

Supplementary Materials of DPO: Dual-Perturbation Optimization for Test-time Adaptation in 3D Object Detection

Anonymous Authors

This supplementary material provides additional descriptions of the proposed DPO, including empirical results and implementation details. Visual aids are also included to enhance understanding of the method. Furthermore, the attached code is available for reference.

- **Sect. 1:** Additional experimental results.
- **Sect. 2:** More implementation details.
- **Sect. 3:** Quantitative study on Waymo \rightarrow KITTI-C task.

1 ADDITIONAL EXPERIMENTAL RESULTS

We adhere to a classical LiDAR-based 3D object detection evaluation, focusing on the car class in the main paper. In addition to this, we also explore the effectiveness of the proposed DPO on two other classes: pedestrians and cyclists. We evaluate TTA baselines and DPO across all difficulty levels for the most challenging transfer task, i.e., composite domain shift, in terms of AP_{3D} . Detailed explanations are provided below.

1.1 Pedestrian Class

We evaluate the effectiveness of DPO for the pedestrian class across all difficulty levels, as shown in Table 1. Notably, our proposed method achieves state-of-the-art performance in terms of mean AP_{3D} , showcasing its effectiveness. When examining specific corruption types, DPO also demonstrates competitive performance. Specifically, for the crosstalk (CrossT.) corruption, DPO improves the AP_{3D} from 40.71% to 42.06%, compared with the strongest baseline SAR, at the hard level. Moreover, our method achieves a 4.1% improvement at the moderate level for the same corruption type. However, there are two exceptions: Beam missing (Beam.) and cross sensor (CrossS.), where a significant number of object points are dropped when generating the corruption, leading to a performance decline for all pseudo-label-based adaptation methods [3], including both CoTTA and DPO. Despite these challenges, our method still manages to handle most corruptions effectively, maintaining a leading mean AP_{3D} .

1.2 Cyclist Class

A similar performance trend is observed in the cyclist class. Our method outperforms the baseline methods at every difficulty level, except for the cross-sensor corruption. Specifically, in terms of mean AP_{3D} , DPO achieves 7.28%, 5.09%, and 5.71% for the easy, moderate, and hard difficulty levels, respectively. Notably, for the snow corruption, our method leads to the greatest improvement over the baseline, increasing from 52.91% to 58.02% AP_{3D} at the easy level. Similarly, a performance increase from 29.34% to 32.05% is achieved at the hard level when facing motion blur (Moti.). For reasons similar to those discussed in Sect. 1.1, DPO underperforms for cross-sensor (CrossS.) corruption, potentially due to failure of the pseudo-labeling strategy when encountering cyclists with

too few points. However, our method represents the best trade-off solution, as it offers the highest mean AP_{3D} .

2 MORE IMPLEMENTATION DETAILS

2.1 Datasets

2.1.1 Waymo. The Waymo open dataset [6] is a large 3D detection dataset for autonomous driving. It contains 798 training sequences with 158,361 LiDAR samples and 202 validation sequences with 40,077 LiDAR samples. The point clouds feature 64 lanes of LiDAR, corresponding to 180,000 points every 0.1 seconds. In DPO, we train the source model on the Waymo training set.

2.1.2 nuScenes. The nuScenes dataset [1] consists of 1,000 driving sequences, divided into 700 for training, 150 for validation, and 150 for testing. Each sequence is approximately 20 seconds long, with a LiDAR frequency of 20 FPS. The dataset provides calibrated vehicle pose information for each LiDAR frame while offering box annotations every ten frames (0.5s). nuScenes uses a 32-lane LiDAR, which generates approximately 30,000 points per frame. In total, there are 28,000 annotated frames for training, 6,000 for validation, and 6,000 for testing. We employ its training set for pre-training the source model for all baselines and the proposed DPO.

2.1.3 KITTI. The KITTI Dataset [2] is widely recognized as a crucial resource for 3D object detection in autonomous driving. The training point clouds are divided into a training split of 3,712 samples and a validation split of 3,769 samples. The dataset categorizes detection difficulty into three levels, defined by criteria of visibility, occlusion, and truncation. The category ‘Easy’ denotes scenarios with no occlusion and a truncation limit of 15%. ‘Moderate’ applies to conditions with partial occlusion and truncation not exceeding 30%. ‘Hard’ encompasses situations with severe occlusion and a truncation threshold of 50%. For evaluating the predicted boxes in 3D object detection, KITTI requires a minimum 3D bounding box overlap of 70% for cars and 50% for pedestrians and cyclists. In this study, where KITTI serves as the target domain, we evaluate all models using the validation split.

