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

Gyusam Chang[1*], Wonseok Roh[1*], Sujin Jang[2], Dongwook Lee[2], Daehyun Ji[2], Gyeongrok Oh[1], Jinsun Park[3], Jinkyu Kim[4 †], Sangpil Kim[1†]

[1]Department of Artificial Intelligence, Korea University, Republic of Korea
[2]Samsung Advanced Institute of Technology (SAIT), Republic of Korea
[3]School of Computer Science and Engineering, Pusan National University, Republic of Korea
[4]Department of Computer Science and Engineering, Korea University, Republic of Korea
{gsjang95, paulroh, dhrudfhr98, jinkyukim, spk7}@korea.ac.kr,
{s.steve.jang, dw12.lee, derek.ji}@samsung.com, jspark@pusan.ac.kr

Abstract

Recent LiDAR-based 3D Object Detection (3DOD) methods show promising results, but they often do not generalize well to target domains outside the source (or training) data distribution. To reduce such domain gaps and thus to make 3DOD models more generalizable, we introduce a novel unsupervised domain adaptation (UDA) method, called CMDA, which (i) leverages visual semantic cues from an image modality (i.e., camera images) as an effective semantic bridge to close the domain gap in the cross-modal Bird's Eye View (BEV) representations. Further, (ii) we also introduce a self-training-based learning strategy, wherein a model is adversarially trained to generate domain-invariant features, which disrupt the discrimination of whether a feature instance comes from a source or an unseen target domain. Overall, our CMDA framework guides the 3DOD model to generate highly informative and domain-adaptive features for novel data distributions. In our extensive experiments with large-scale benchmarks, such as nuScenes, Waymo, and KITTI, those mentioned above provide significant performance gains for UDA tasks, achieving state-of-the-art performance.

Introduction

3D Object Detection (3DOD) is one of the fundamental computer vision problems and plays a crucial role in real-world applications such as autonomous driving and robotics (Qian, Lai, and Li 2022; Zhu et al. 2014). Recent studies (Shi, Wang, and Li 2019; Wang et al. 2022; Roh et al. 2022) have achieved significant advancements in 3DOD with large-scale benchmarks and precise 3D vision sensors. Especially, LiDAR-based approaches (Liang et al. 2022; Liu et al. 2023; Yin, Zhou, and Krahenbuhl 2021) have demonstrated state-of-the-art performance by leveraging precise 3D geometric information (*i.e.*, object location and size) from point clouds. However, despite these breakthroughs, most 3DOD works face significant performance drops when tested on previously unseen data distributions due to inevitable domain shift issues (*e.g.*, variations in point density, weather conditions, and geographic locations).

To address these challenges, recent approaches in LiDAR-based Unsupervised Domain Adaptation (UDA) primarily focus on effectively leveraging precise geometric information from point clouds (Wang et al. 2021) or self-training strategies with pseudo-labels (Yang et al. 2021). However, these methods face challenges in learning domain-agnostic contextual information (e.g., colors, textures, and object appearances) relying solely on geometric LiDAR features.

To supplement the absence of semantic information, we introduce Cross-Modality Knowledge Interaction (CMKI), leveraging the contextual details presented in RGB images to guide the learning of rich semantic cues in LiDAR-based geometric features. Recent studies on multi-modal fusion (Zhang, Chen, and Huang 2022; Bai et al. 2022; Liu et al. 2023) demonstrate that properly complementing 3D point clouds and 2D images with each other enhances overall detection accuracy. Due to the individualized multi-modality sensor configuration for each dataset, these methods are still limited in the UDA task. To tackle these issues, we advocate for leveraging optimal joint representation, Bird's-Eye-View (BEV), facilitating the transfer of deep semantic clues from 2D image-based features to 3D LiDAR-based features. Here, the context details of each image pixel can serve as discriminative semantic priors for improved 3DOD performance. Finally, CMKI enables producing highly informative features by softly associating multi-modal cues. We empirically found that our image-assisted approach effectively overcomes domain shift. To the best of our knowledge, we are the first to adopt the usefulness of multi-modality for UDA on 3DOD.

In addition to utilizing the fine detail of 2D images, we focus on smartly extending the standard self-training approach (Yang et al. 2021) to adapt to the previously unseen target data distributions. We propose self-training-based learning strategy with Cross-Domain Adversarial Network (CDAN) to relieve the distinct representational gap between source and target data. To ensure an explicit connection across domains, we first introduce the point cloud mix-up technique, which swaps points sector with random

*These authors contributed equally.
†Corresponding authors.

azimuth angles. Then, we further apply adversarial regularization to reduce the representational gap across domains, guiding the model to learn domain-invariant information. Besides, we design a function that minimizes independent BEV grid-wise entropy to suppress ambiguous and uncertain features derived from mixed inputs. Ultimately, our domain-adaptive adversarial self-training approach is now robust in various cross-domain scenarios.

Given landmark datasets in 3DOD, nuScenes (Caesar et al. 2020), Waymo (Sun et al. 2020), and KITTI (Geiger, Lenz, and Urtasun 2012), we validate the generalizability and effectiveness of our novel UDA framework CMDA. Above all, our proposed framework outperforms the existing state-of-the-art methods on UDA for LiDAR-based 3DOD. To summarize, our main contributions are as follows:

- We propose a novel image-assisted unsupervised domain adaptation approach, called CMDA with Cross-Modality Knowledge Interaction (CMKI) to yield highly informative features by softly associating multi-modal cues in joint BEV space. To the best of our knowledge, we are the first work that introduces leveraging the semantics of 2D images for UDA on LiDAR-based 3DOD.
- We design a practical self-training paradigm with Cross-Domain Adversarial Network (CDAN) to relieve the representational gap across domains effectively. Specifically, our approach adversarially constrains the network from learning domain-invariant cues.
- We analyze the effectiveness of our proposed method on multiple challenging benchmarks, including nuScenes, KITTI and Waymo. Extensive experiments on various cross-domain adaptation scenarios validate that our proposed method achieves new State-of-the-Art performance for UDA on 3DOD.

