130 samples in the training set and 6,019 samples in the validation set. Point clouds within nuScenes are captured by a 32-beam LiDAR in Boston and Singapore, under diverse weather conditions. To ensure consistency with previous works, we access the performance of transferring 3D detectors across different LiDAR beams by treating all 28,130 training scenes as the target domain.

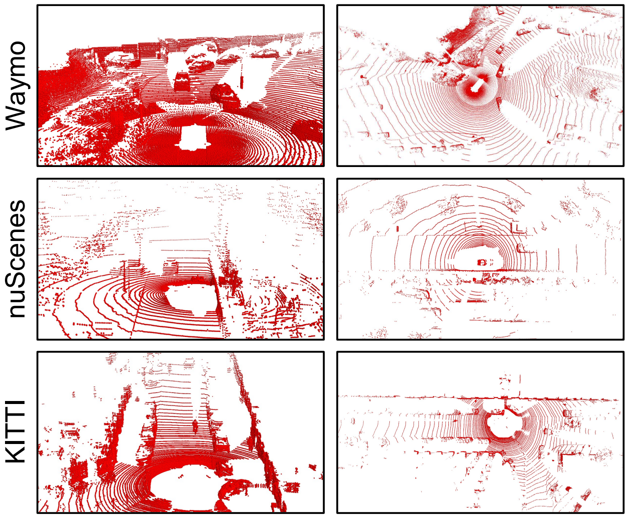

Figure 1. Visualizations of the point clouds for different datasets. Left: Frontal view. Right: Bird's-eye view.

KITTI. As a popular autonomous driving dataset, KITTI [2] contains 7,481 labeled frames for training and 7,518 unlabeled frames for testing. The point clouds of KITTI are captured by a 64-beam Velodyne LiDAR in Karlsruhe, Germany. Following previous approaches, we partition the training frames into two distinct sets: the *train* split, comprising 3,712 samples, and the *val* split, consisting of 3,769 samples.

C. More Implementation Details

Co-training Framework. We follow the default settings of ONCE [4], an open-source 3D detection codebase, to construct the co-training framework in GPA-3D. Specifically, this co-training framework feeds an equal number of point clouds from both source and target domains into the 3D detector in each mini-batch. The outputs generated by the detector are then used for loss computation, with the supervision of ground truth and pseudo-labels, respectively. The calculated losses are subsequently summed together to update the detector parameters and prototypes via the back-propagation method.

Table 1. Comparison with previous works on the pedestrian category. The adaptation scenario is nuScenes → KITTI, and the base detector is SECOND-IoU [8]. For fair comparison, the results are cited from the original paper of ST3D++ [10].

Method	AP_{BEV}	Closed Gap	AP_{3D}	Closed Gap
Source Only	39.95	-	34.57	-
SN [7]	38.91	-16.07%	34.36	-3.11%
ST3D [9]	44.00	$+60.36\%$	42.60	$+118.79\%$
ST3D++ [10]	47.20	$+108.41\%$	43.96	$+138.91\%$
GPA-3D (ours)	**48.17**	**+122.97%**	**45.20**	**+157.25%**
Improvement	*+0.97*	*+14.56%*	*+1.3*	*+18.34%*
Oracle	46.64	-	41.33	-

Detection Architecture. To ensure fair comparisons, we adopt the default configurations of ST3D [9] and ST3D++ [10] to set the voxel size in SECOND-IoU [8] and PointPillars [3] to (0.1m, 0.1m, 0.15m) and (0.2m, 0.2m), respectively. Furthermore, for all datasets utilized in our experiments, we shift the coordinate origins to the ground plane, and separately set the detection ranges of X, Y, Z axes to [-75.2m, 75.2m], [-75.2m, 75.2m], and [-2m, 4m].

Hyper-parameters in GPA-3D. For the geometry-aware prototype alignment, we set the length M_i of the feature sequences to be equal to the number of foreground areas in the i-th BEV feature map. Additionally, we set the prototype numbers to 8 and 4 for the adaptation scenarios of Waymo → KITTI and Waymo → nuScenes, respectively. For the soft contrast loss, we determine the balance coefficients β_1, β_2, and β_3 to be 5, 1, and 5, respectively. In our implementation, we perform the instance replacement augmentation with the probability p_{IRA} of 0.25.

D. Exploration Studies

Extend GPA-3D to Multiple Categories. For autonomous driving vehicles, the detection of pedestrians on the road is also a crucial aspect. In fact, it is easy and effective to extend GPA-3D to other classes. Compared to cars, the geometric variations of pedestrians are smaller, thus we reduce the prototype numbers to 3 for pedestrian. As shown in Tab. 1, GPA-3D improves the pedestrian detection performances to 48.17% AP_{BEV} and 45.20% AP_{3D}, surpassing previous state-of-the-art methods. Compared to ST3D++ [10], our approach achieves 0.97% and 1.3% gains in terms of AP_{BEV} and AP_{3D}, respectively. These improvements demonstrate that GPA-3D has consistent effectiveness on the pedestrian detection.