2.1.4 KITTI-C. The robustness of 3D perception systems against natural corruptions, which arise due to environmental and sensor-related anomalies, is crucial for safety-critical applications. While existing large-scale 3D perception datasets are often meticulously curated to exclude such anomalies, this does not accurately represent the operational reliability of perception models. KITTI-C [3] is the first comprehensive benchmark designed to assess the robustness of 3D detectors in scenarios involving out-of-distribution natural corruptions encountered in real-world environments. It specifically investigates three major sources of corruption likely to impact real-world deployments: 1) severe weather conditions such as fog, rain (Wet.), and snow, which affect laser pulse dynamics

Table 1: TTA-3OD results (easy/moderate/hard AP_{3D}) of pedestrian class under the composite domain shift (Waymo → KITTI-C) at heavy corruption level.

	No Adaptation	Tent	CoTTA	SAR	MemCLR	DPO
Fog	30.48/26.15/23.61	31.22/26.68/23.99	31.29/26.69/24.05	30.68/25.94/23.70	30.51/26.02/23.77	33.25/27.83/25.23
Wet.	49.10/44.44/41.74	49.13/44.58/41.85	49.14/45.01/42.23	49.18/44.59/41.97	49.09/44.55/41.81	50.31/45.27/42.31
Snow	47.22/42.26/39.19	47.55/42.79/39.44	46.30/41.62/38.11	47.42/42.85/39.54	47.68/42.78/39.45	48.06/43.61/40.11
Moti.	27.18/25.02/23.29	27.47/25.25/23.43	27.28/25.43/23.41	27.34/25.15/23.31	27.44/25.19/23.36	27.48/25.57/23.60
Beam.	32.47/27.89/25.27	34.50/30.55/28.18	32.22/27.41/25.13	34.83/30.74/28.54	34.53/30.30/28.16	34.29/30.42/28.13
CrossT.	47.42/43.08/40.37	47.66/43.37/40.51	47.76/43.29/40.39	48.13/43.65/40.71	47.87/43.58/40.48	50.38/45.43/42.06
Inc.	49.28/44.79/42.21	49.18/44.80/42.11	49.36/45.39/42.77	49.22/44.70/42.24	49.01/44.76/42.11	50.83/46.06/43.02
CrossS.	22.46/18.40/16.08	27.70/22.82/20.30	22.11/17.88/15.98	27.99/23.20/21.36	27.23/23.70/21.63	25.32/20.93/18.99
Mean	38.20/34.00/31.47	39.30/35.11/32.48	38.18/34.09/31.52	39.35/35.10/32.67	39.17/35.11/32.60	39.99/35.64/32.93

Table 2: TTA-3OD results (easy/moderate/hard AP_{3D}) of cyclist class under the composite domain shift (Waymo → KITTI-C) at heavy corruption level.

	No Adaptation	Tent	CoTTA	SAR	MemCLR	DPO
Fog	21.15/17.91/16.66	23.62/19.21/18.33	22.60/18.74/17.57	23.49/19.02/18.10	23.43/19.01/17.86	23.83/19.61/18.64
Wet.	60.36/49.61/47.20	59.72/48.57/45.96	61.36/49.27/47.04	57.43/46.48/43.96	57.76/46.34/44.79	62.64/50.78/48.50
Snow	48.87/40.37/37.96	52.81/42.25/40.11	52.91/41.55/39.09	52.09/41.89/39.26	52.24/41.71/39.28	58.02/44.09/42.18
Moti.	34.62/29.25/27.33	36.79/29.04/27.31	40.37/31.18/29.34	37.53/29.78/28.12	38.19/29.75/28.19	44.03/34.23/32.05
Beam.	32.48/22.42/21.26	36.08/25.03/23.85	30.89/21.37/20.42	37.16/26.21/24.74	36.34/25.35/24.32	38.65/26.78/25.45
CrossT.	59.56/48.75/46.20	58.72/49.26/46.51	62.14/49.13/46.21	58.66/49.16/46.61	58.68/48.87/46.40	63.57/51.07/48.22
Inc.	59.62/49.03/46.82	59.14/47.87/45.11	59.89/47.62/45.44	58.86/47.93/45.51	58.91/48.28/45.68	62.34/50.18/47.73
CrossS.	18.38/11.40/10.93	24.84/15.04/14.66	20.98/12.77/12.28	26.19/15.29/14.95	25.46/15.06/14.49	24.23/14.28/13.99
Mean	41.88/33.59/31.79	43.96/34.53/32.73	43.89/33.95/32.17	43.93/34.47/32.66	43.88/34.30/32.63	47.16/36.38/34.60