Related Work

LiDAR-based 3D Object Detection. In early 3D object detection tasks (Chen et al. 2017; Ku et al. 2018) which uses point clouds focus on projecting point clouds into 2D feature space by minimizing the loss of spatial information. Recent LiDAR-based 3D object detection works can be categorized as two different approaches: voxel-grid representation and point-based methods. First, the voxel-based approach (Zhou and Tuzel 2018; Yan, Mao, and Li 2018; Yang, Liang, and Urtasun 2018; Lang et al. 2019; Shi et al. 2020; Deng et al. 2021) converts point cloud data into voxel representation that is compatible with vanilla Convolution Neural Network(CNN). Also, because the point cloud is sparsely distributed over the whole image, voxel representation constructed from different point sets is more efficient. Although voxel representation is versatile and shows competitive performance in 3D object detection, loss of fine-grained information is inevitable. Differently, to handle this problem, point-based approaches (Yang et al. 2019; Shi, Wang, and Li 2019; Yang et al. 2020) directly use 3D point cloud data to leverage more accurate geometry information than previous methods. In our works, we adopt SECOND (Yan, Mao, and Li 2018) and PV-RCNN (Shi et al. 2020) as baseline models that are the representative networks in 3D object detection to demonstrate the effectiveness of extracting domain invariant features with our proposed methods.

Unsupervised Domain Adaptation (UDA) for LiDAR-based 3D Object Detection. To generalize LiDAR-based 3D Object Detection for autonomous driving, Unsupervised Domain Adaptation addresses performance drop between a labeled source dataset and an unlabeled target dataset. In early UDA for LiDAR-based 3D object detection, Y. Wang *et al.* (Wang et al. 2020) propose to mitigate the inductive bias of box scale by unfamiliar objects exploiting Statistical Normalization (SN). ST3D (Yang et al. 2021) applied Random Object Scaling (ROS) and a novel self-training framework in the data pipeline to demonstrate efficiency in the target scenario. Turning to the domain of point cloud resolution, recent studies suggest various methods to complement the sparsity of point clouds. SPG (Xu et al. 2021) enriches the missing points by employing efficient point generation. 3D-CoCo (Yihan et al. 2021) utilizes domain alignment between source and target to extract robust features from unlabeled point clouds. LiDAR Distillation (Wei et al. 2022) generates pseudo sparse point sets leveraging spherical coordinates and transfers the knowledge of the source, effectively reducing the domain gap.

Method

In this section, we present a novel Unsupervised Domain Adaptation (UDA) framework CMDA for LiDAR-based 3D object detection (3DOD). We advocate leveraging multi-modal inputs during the training phase to enhance the generalizability across diverse domains. Specifically, we encourage the LiDAR BEV features to learn rich-semantic knowledge from camera BEV features and explicitly guide such cross-modal learning via cross-domain adversarial pipeline, achieving generalized perception against unseen target conditions. We first provide an overview of our framework and then present technical details of the proposed methods as follows: (1) Cross-Modality Knowledge Interaction (CMKI) and (2) Cross-Domain Adversarial Network (CDAN).

Overview

We illustrate an overview of our framework in Alg. 1 and Fig. 1, which aims to maximize the potential of each modality in guiding the 3DOD models for improved generalizability. During the training phase, the model takes multi-view images $I = \{i_1, i_2, ..., i_{N^I}\} \in \mathbb{R}^{N^I \times H \times W \times 3}$ and 3D point clouds $P = \{p_1, p_2, ..., p_{N^P}\} \in \mathbb{R}^{N^P \times 3}$ as inputs, and outputs a set of 3D bounding boxes $\hat{L}$. Our primary goal is to effectively transfer a LiDAR-based 3DOD model trained on labeled source domain data $\{(P_i^s, I_i^s, L_i^s)\}_{i=1}^{N_s}$ to the unlabeled target domain data $\{(P_i^t, L_i^t)\}_{i=1}^{N_t}$. Here, P_i^s, I_i^s, and L_i^s represent the i-th point clouds, multi-view images, and their corresponding ground truth labels from the source domain. Similarly, P_i^t and L_i^t denote the i-th point clouds and their corresponding pseudo-label from the target domain. N_s and N_t indicate the number of samples from the source and