Why Could Adaptation Method Outperforms the Oracle. In the adaptation scenario of Waymo → KITTI, the AP_{BEV} of GPA-3D has surpassed that of the Oracle method, which is fully supervised by the ground truth of KITTI dataset. We attribute the reason into two aspects. *1) Label-insufficient target domain:* Compared to Waymo, KITTI is a relatively label-insufficient dataset (7,000 *vs.* 150,000). The limited annotations affect the performance of Oracle. *2) Stronger generalization ability:* Our method reduces the feature discrepancy across domains, bringing stronger generalization ability. This makes it easier for model to apply the knowledge learned from source domain to the target domain, thereby improving the final performance.

Table 2. Analysis of different alignment schemes in GPA-3D on Waymo → nuScenes. **Conv.** indicates that an extra branch with three convolution layers are attached to the BEV features for alignment. **Pre.** means to align the intermediate features from the backbone network. **BEV** is the BEV-level alignment in GPA-3D.

Method	w/o align	Conv.	Pre.	BEV
AP_{BEV} / AP_{3D}	35.34 / 20.13	35.92 / 22.37	35.72 / 22.13	**37.25 / 22.54**

Table 3. Comparison on nuScenes → KITTI with PointRCNN [5].

Method	SF-UDA^{3D}	Dreaming	MLC-Net	ST3D++	GPA-3D
Reference	[3DV'20]	[ICRA'22]	[ICCV'21]	[TPAMI'22]	(ours)
0.7 IoU AP_{3D}	54.5	-	55.42	67.51	**67.77**
0.5 IoU AP_{3D}	-	70.3	-	79.93	**81.06**

Analysis of Different Alignment Schemes. We investigate the effects of different alignment schemes in GPA-3D, as shown in Tab. 2. Without alignment, the adaptation performance degrades due to the distributional discrepancy in the feature space. Compared with the policies of Conv. and Pre., our BEV-level alignment achieves superior results, indicating the effectiveness of our approach in directly dealing with the distributional discrepancy problem at BEV features.

Extend GPA-3D to Point-based Architecture. We also try to extend GPA-3D to a point-based 3D detector, PointRCNN [5]. For the point-wise features, we assign prototypes to them based on the geometric information of the objects to which they belong. The results on nuScens→KITTI demonstrate that GPA-3D has the potential to be applied to point-based detectors with minor adjustments.

Qualitative Results. We present more visualizations on the adaptation scenarios of Waymo → KITTI and Waymo → nuScenes in Fig. 2. These qualitative results demonstrate the effectiveness of GPA-3D in improving adaptation performance via reducing the false positive predictions and enhancing the regression accuracy. To further validate the efficacy of our GPA-3D, we employ the t-SNE method to visualize the feature distributions of different approaches, as illustrated in Fig. 3. The results clearly show that GPA-3D clusters the features of the same category in different domains, while also separates the features of different categories. This indicates that GPA-3D provides better alignment of features and facilitates the transferring across domains.

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 *CVPR*, pages 11621–11631, 2020. 1

[2] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361, 2012. 1

[3] 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*, pages 12697–12705, 2019. 2

[4] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Jie Yu, et al. One million scenes for autonomous driving: Once dataset. In *NIPS*, 2021. 1

[5] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3D object proposal generation and detection from point cloud. In *CVPR*, pages 770–779, 2019. 2

[6] 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 *CVPR*, pages 2446–2454, 2020. 1

[7] Yan Wang, Xiangyu Chen, Yurong You, Li Erran Li, Bharath Hariharan, Mark Campbell, Kilian Q Weinberger, and Wei-Lun Chao. Train in germany, test in the usa: Making 3D object detectors generalize. In *CVPR*, pages 11713–11723, 2020. 2

[8] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 2, 4

[9] Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. St3d: Self-training for unsupervised domain adaptation on 3D object detection. In *CVPR*, pages 10368–10378, 2021. 1, 2

[10] Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. St3d++: Denoised self-training for unsupervised domain adaptation on 3D object detection. *TPAMI*, 2022. 1, 2

