Appendix
CMDA: Cross-Modal and Domain Adversarial Adaptation for LiDAR-based 3D Object Detection

Anonymous submission

Overview

In this Appendix, we supply further explanations and visualizations of our main paper, "CMDA: Cross-Modal and Domain Adversarial Adaptation for LiDAR-based 3D Object Detection". Initially, we outline the procedure for generating the BEV feature map from 3D point clouds. Subsequently, we provide details about the implementation and large-scale datasets, including analysis of LiDAR sensors that induce domain shifts across datasets. Also, we conduct additional experimental studies to validate the effectiveness of our novel unsupervised domain adaptation (UDA) framework CMDA. Moreover, we supply more qualitative analysis with diverse scenarios. The code will be released soon after internal discussion.

Points-to-BEV feature map Generation.

Our LiDAR stream starts from common approaches (Deng et al. 2021; Li et al. 2022; Shi et al. 2020; Yan, Mao, and Li 2018) to extract the BEV features from the LiDAR points. The points $P \in \mathbb{R}^{N^p \times 3}$ are first divided into uniformly spaced 3D voxel grids, which are input into a voxel backbone to extract voxel features (illustrated in Fig.1 of our main paper). Then, the voxel features $F_P^{vox} \in \mathbb{R}^{\hat{X} \times Y \times Z \times C}$ are compressed along the height dimension to produce the feature map $F_P^{bev} \in \mathbb{R}^{X \times Y \times ZC}$ in the BEV space, where X, Y, Z, and C represent width, length, height, and the number of channels, respectively.

Experimental Setup.

Datasets.

We evaluate adaptation performance using landmark benchmarks: nuScenes (Caesar et al. 2020), Waymo (Sun et al. 2020), and KITTI (Geiger, Lenz, and Urtasun 2012).

Waymo. The Waymo dataset (Sun et al. 2020) consists of high-quality and large-scale data with 230K frames from all 1,150 scenes using multiple LiDAR scanners and cameras. Furthermore, for the generalization purpose, Waymo is recorded at diverse cities, weather conditions, and times. For object detection in 2D or 3D, Waymo provides point cloud-annotated 3D bounding boxes as 3D data pairs and RGB image-annotated 2D bounding boxes as 2D data pairs.

nuScenes. The nuScenes dataset (Caesar et al. 2020) uses six cameras that cover a full 360-degree range of view and a single LiDAR sensor to obtain 40K frames from all 1,000 scenes. The nuScenes frames are captured in the same manner as Waymo dataset for the data diversity. But unlike Waymo, nuScenes provides labels only for the point cloud data with 23 classes of 3D bounding boxes.

KITTI. The KITTI dataset (Geiger, Lenz, and Urtasun 2012) consists of the point cloud from a single LiDAR sensor and front camera images. Also, compared with Waymo and nuScenes dataset, KITTI is recorded at only the day time and provides 15K frames from all 22 scenes, which is relatively smaller than others. KITTI also provides ground truth correspondance to the point cloud and images labeled with 28 classes for 2D and 3D object detection.

Analysis of Domain Shift between Datasets.

In Tab. 1, we compare large-scale benchmarks, which are primarily used for 3D object detection tasks. There are many differences between them, including LiDAR type, Beam Angles, Points per Beam, Camera View, and Location. While every difference causes domain shifts, variations in the sensor configurations or country-specific factors induce a substantial domain gap across datasets.

Table 1: Dataset details. Note that each statistical information is calculated from the whole dataset. Beam Angles indicates the vertical field of view (VFOV) of 3D sensors.

Dataset	LiDAR Type	Beam Angles	Points per Beam	Camera View	Location
nuScenes	32-beam	[-30.0°, 10.0°]	1,084	Multi-view	USA and Singapore
KITTI	64-beam	[-23.6°, 3.2°]	1,863	Single-view	Germany
Waymo	64-beam	[-18.0°, 2.0°]	2,258	Multi-view	USA

First, transferring knowledge from high-beam data to low-beam data is challenging due to the loss of details. Therefore, addressing this issue becomes crucial in order to fully utilize the potential of large-scale datasets collected using high-beam sensors. Also, each dataset contains data from diverse nations: USA, Singapore and Germany. These discrepancies in the local environment for data collection are significant fatal causes of box scale errors.