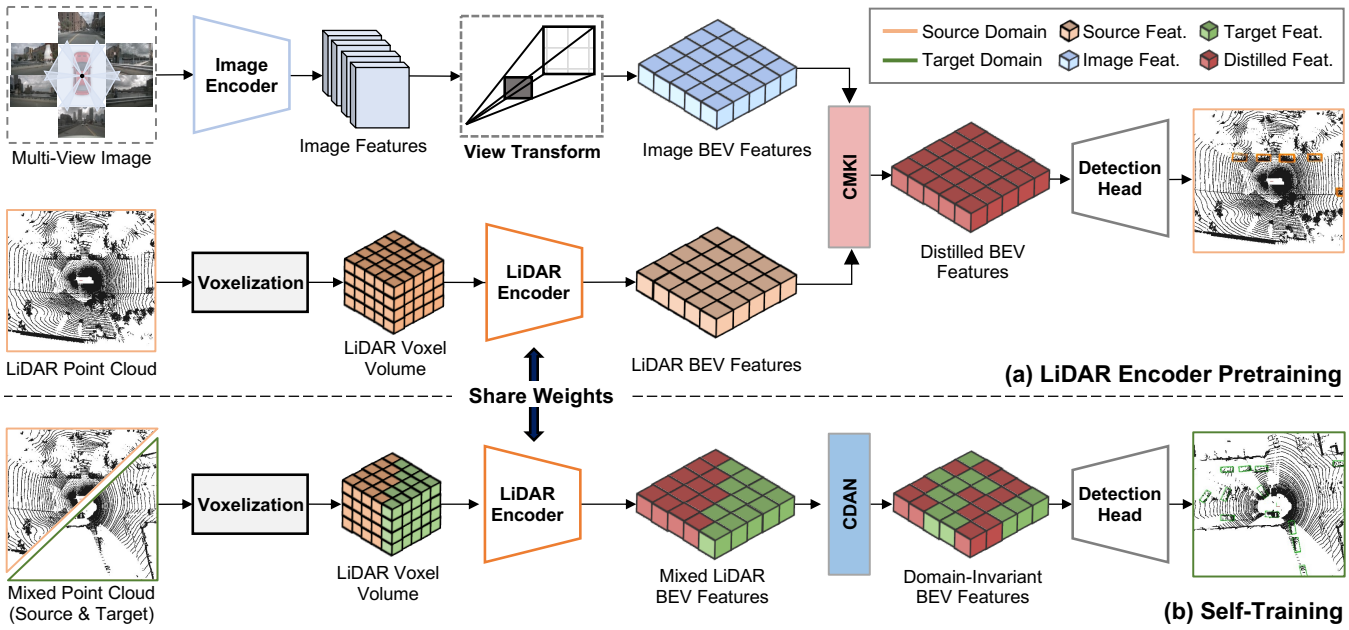

Figure 1: An overview of our architecture. Our framework consists of two main steps. (a) Cross-Modal LiDAR Encoder Pre-Training: aligning spatially paired image-based and LiDAR-based BEV representations for cross-modal BEV feature learning. This allows the LiDAR encoder to learn modality-specific visual semantic information from the image features. (b) Cross-Domain LiDAR-Only Self-Training: learning domain-invariant features through adversarial regularization of the LiDAR encoder, ultimately reducing the representation gap between source and target domains.

the target domain, respectively. Note that we do not use the target domain labels during training.

Cross-Modality Knowledge Interaction (CMKI)

Recently, multi-modal 3DOD mechanisms (Zhang, Chen, and Huang 2022; Bai et al. 2022; Vora et al. 2020) have highlighted the benefits of synergistically complementing geometric losses from 2D images with semantic losses from 3D point clouds. Despite their potential benefits, the introduction of multi-modalities to address both geometric and semantic domain gaps has received limited attention in the field of UDA for 3DOD tasks. In this work, we delve into the effectiveness of interactions between different modalities (*i.e.*, camera images and LiDAR point clouds), aiming to enhance the BEV feature quality and improve detection performance on previously unseen target data distributions.

Optimal Joint Representation. Precise geometric alignment is essential to ensure the quality of both image and point cloud features. Although existing association techniques (Chen et al. 2022; Vora et al. 2020) with calibration matrices employ multi-modal information, they do not fully take advantage of the deep semantic clues from the images due to the non-homogeneous nature of features and their representations. For example, camera features are encoded in single or multiple-perspective views, whereas LiDAR features are expressed in the BEV space (Bai et al. 2022). Hence, we are motivated to find an optimal joint representation to facilitate effective cross-modal knowledge interaction and to investigate its impact on the UDA task. Inspired by recent multi-modal fusion (Liu et al. 2023; Liang et al. 2022), we adopt BEV feature representations and aim to transfer valuable modality-specific cues between them.

Cross-Modal BEV Feature Map Generation. Inspired by the prevalent work, Lift-Splat-Shoot (LSS) (Philion and Fidler 2020), our camera stream (illustrated in Fig. 2) transforms RGB images into high-level BEV representations. First, the image encoder extracts rich-semantic visual features $F_I \in \mathbb{R}^{N^I \times H \times W \times C}$ from the multi-view images $I \in \mathbb{R}^{N^I \times H \times W \times 3}$. To construct the BEV feature, we apply a view transform module that links the 2D image coordinate to the 3D world coordinate. For each pixel, we densely predict representations at all possible depths $D_{depth} \in \mathbb{R}^{H \times W \times D}$ in a classification manner, where D denotes the discrete depth bins. We then complete the frustum-shaped voxels of contextual features by calculating the outer product of D_{depth} and F_I. Given the camera parameters, we obtain a pseudo voxel via the interpolation process, which is fed into a voxel backbone to extract features $F_I^{vox} \in \mathbb{R}^{X \times Y \times Z \times C}$. Then, F_I^{vox} are compressed along the height axis to yield the image-based BEV feature map $F_I^{bev} \in \mathbb{R}^{X \times Y \times ZC}$.

To generate the BEV feature map $F_P^{bev} \in \mathbb{R}^{X \times Y \times ZC}$ from the 3D LiDAR point clouds, we follow standard voxel-based height compression method (Zhou and Tuzel 2018).

Cross-Modal Knowledge Interaction in BEV Features. As reported by previous studies (Wang et al. 2020; Yang et al. 2021; Xu et al. 2021; Wei et al. 2022), UDA performance is significantly enhanced by leveraging precise ge-

Algorithm 1: Overview of our framework CMDA.