through back-scattering, attenuation, and reflection; 2) external disturbances including bumpy surfaces, dust, and insects, which can cause motion blur (Moti.) and missing LiDAR beams (Beam.); and 3) internal sensor failures like incomplete echo (Inc.) or misidentification of dark-colored objects and sensor crosstalk (Cross.T), which may compromise 3D perception accuracy. Additionally, understanding cross-sensor discrepancies is essential to mitigate risks associated with sudden failures due to changes in sensor configurations (Cross.S).

2.2 Additional Implementation Details

Tent and SAR [5, 8] utilize entropy minimization to optimize the batchnorm layers during test time. Therefore, we calculate the entropy loss by summing the classification logits for all proposals at the first detection stage. **CoTTA** [9] follows a mean-teacher framework. Although a broad range of data augmentations is typically required to generalize the model to various corruptions, our empirical evidence from test-time adaptation for 3D object detection (TTA-3OD) suggests that most augmentations do not improve—and may even impair—performance. The sole exception is random world scaling. As a result, we adopt random world scaling as our primary strategy, in accordance with [4], applying strong scaling (0.9 to 1.1) and weak scaling (0.95 to 1.05), respectively. Regarding pseudo-labeling, we directly apply strategies tailored for 3D object detection from [10, 11] to enhance self-training in CoTTA. Similar to CoTTA, we adopt the same augmentation strategy for **MemCLR** [7], which

was originally tailored for image-based 2D object detection, and extend it to 3D detection scenarios. This involves reading and writing pooled region of interest (RoI) features extracted during the second detection stage and computing the memory-based contrastive loss. We apply all hyperparameters from the original paper by default. Besides, The proposed **DPO** is a pseudo-labeling-based self-training approach for TTA-3OD. We leverage the self-training paradigm and augmentation strategies from prior works [4, 10, 11]. The complete configuration files and implementation code are included in this supplementary material.

3 QUANTITATIVE STUDY

Figure 1 visualizes the box predictions from the source pre-trained 3D detector, the proposed DPO, and the ground truth labels. The detection model, pre-trained on Waymo, is adapted to KITTI-C under conditions simulating heavy snowfall, where many noisy green points are distributed throughout the point clouds. The last row displays images of the same testing scenes with projected 2D ground truth boxes. All detected instances in the point clouds are enclosed in blue 3D boxes. Intuitively, DPO demonstrates its ability to better align with the ground truth labels, evidenced by more accurate locations and fewer false positives. In comparison, direct inference often results in a greater number of boxes that do not contain actual objects, caused by a significant domain shift (i.e., cross-dataset plus heavy snow). Additionally, in the first column, a car obscured behind the white car on the left is missed by the ground