Moreover, we observe the inductive bias deriving from annotation policies. For example, Waymo and KITTI have

only one class for "vehicles", whereas nuScenes contains multiple classes, including "car", "truck", "bus", and "construction vehicle". Ultimately, the adaptation of uniform-labeled to previously unseen various-labeled domains often results in poor generalization performance, detecting multiple false positives, and vice versa.

Implementation Details.

We validate the generalizability of our core modules on LiDAR-based 3d object detection models, specifically SECOND-IoU (Yan, Mao, and Li 2018) and PV-RCNN (Shi et al. 2020). Our camera stream processes single or multi-view images of size $(640, 960)$, which are converted through padding and rescaling, and generates the BEV feature map through the Images-to-BEV view transform process. Furthermore, we train our image-assisted source pre-trained model for adversarial domain adaptive self-training on 15 epochs using the Adam optimizer and a learning rate of 1.5×10^{-3}. Note that we use grid search to identify optimal hyperparameters for our approach.

Additional Experimental Results.

In addition to various ablative studies of our main paper, we conduct extensive experiments on our proposed modules. We focus on the self-training process with CDAN, which smartly extends the standard self-training approach (Yang et al. 2021, 2022) to adapt effectively to unfamiliar target data distributions. Our novel domain-adaptive self-training approach adversarially pilots the network to restrict learning domain-invariant cues and enhance the accuracy of pseudo-labels. Specifically, we fool the discriminator to relieve the representational gap between the source vs. target. We first investigate the discriminator in more detail.

Table 2: Quantitative examination on the number of instance-level features $\mathbf{f}_i$, which inputs to adversarial domain discriminator. We evaluate each module in Waymo → nuScenes setting on SECOND-IoU (Yan, Mao, and Li 2018). Note that we highlight the best in **bold** for visibility.

Task	Number of Features	SECOND-IoU	
		BEV AP ↑	3D AP ↑
Waymo (Caesar et al. 2020) → nuScenes (Sun et al. 2020)	1,000	42.18	23.02
	2,000	42.32	24.33
	3,000	**42.81**	**24.64**
	4,000	42.52	24.15
	5,000	42.07	23.59

The discriminator takes instance-level features $\mathbf{f}_i$ from the detection head and predicts where each instance is located in source or target regions. In Tab. 2, we study how the number of instance-level features affects the performance of the self-training process. To discover the optimal number of features, we fix the relevant hyperparameters, such as cost coefficient. Here, we set the weight λ_d of discriminator loss $\mathcal{L}_d$ as 0.05. We then employ a different number of instance-level features for each 1,000 from 1,000 to 5,000 during self-training. As shown in Tab. 2, we confirm that the optimal number for the best performance is 3,000; if it is higher or lower than this, performance improvement becomes less significant.

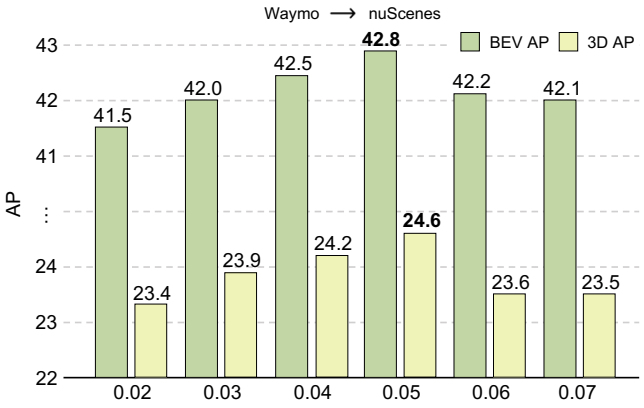

Figure 1: Ablative investigation on the matching cost coefficient λ_d of domain adaptive discriminator loss $\mathcal{L}_d$. Here, we set the number of instances features, the discriminator input, to 3,000. We evaluate each module in Waymo → nuScenes setting on SECOND-IoU (Yan, Mao, and Li 2018).