Input: Source labeled data $\{(P_i^s, I_i^s, L_i^s)\}_{i=1}^{N_s}$ and target pseudo-labeled data $\{(P_i^t, L_i^t)\}_{i=1}^{N_t}$.
Result: Robust 3D detector for the target domain.
Procedure:
`# LiDAR Encoder Pretraining.`
while $i = N_s$ **do**
  Transform 3D points P_i^s to BEV feature F_P^{bev}.
  Transform 2D images I_i^s to BEV feature F_I^{bev}.
  Guide F_P^{bev} to contain semantic clues from F_I^{bev}.
end
`# Self-Training.`
while $i = N_t$ **do**
  Generate pseudo label L_i^t for self-training.
  Mix source point sector with target point sector according to Eq. 2.
  Extract instance-level features $\mathbf{f}_i$ using mixed point P^{mix}.
  Generate Domain labels y_r based on location.
  Close the representational gap using adversarial discriminator ϕ_D.
end

ometric details from 3D point cloud data to guide feature-level adaptation. Although point clouds provide geometrically informative cues, they are limited in generating rich semantic information such as colors, textures, and the appearance of target objects and backgrounds. To complement such a lack of contextual information, we are motivated to exploit the fine detail of RGB images as discriminative semantic priors for improved 3DOD performance. To this end, we focus on transferring the rich semantic knowledge from image-based features to LiDAR-based features. Based on the joint BEV representations between modalities, we formulate cross-modal knowledge interaction with:

$$\mathcal{L}_{cmki} = \frac{1}{XY}\sum_{i=1}^{X}\sum_{j=1}^{Y}\left\|F_P^{bev}(i,j) - F_I^{bev}(i,j)\right\|_2, \quad (1)$$

where X, Y denotes the width and length of the BEV feature map and $\|\cdot\|_2$ is the L_2 norm. By minimizing $\mathcal{L}_{cmki}$, we optimize 3D LiDAR-based features to contain highly informative semantic clues from 2D image-based features. Our BEV-based cross-modal knowledge interaction establishes valuable connections between input modalities and consistently yields improvements across various cross-domain deployments, as demonstrated in Tables 1 and 2.

Cross-Domain Adversarial Network (CDAN)

In the field of UDA, self-training strategies (Xie et al. 2020; Yang et al. 2021; Yihan et al. 2021; Zou et al. 2018) with target pseudo-labels have significantly enhanced the performance of 3DOD models on unsupervised environments. However, we empirically discover that addressing the representational gap between source and target domains still poses challenges. Concretely, these approaches struggle to accurately recognize target objects composed of less familiar point samples and lead to generating low-quality target

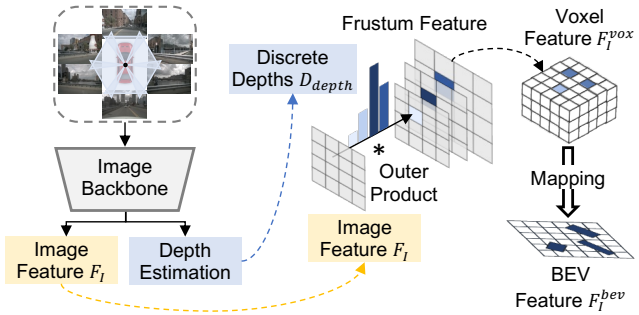

Figure 2: An overview of our Images-to-BEV View Transform module. We first transform multi-view images into voxel-wise representations F_I^{vox} by simultaneously leveraging F_I and D_{depth}, yielding a BEV representation F_I^{bev}.

pseudo-labels as illustrated in Fig. 5(a). To this end, we propose a domain-adaptive adversarial self-training approach, as shown in Fig. 3, to enhance the learning of domain-agnostic features and improve the accuracy of pseudo-labels.

Cross-Domain Mix-up. To reduce the distributional shift between domains during the self-training process, we first switch the point cloud sectors of the source and target domain scenes. We cut both point cloud sectors and corresponding labels at the identical azimuth angle and swap them with each other, following (Xiao et al. 2022). Note that the azimuth angle θ is randomly set within a specific range for each iteration to avoid inductive biases. Our mix-up process is formulated as follows:

$$(P^{mix}, L^{mix}) = M_s^{\theta}(P^s, L^s) \oplus M_t^{2\pi-\theta}(P^t, L^t), \quad (2)$$

where M_s and M_t denote binary mask to filter points within the azimuth angle θ; and L^t represents target pseudo-labels. Next, the pre-trained encoder takes the mixed point cloud P^{mix} as input and generates LiDAR-based BEV features. Our mix-up strategy directly introduces cross-domain data instances to facilitate learning domain-invariant features, ultimately leading to improved adaptation performance.

Domain Adaptive Discriminator. We introduce the domain adaptive adversarial discriminator to implicitly reduce the representational gap between the source and the target within the shared embedding space. Specifically, we guide the detection head to learn generalized information during the self-training through an adversarial learning paradigm using Gradient Reversal Layer (Ganin et al. 2016). Our cross-domain discriminator ϕ_D tries to classify the domain of instance-level features $\mathbf{f}_i$ for $i = \{1, 2, \ldots, |N_f|\}$ from the detection head. Each feature instance is labeled with y_r based on its coordinates to indicate whether it belongs to the source or target regions. We train ϕ_D to discriminate the domain of each instance based on the following loss function:

$$\mathcal{L}_d = -\mathbb{E}_{f,y_r\sim\mathbb{D}}\left[\sum_{r\in\mathcal{R}} y_r \log \phi_D(f)_r\right], \quad (3)$$

where $\mathbb{E}_{f,y_r\sim\mathbb{D}}$ indicates an expectation over samples (f, y_r) drawn from the input data distribution $\mathbb{D}$. While the discriminator ϕ_D is trained to identify the domain of each instance accurately, the 3DOD model produces instance-level

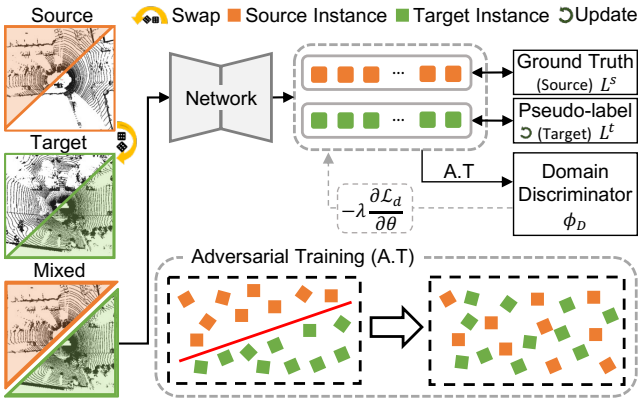

Figure 3: An overview of our cross-domain self-training step. Given a mixed point scene (source-domain points replace target-domain points in a randomly chosen region), our domain discriminator is adversarially trained to classify whether an object is from source or target domains.