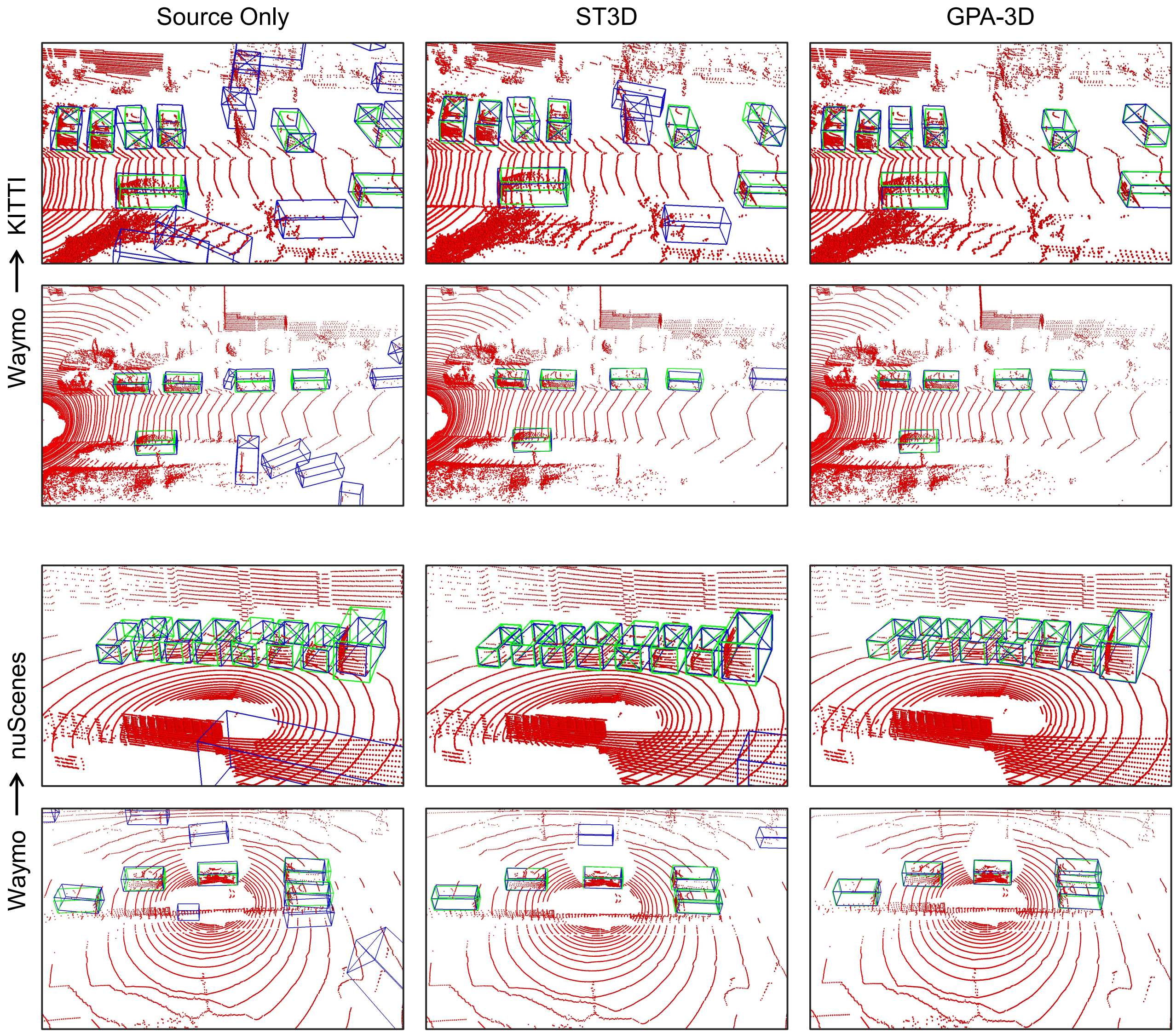

Figure 2. Qualitative results on the adaptation scenarios of Waymo → KITTI and Waymo → nuScenes. For each box, we use the X to specify the orientation. The predicted results and ground truths are painted in blue and green, respectively.

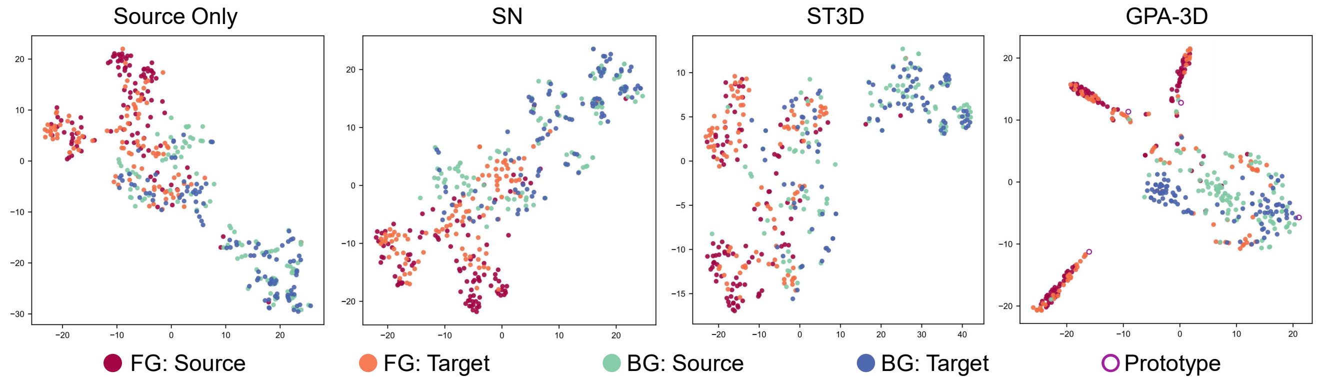

Figure 3. The t-SNE visualization of different methods on Waymo → nuScenes. SECOND-IoU [8] is adopted as the base detector.