Then, we explore the impact of the coefficient λ_d of cost term $\mathcal{L}_d$ on our adaptation performance. Since the discriminator is also affected by the number of input features, we keep the number of input features at 3,000. We execute experiments independently several times, increasing the weight value by 0.01 from 0.02 to 0.07. As shown in Fig. 1, the performance gains as the weight increases, peaks (**42.8 / 24.6**) at 0.05, and then declines.

Table 3: Quantitative examination on λ_d and λ_{ent} in nuScenes → Waymo setting on PV-RCNN (Shi et al. 2020). We report moderate BEV AP and 3D AP of the car category at IoU = 0.5. Note that we highlight the best in **bold**.

Task	CDAN		PV-RCNN	
	λ_d	λ_{ent}	BEV AP ↑	3D AP ↑
nuScenes (Caesar et al. 2020) → Waymo (Sun et al. 2020)	-	-	53.04	42.75
	0	0.5	53.35	43.18
	0	1.0	53.75	43.99
	0.5	0	56.53	45.25
	1.0	0	56.57	45.30
	0.5	1.0	57.46	45.01
	1.0	0.5	58.20	45.51
	1.0	1.0	**58.57**	**45.58**

Next, we concentrate on the ideal balance between the two loss terms, $\mathcal{L}_d$ and $\mathcal{L}_{ent}$, utilized in the CDAN. Each $\mathcal{L}_d$ is the discriminator loss, and $\mathcal{L}_{ent}$ is BEV grid-wise entropy loss for suppressing the uncertainty of features and improv-

Table 4: Quantitative examination on the azimuth angle θ of cross-domain mix-up strategy in nuScenes → KITTI and nuScenes → Waymo setting on PV-RCNN (Shi et al. 2020). We report the azimuth angle and the corresponding ratio within a 3D point cloud scene. Evaluation metrics include moderate BEV AP / 3D AP of the car category at IoU = 0.5. **Bold** indicates the best.

Task	Azimuth angle θ					
	0° (0%)	54° (15%)	108° (30%)	162° (45%)	216° (60%)	Random (0% - 40%)
nuScenes (Caesar et al. 2020) → KITTI (Geiger, Lenz, and Urtasun 2012)	83.66 / 73.25	84.10 / 73.53	84.69 / 74.52	83.38 / 72.82	80.93 / 70.82	**84.85 / 75.02**
nuScenes (Caesar et al. 2020) → Waymo (Sun et al. 2020)	56.26 / 44.73	56.56 / 45.16	57.07 / 45.27	56.04 / 43.55	53.19 / 41.88	**58.57 / 45.58**

ing the reliability of predictions. In Tab 3, we evaluate performance while varying the weight value of each loss term, λ_d and λ_{ent}. First, if only one of the two loss terms is applied, the performance is higher when the weight is 1.0 than when it is 0.5. In particular, we observe that $\mathcal{L}_{ent}$ significantly contributes to the geometrical reliability of the network, increasing **+0.71**% and **+1.24**% in BEV AP and 3D AP. Also, $\mathcal{L}_d$ boosts both BEV AP and 3D AP by **+3.53**% and **+2.55**%, respectively, encouraging the broad recognition of unseen datasets. Subsequently, if both loss terms are applied, the performance is the highest (**58.57**% / **45.58**%) when both weights are 1.0. The suitable harmony of the two loss terms allows the model most stably adapts to the target.

Furthermore, we explore our cross-domain 3D point cloud mix-up strategy, effectively bridging the distributional gap. Specifically, we measure the generalizability based on the azimuth angle θ during the swapping process. We conduct experiments independently several times, increasing the angle by 15% from 0% to 60%. As shown in Tab. 4, we observe that a constant azimuth angle causes inductive bias, which limits the learning of domain-agnostic features. In particular, excessive swapping interferes with the adaptive potential, extracting ambiguous features. However, randomly setting the azimuth angle within a specific range (0% - 40%) for iteration encourages learning domain-invariant features, improving adaptation performance: achieving a maximum +2.07% / +1.77% AP gain compared to $\theta = 0^\circ$.