features to fool the discriminator in distinguishing their domains (*i.e.*, negative loss function $-\mathcal{L}_d$). This way, we use adversarial guidance to encourage the 3DOD model to learn domain-agnostic features. Furthermore, in order to mitigate ambiguous features induced by randomly mixed domain scenes and enhance prediction confidence, we regularize the network using BEV grid-wise entropy loss $\mathcal{L}_{ent}$:

$$\mathcal{L}_{ent} = \frac{-1}{\log ZC} \sum_{i=1}^{X} \sum_{j=1}^{Y} \sum_{c=1}^{ZC} F_P^{bev}(i,j,c) \log F_P^{bev}(i,j,c) \tag{4}$$

Finally, we advance a simple self-training stage to improve generalizability across various target domains with the following loss $\mathcal{L}_{cdan}$ as a sum of the two losses $\mathcal{L}_d$ and $\mathcal{L}_{ent}$:

$$\mathcal{L}_{cdan} = \lambda_d \mathcal{L}_d + \lambda_{ent} \mathcal{L}_{ent} \tag{5}$$

Loss Function. We also leverage conventional loss term $\mathcal{L}_{det}$ associated with regression of 3D bounding box parameters and classification of object categories. Taking all loss functions together, our learning objective is:

$$\mathcal{L}_{total} = \lambda_{det}\mathcal{L}_{det} + \mathcal{T}_{cmki}\lambda_{cmki}\mathcal{L}_{cmki} + \mathcal{T}_{cdan}\lambda_{cdan}\mathcal{L}_{cdan} \tag{6}$$

where λ is a hyperparameter derived from grid searches to handle the strength of each loss term. In addition, we employ the binary toggle $\mathcal{T}$ for carefully scheduled training processes, where $(\mathcal{T}_{cmki}, \mathcal{T}_{cdan})$ set $(1, 0)$ for source training and $(0, 1)$ for self-training.

Experiments

Datasets. We evaluate overall performance on landmark datasets for 3D object detection task: nuScenes (Caesar et al. 2020), Waymo (Sun et al. 2020), and KITTI (Geiger, Lenz, and Urtasun 2012). The three datasets have different point cloud ranges and specifications. Hence, we convert them to a unified range $[-75.2, -75.2, -2, 75.2, 75.2, 4]$ and adopt only seven parameters to achieve consistent training results under the same conditions: center locations (x, y, z), box size (l, w, h), and heading angle δ.

Evaluation Metrics. We follow the KITTI evaluation metric for consistent evaluation across datasets. Also, we adopt the 360-degree surrounding view configuration for evaluation, apart from the KITTI dataset, which only offers the annotations in the front view. We report the Average Precision (AP) over 40 recall positions and 0.7 IoU thresholds for both the BEV IoUs and 3D IoUs. To offer empirical lower and upper bounds on adaptation performance, we present three additional reference points: **Direct Transfer**—evaluating the source domain pretrained model directly on the target domain, **Oracle**—the fully supervised model trained on the target domain, and **Closed Gap**—representing the hypothetical closed gap by

$$\text{Closed Gap} = \frac{\text{AP}_{\text{model}} - \text{AP}_{\text{Direct Transfer}}}{\text{AP}_{\text{Oracle}} - \text{AP}_{\text{Direct Transfer}}} \times 100\%. \tag{7}$$

Performance Comparison with SOTA Approaches. As shown in Tab. 1, we quantitatively compare our proposed framework with existing state-of-the-art methods, which include Statistical Normalization (SN) (Wang et al. 2020), ST3D (Yang et al. 2021), ST3D++ (Yang et al. 2022), LiDAR Distillation (LD) (Wei et al. 2022), and DTS (Hu, Liu, and Hu 2023). SN applies statistical normalization to reduce the inductive bias of box scales in the cross-domain setting. ST3D improves the effectiveness of the self-training process with data augmentation, while LD mitigates the beam-induced dense-to-sparse density shift by generating pseudo points. These methods demonstrate notable capacity but still rely on geometric information from 3D sensors and often face challenges in effectively adapting to unseen target domains. To overcome these limitations, we introduce CMDA, featuring Cross-Modality Knowledge Interaction (CMKI) and Cross-Domain Adversarial Network (CDAN).

In Tab. 1, we observe that our CMDA generally outperforms the other five methods in all metrics, including BEV AP, 3D AP, and Closed Gap. Following existing work, we evaluate UDA performance in three different scenarios: (1) nuScenes (Caesar et al. 2020) → Waymo (Sun et al. 2020), (2) nuScenes → KITTI (Geiger, Lenz, and Urtasun 2012), and (3) Waymo → nuScenes. The performance gain of CMDA is more apparent in scenarios (1) and (3), which utilize multi-view camera images, and thus benefit from highly instructive visual details. More importantly, in the dense to sparse subdomain shift setting, *i.e.*, Waymo → nuScenes, CMDA achieves a substantial performance improvement of *Closed Gap*, by up to +52.19%/+41.97% on SECOND-IoU (Yan, Mao, and Li 2018), and +53.41%/+28.91% on PV-RCNN (Shi et al. 2020) for BEV AP / 3D AP. These promising scores demonstrate that our framework can effectively boost 3DOD performance in the unsupervised target domain, even with fewer LiDAR sensor beams. Note that SN and LD are unsuitable for nuScenes → Waymo task and are therefore excluded for a fair comparison.