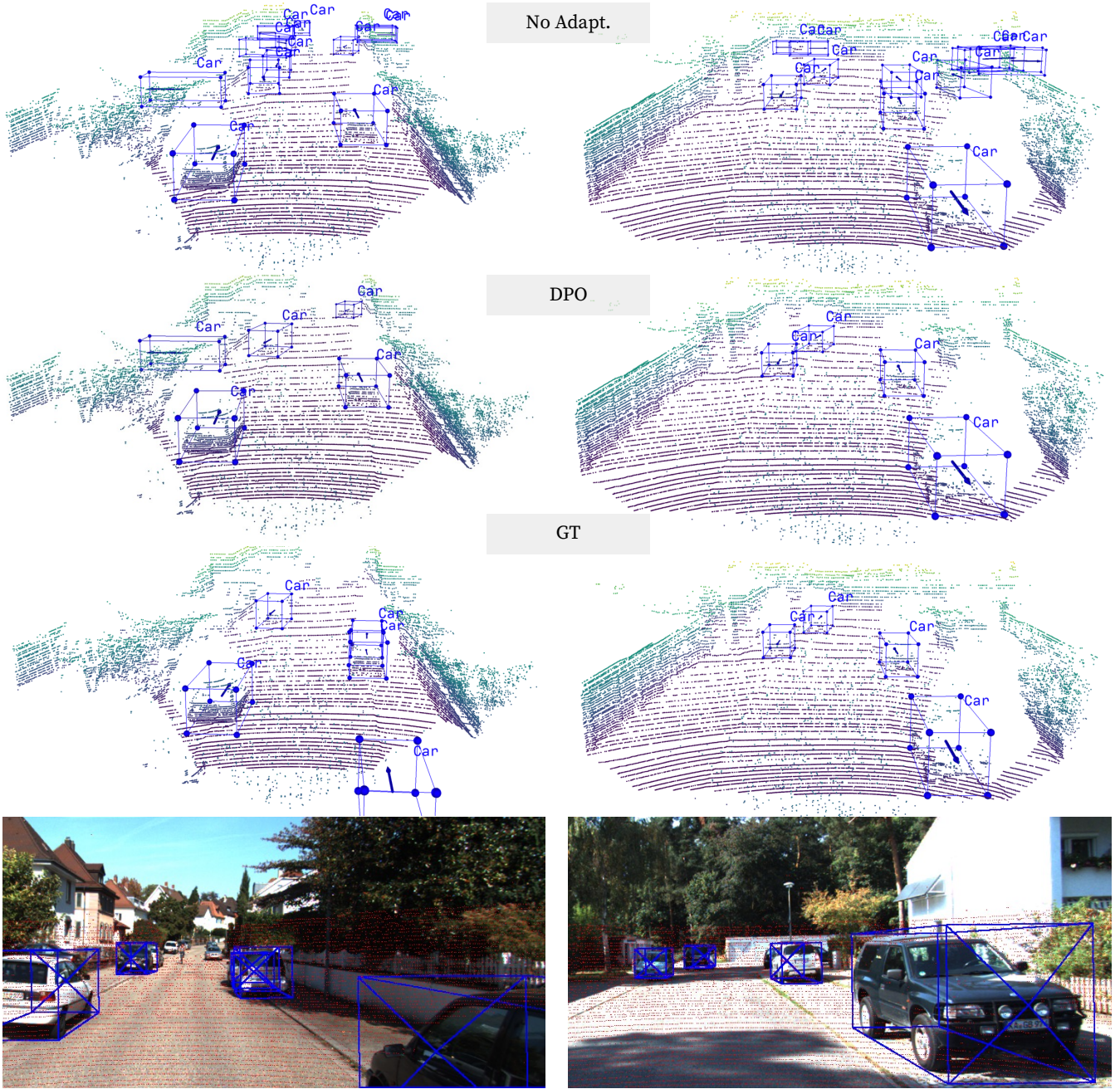


Figure 1: Visualization of box predictions comparing direct inference (No Adapt.), the proposed DPO, and the ground truth labels, across a composite domain shift scenario (Waymo → KITTI-C) under heavy snow conditions.

truth but detected by both DPO and direct inference. While direct inference achieves high recall, it does so at the cost of numerous false positives (*i.e.*, boxes without actual objects). Conversely, the proposed DPO not only demonstrates high recall but also maintains high precision, effectively reducing false positives and confirming its effectiveness in test-time adaptation for 3D object detection.

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. 2020. nuScenes: A Multimodal Dataset for Autonomous Driving. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 11618–11628.
- [2] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3354–3361.
- [3] Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. 2023. Robo3D: Towards Robust and Reliable 3D Perception against Corruptions. In *Proc. International Conference on Computer Vision (ICCV)*. 19937–19949.
- [4] Zhipeng Luo, Zhongang Cai, Changqing Zhou, Gongjie Zhang, Haiyu Zhao, Shuai Yi, Shijian Lu, Hongsheng Li, Shanghang Zhang, and Ziwei Liu. 2021. Unsupervised Domain Adaptive 3D Detection with Multi-Level Consistency. In *Proc. International Conference on Computer Vision (ICCV)*. 8846–8855.
- [5] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiqian Wen, Yaofu Chen, Peilin Zhao, and Mingkui Tan. 2023. Towards Stable Test-time Adaptation in Dynamic Wild World. In *Proc. International Conference on Learning Representations (ICLR)*.
- [6] Pei Sun, Henrik Kretschmar, 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. 2020. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2443–2451.
- [7] Vibashan VS, Poojan Oza, and Vishal M. Patel. 2023. Towards Online Domain Adaptive Object Detection. In *Proc. Winter Conference on Applications of Computer Vision (WACV)*. 478–488.
- [8] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *Proc. International Conference on Learning Representations (ICLR)*.
- [9] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. 2022. Continual Test-Time Domain Adaptation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7191–7201.
- [10] Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. 2021. ST3D: Self-Training for Unsupervised Domain Adaptation on 3D Object Detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 10368–10378.
- [11] Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. 2022. ST3D++: denoised self-training for unsupervised domain adaptation on 3D object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).