We ultimately design our adaptive network with optimal parameters and validate the remarkable efficiency of CDAN in the target. Most importantly, the adversarial discriminator thoroughly handles the domain shift effect by misleading the network into incorrectly determining which domain instances belong to. Finally, we achieve the current state-of-the-art through our robust domain adaptive framework.

Qualitative Analysis.

In this section, we describe qualitative visualizations of our proposd model. In Fig. 2, we capture various cases of detection on Waymo (Sun et al. 2020) → nuScenes (Caesar et al. 2020) setting. In scene 1 and 2, Direct Transfer (magenta) and ST3D (red) model output larger boxes than ground-truth boxes (blue), while our predicted boxes (green) are relatively similar. Although the bounding boxes of Waymo are larger than nuScenes by (0.16, 0.15, 0.06), resulting in domain shift, our model overcomes this difficulty and adapts better than other baselines. In scene 3 and 4, we observe that DT and ST3D struggle when adapting from uniform-labeled (vehicle) to various-labeled (car, truck, bus, trailer, and construction vehicle) domains. Also, in scene 5, DT and ST3D fail to perceive extremely distant objects in the target distribution utilizing low-beam LiDAR. Conversely, our methodology relatively overcomes such issues compared to existing studies. Significantly, it is noteworthy that our module reliably recognizes the challenging samples of the target without regularizing data distributions.

In addition to our main paper, we validate the potential of our suggested cross-domain adaptive discriminator in Fig. 3. Here, we compare with Direct Transfer (DT). We observe that DT is usually limited to understanding the target domain. As we see in the first row (DT), the instance-level features of each two domain form clearly distinct clusters. To relieve this problem, our CDAN explicitly guides the network to generate domain-variance features and maximizes generalivability to unseen data distributions. The results of the second row, where the features of the two domains are harmoniously distributed, demonstrate that our module effectively guides the whole network.