Remarkably, our CMDA framework also achieves higher adaptation scores when utilizing single-view camera images, *i.e.*, nuScenes → KITTI. In this case, CMDA with SECOND-IoU achieves +96.31% / +91.90% of *Closed Gap*. Overall, our CMDA framework effectively reduces the distributional shift between the source and target domains,

Task	Model	SECOND-IoU (Yan, Mao, and Li 2018)		PV-RCNN (Shi et al. 2020)	
		BEV AP ↑ / 3D AP ↑	Closed Gap ↑	BEV AP ↑ / 3D AP ↑	Closed Gap ↑
nuScenes → Waymo	Direct Transfer	39.18 / 20.78		41.30 / 25.89	
	ST3D (Yang et al. 2021)	45.35 / 27.12	+21.62% / +19.08%	52.50 / 36.21	+38.63% / +31.07%
	ST3D++ (Yang et al. 2022)	44.87 / 25.79	+19.94% / +15.08%	–.– / –.–	–.–% / –.–%
	CMDA (Ours)	**46.79 / 29.42**	**+26.66% / +26.00%**	**58.57 / 45.58**	**+59.57% / +59.29%**
	Oracle	67.72 / 54.01		70.29 / 59.10	
nuScenes → KITTI	Direct Transfer	51.84 / 17.92		68.15 / 37.17	
	SN (Wang et al. 2020)	40.03 / 21.23	-37.55% / +05.96%	60.48 / 49.47	-36.82% / +27.13%
	ST3D (Yang et al. 2021)	75.94 / 54.13	+76.63% / +59.50%	78.36 / 70.85	+49.02% / +74.30%
	ST3D++ (Yang et al. 2022)	80.52 / 62.37	+91.19% / +80.05%	–.– / –.–	–.–% / –.–%
	DTS (Hu, Liu, and Hu 2023)	81.40 / 66.60	+93.99% / +87.66%	83.90 / 71.80	+75.61% / +76.40%
	CMDA (Ours)	**82.13 / 68.95**	**+96.31% / +91.90%**	**84.85 / 75.02**	**+80.17% / +83.50%**
	Oracle	83.29 / 73.45		88.98 / 82.50	
Waymo → nuScenes	Direct Transfer	32.91 / 17.24		34.50 / 21.47	
	SN (Wang et al. 2020)	33.23 / 18.57	+01.69% / +07.54%	34.22 / 22.29	-01.50% / +04.80%
	ST3D (Yang et al. 2021)	35.92 / 20.19	+15.87% / +16.73%	36.42 / 22.99	+10.32% / +08.89%
	ST3D++ (Yang et al. 2022)	35.73 / 20.90	+14.87% / +20.76%	–.– / –.–	–.–% / –.–%
	LD (Wei et al. 2022)	40.66 / 22.86	+40.85% / +31.88%	43.31 / 25.63	+47.34% / +24.34%
	DTS (Hu, Liu, and Hu 2023)	41.20 / 23.00	+43.70% / +32.67%	44.00 / 26.20	+51.04% / +27.68%
	CMDA (Ours w/ LD)	**42.81 / 24.64**	**+52.19% / +41.97%**	**44.44 / 26.41**	**+53.41% / +28.91%**
	Oracle	51.88 / 34.87		53.11 / 38.56	

Table 1: Comparisons of Unsupervised Domain Adaptation (UDA) performance with state-of-the-art approaches, including SN (Wang et al. 2020), ST3D (Yang et al. 2021), ST3D++ (Yang et al. 2022), LiDAR Distillation (LD) (Wei et al. 2022) and DTS (Hu, Liu, and Hu 2023). For fair comparisons, we train our LiDAR-based object detector with two baseline methods: SECOND (Yan, Mao, and Li 2018) and PV-RCNN (Shi et al. 2020). We report UDA performance in three popular benchmarks: nuScenes (Caesar et al. 2020) → Waymo (Sun et al. 2020), nuScenes → KITTI (Geiger, Lenz, and Urtasun 2012), and Waymo → nuScenes. Evaluation metrics include moderate BEV AP and 3D AP (IoU threshold=0.7) and Closed Gap for car objects.

leading to new state-of-the-art performance.

Qualitative Analyses

t-SNE Analysis. To assess the extent of the domain gap, we provide t-SNE (van der Maaten and Hinton 2008) visualizations of the feature space learned from both the source (red) and target (blue) domains. As shown in Fig. 4, ST3D exhibits distinct clusters for the source and target domains, whereas CMDA results in a harmoniously dispersed feature space encompassing both target and source domains. These qualitative findings confirm that CMDA effectively encourages the model to learn domain-invariant features.

Impact of Utilizing Visual Semantic Priors. To validate the effectiveness of the semantic priors learned from the image-based BEV features, we present additional experimental results and qualitative analyses. In Fig. 5 (left), we perform a statistical evaluation of the perception capacity based on various point densities per object. CMDA effectively detects objects even with relatively sparse points. Fig. 5 (right) shows that CMDA achieves improved detection accuracy (mAP%), particularly for distant objects.

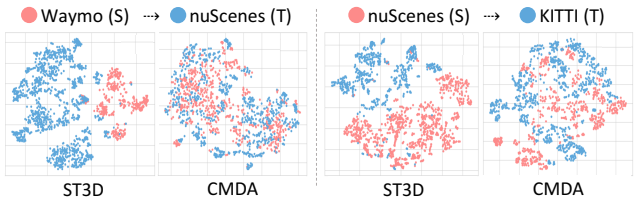

Figure 4: t-SNE (van der Maaten and Hinton 2008) visualizations of source (S, red) and target (T, blue) domains' LiDAR-based BEV feature distribution.

Further, Scene 1 of Fig. 6 provides a notable example of improved detection accuracy for distant objects. ST3D (Yang et al. 2021) fails to detect a relatively distant object, whereas CMDA successfully detects it. Also, in Scenes 1 and 2, ST3D struggles to adapt from uniform-labeled (vehicle) to various-labeled (car, truck, bus, construction vehicle, etc.) domains. In contrast, ours effectively discriminates "cars" from "construction vehicles" and

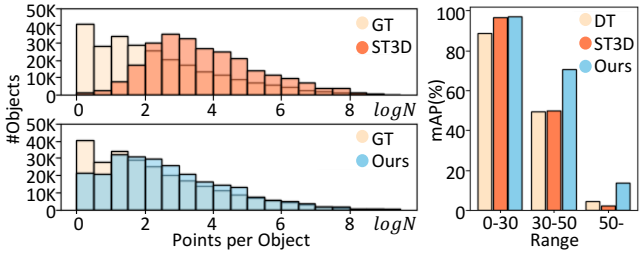

Figure 5: Statistical analyses of detection results: (left) perception capacity for the number of points per object and (right) accuracy comparison across the range.

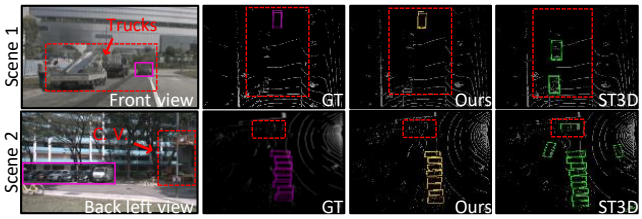

Figure 6: Qualitative visualization of Waymo → nuScenes adaptation. Magenta, green, and yellow represent Ground Truth, ST3D, and Ours. For better understanding, we visualize corresponding camera views along with the red dotted line showing the region where the domain shift is prominent.

"trucks". These findings confirm the effectiveness of utilizing the visual semantic priors jointly learned from image-based features to improve the overall UDA performance.

Ablation Studies

Effect of CDAN and CMKI. In Tab. 2, we evaluate CMKI and CDAN in various adaptation configurations on SECOND-IoU (Yan, Mao, and Li 2018) and PV-RCNN (Shi et al. 2020). ST3D (Yang et al. 2021) denotes a baseline self-training method, and the first row in each setting indicates *Direct Transfer*. To investigate the sole effect of each approach, we deliberately did not use any augmentation strategies (*e.g.*, LD (Wei et al. 2022), SN (Wang et al. 2020)). Tab. 2 demonstrates that the addition of CMKI and CDAN improves adaptive capability across all experiments. Notably, CMKI emphasizes the significance of rich semantic knowledge in achieving generalized recognition, narrowing the gaps by up to +12.02% in BEV AP and +19.30% in 3D AP, when compared to *Direct Transfer*. CDAN further enhances the generalizability by learning domain-agnostic BEV features through the adversarial discriminator, achieving improvements of up to +30.29% in BEV AP and +51.03% in 3D AP compared to *Direct Transfer*. These results prove the power of our CMDA framework in substantially enhancing the quality of UDA for 3DOD.

CDAN vs. Contrastive Learning (CL). To validate the effectiveness of CDAN, in Tab. 3, we provide a comparison with Contrastive Learning (CL)-based adaptation approach following 3D-CoCo (Yihan et al. 2021). For a fair comparison, we employ the identical source pre-trained weights and apply each learning strategy on instances from the detection head during self-training. While the CL-based method, with well-matched positive/negative pairs, enhances the baseline self-training approach (ST3D), it exhibits limited improvements compared to CDAN due to implicit issues such as sample discrepancy or precision errors. Unlike the CL-based approach, CDAN benefits significantly from adversarial mechanism and successfully tackle these challenges, producing stable adaptation effects; up to +28.29% in BEV AP / +45.75% in 3D AP compared to *Direct Transfer*.

Task	Method			SECOND-IoU	PV-RCNN
	ST3D	CMKI	CDAN	BEV AP / 3D AP	BEV AP / 3D AP
nuScenes → Waymo	-	-	-	39.18 / 20.78	41.30 / 25.89
	-	✓	-	44.41 / 22.26	47.28 / 28.57
	✓	-	-	45.35 / 27.12	52.50 / 36.21
	✓	✓	-	45.43 / 28.63	53.04 / 42.75
	✓	✓	✓	**46.79 / 29.42**	**58.57 / 45.58**
nuScenes → KITTI	-	-	-	51.84 / 17.92	68.15 / 37.17
	-	✓	-	63.86 / 37.22	72.12 / 40.17
	✓	-	-	75.94 / 54.13	78.36 / 70.85
	✓	✓	-	78.52 / 60.04	82.43 / 72.20
	✓	✓	✓	**82.13 / 68.95**	**84.85 / 75.02**

Table 2: Ablation study to see the effect of CMKI and CDAN. We report BEV AP and 3D AP (IoU=0.7) in the following domain adaption scenarios: (i) nuScenes (Caesar et al. 2020) → Waymo (Sun et al. 2020) and (ii) nuScenes → KITTI (Geiger, Lenz, and Urtasun 2012).

Task	Method			SECOND-IoU	PV-RCNN
	ST3D	CL	CDAN	BEV AP ↑ / 3D AP ↑	BEV AP ↑ / 3D AP ↑
nuScenes → KITTI	-	-	-	51.84 / 17.92	68.15 / 37.17
	✓	-	-	75.94 / 54.13	78.36 / 70.85
	✓	✓	-	79.20 / 58.20	80.99 / 71.51
	✓	-	✓	**80.13 / 63.67**	**83.27 / 73.05**

Table 3: Comparisons of domain adaptation performance with Contrastive Learning (CL) approach in nuScenes (Caesar et al. 2020) → KITTI (Geiger, Lenz, and Urtasun 2012).