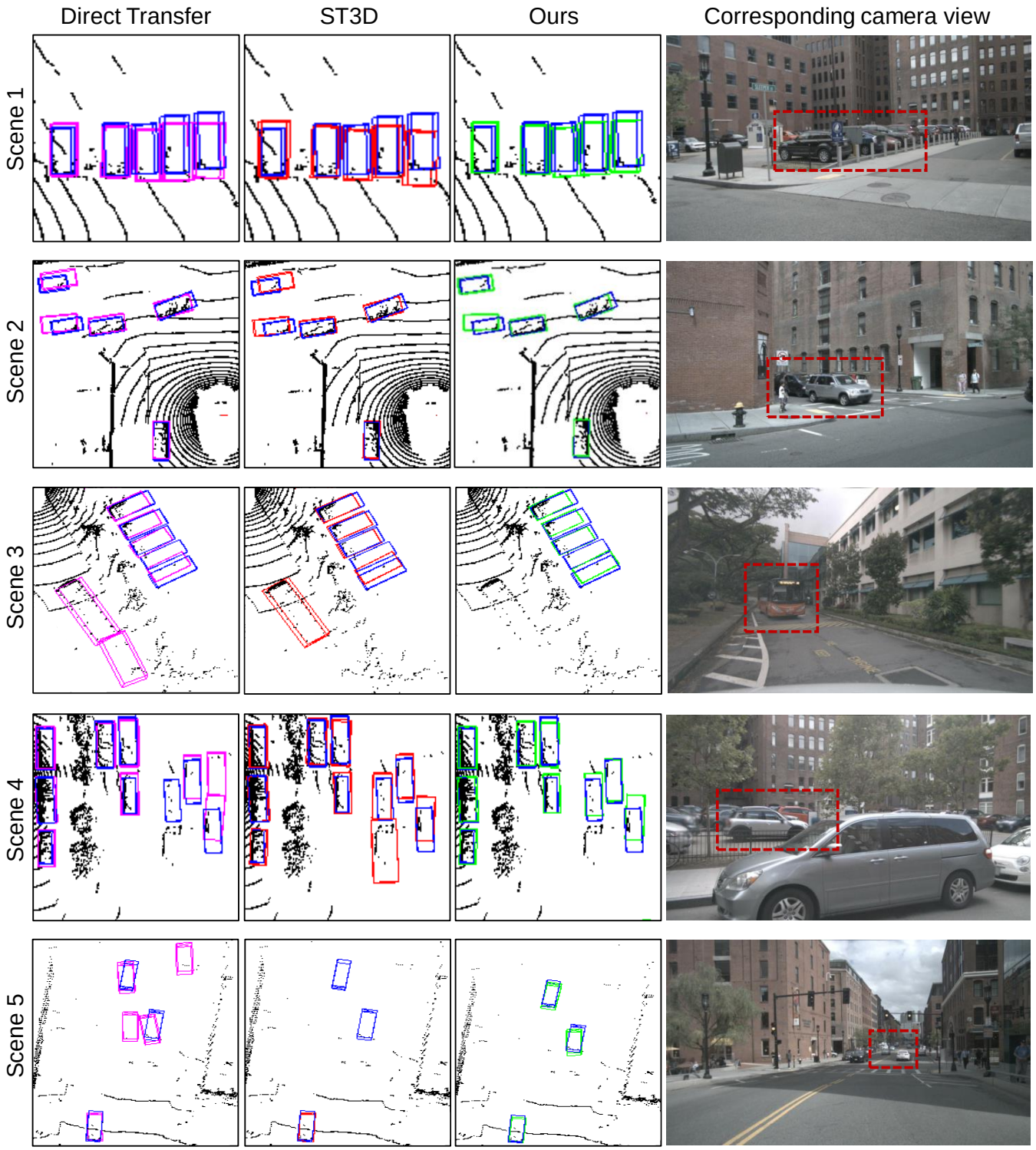

Figure 2: Qualitative visualization of subdomain shift in Waymo → nuScenes. Blue, magenta, red, and green represent ground truth, Direct Transfer, ST3D, and Ours, respectively. For better understanding, we visualize corresponding camera views along with the red dotted line showing the region where the domain shift is prominent.