Conclusion

In this work, we introduce a novel unsupervised domain adaptation approach, called CMDA, to improve the generalization power of existing LiDAR-based 3D object detection models. To reduce the gap between source and target (where its labels are not accessible during training) domains, we propose two main steps: (i) Cross-modal LiDAR Encoder Pre-training and (ii) Cross-Domain LiDAR-Only Self-Training. In (i), a pair of image-based and LiDAR-based BEV features is aligned to learn modality-agnostic (and thus more domain-invariant) features. Further, in (ii), we apply an adversarial regularization to reduce the representation gap between source and target domains. Our extensive experiments on large-scale datasets demonstrate the effectiveness of our proposed method in various cross-domain adaptation scenarios, achieving state-of-the-art UDA performance.

Acknowledgements

This work was primarily supported by Samsung Advanced Institute of Technology (SAIT) (85%) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-00079, Artificial Intelligence Graduate School Program (Korea University), 15%).

References

Bai, X.; Hu, Z.; Zhu, X.; Huang, Q.; Chen, Y.; Fu, H.; and Tai, C.-L. 2022. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1090–1099.

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.

Chen, X.; Ma, H.; Wan, J.; Li, B.; and Xia, T. 2017. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 1907–1915.

Chen, Y.; Li, Y.; Zhang, X.; Sun, J.; and Jia, J. 2022. Focal Sparse Convolutional Networks for 3D Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

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.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1): 2096–2030.

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)*.

Hu, Q.; Liu, D.; and Hu, W. 2023. Density-Insensitive Unsupervised Domain Adaption on 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17556–17566.

Ku, J.; Mozifian, M.; Lee, J.; Harakeh, A.; and Waslander, S. L. 2018. Joint 3d proposal generation and object detection from view aggregation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1–8. IEEE.

Lang, A. H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; and Beijbom, O. 2019. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12697–12705.

Liang, T.; Xie, H.; Yu, K.; Xia, Z.; Lin, Z.; Wang, Y.; Tang, T.; Wang, B.; and Tang, Z. 2022. Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in Neural Information Processing Systems*, 35: 10421–10434.

Liu, Z.; Tang, H.; Amini, A.; Yang, X.; Mao, H.; Rus, D. L.; and Han, S. 2023. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2774–2781. IEEE.

Philion, J.; and Fidler, S. 2020. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, 194–210. Springer.

Qian, R.; Lai, X.; and Li, X. 2022. 3D object detection for autonomous driving: A survey. *Pattern Recognition*, 130: 108796.

Roh, W.; Chang, G.; Moon, S.; Nam, G.; Kim, C.; Kim, Y.; Kim, S.; and Kim, J. 2022. Ora3d: Overlap region aware multi-view 3d object detection. *arXiv preprint arXiv:2207.00865*.

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.

Shi, S.; Wang, X.; and Li, H. 2019. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 770–779.

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.

Vora, S.; Lang, A. H.; Helou, B.; and Beijbom, O. 2020. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4604–4612.

Wang, C.; Ma, C.; Zhu, M.; and Yang, X. 2021. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11794–11803.

Wang, Y.; Chen, X.; You, Y.; Li, L. E.; Hariharan, B.; Campbell, M.; Weinberger, K. Q.; and Chao, W.-L. 2020. Train in germany, test in the usa: Making 3d object detectors generalize. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11713–11723.

Wang, Y.; Guizilini, V. C.; Zhang, T.; Wang, Y.; Zhao, H.; and Solomon, J. 2022. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, 180–191. PMLR.

Wei, Y.; Wei, Z.; Rao, Y.; Li, J.; Zhou, J.; and Lu, J. 2022. Lidar distillation: Bridging the beam-induced domain gap for 3d object detection. *arXiv preprint arXiv:2203.14956*.

Xiao, A.; Huang, J.; Guan, D.; Cui, K.; Lu, S.; and Shao, L. 2022. PolarMix: A General Data Augmentation Technique for LiDAR Point Clouds. *arXiv preprint arXiv:2208.00223*.

Xie, Q.; Luong, M.-T.; Hovy, E.; and Le, Q. V. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10687–10698.

Xu, Q.; Zhou, Y.; Wang, W.; Qi, C. R.; and Anguelov, D. 2021. Spg: Unsupervised domain adaptation for 3d object detection via semantic point generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15446–15456.

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

Yang, B.; Liang, M.; and Urtasun, R. 2018. Hdnet: Exploiting hd maps for 3d object detection. In *Conference on Robot Learning*, 146–155. PMLR.

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.

Yang, Z.; Sun, Y.; Liu, S.; and Jia, J. 2020. 3dssd: Point-based 3d single stage object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11040–11048.

Yang, Z.; Sun, Y.; Liu, S.; Shen, X.; and Jia, J. 2019. Std: Sparse-to-dense 3d object detector for point cloud. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1951–1960.

Yihan, Z.; Wang, C.; Wang, Y.; Xu, H.; Ye, C.; Yang, Z.; and Ma, C. 2021. Learning transferable features for point cloud detection via 3d contrastive co-training. *Advances in Neural Information Processing Systems*, 34: 21493–21504.

Yin, T.; Zhou, X.; and Krahenbuhl, P. 2021. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11784–11793.

Zhang, Y.; Chen, J.; and Huang, D. 2022. CAT-Det: Contrastively Augmented Transformer for Multi-modal 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 908–917.

Zhou, Y.; and Tuzel, O. 2018. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4490–4499.

Zhu, M.; Derpanis, K. G.; Yang, Y.; Brahmbhatt, S.; Zhang, M.; Phillips, C.; Lecce, M.; and Daniilidis, K. 2014. Single image 3D object detection and pose estimation for grasping. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 3936–3943. IEEE.

Zou, Y.; Yu, Z.; Kumar, B.; and Wang, J. 2018. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, 289–305.