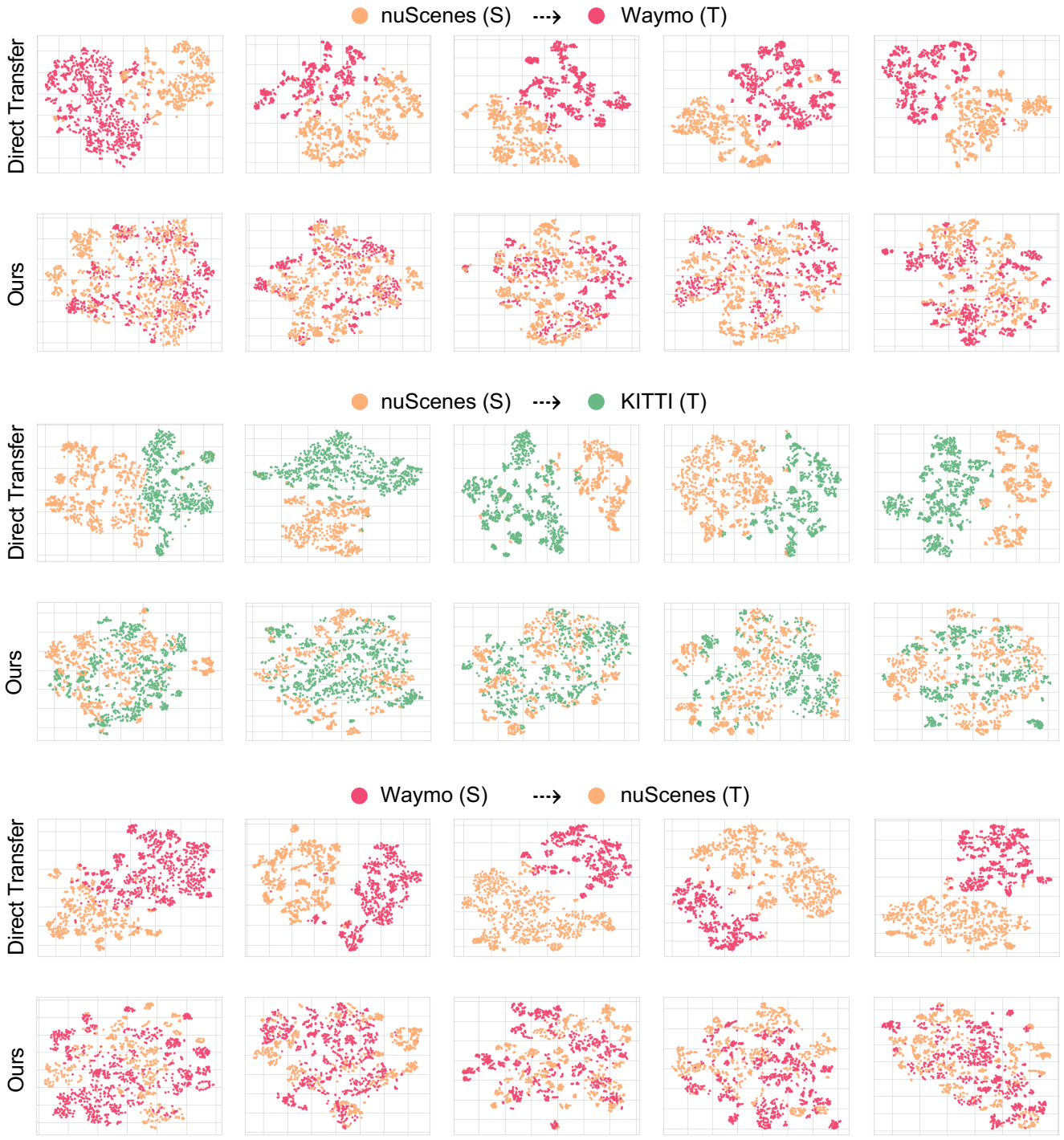

Figure 3: t-SNE (van der Maaten and Hinton 2008) visualization of source (S) and target (T) domains' features. To verify the effectiveness of the cross-domain adversarial discriminator, we compare the instance-level features of the Direct Transfer model (DT) with those of CMDA (ours). Compared to the DT, which produces distinct clusters for source and target domains, ours creates more domain-invariant feature spaces. Here, we confirm that our self-training approach with CDAN effectively overcomes the representational gap between the two domains. Best viewed in color.

References

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

Deng, J.; Shi, S.; Li, P.; Zhou, W.; Zhang, Y.; and Li, H. 2021. Voxel r-cnn: Towards high performance voxel-based 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1201–1209.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, Y.; Qi, X.; Chen, Y.; Wang, L.; Li, Z.; Sun, J.; and Jia, J. 2022. Voxel Field Fusion for 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1120–1129.

Shi, S.; Guo, C.; Jiang, L.; Wang, Z.; Shi, J.; Wang, X.; and Li, H. 2020. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10529–10538.

Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; 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.

van der Maaten, L.; and Hinton, G. 2008. Visualizing High-Dimensional Data Using t-629 SNE. *Journal of Machine Learning Research*, 9(2579-2605): 630.

Yan, Y.; Mao, Y.; and Li, B. 2018. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10): 3337.

Yang, J.; Shi, S.; Wang, Z.; Li, H.; and Qi, X. 2021. St3d: Self-training for unsupervised domain adaptation on 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10368–10378.

Yang, J.; Shi, S.; Wang, Z.; Li, H.; and Qi, X. 2022. ST3D++: Denoised Self-Training for Unsupervised Domain Adaptation on 3D Object Detection. *IEEE transactions on pattern analysis and machine intelligence*, 45(5): 6354–6371.