

Bi3D: Bi-domain Active Learning for Cross-domain 3D Object Detection

Jiakang Yuan^{*,1}, Bo Zhang^{†,2}, Xiangchao Yan², Tao Chen^{†,1}, Botian Shi², Yikang Li², Yu Qiao²

¹School of Information Science and Technology, Fudan University

²Shanghai AI Laboratory

jkyuan22@m.fudan.edu.cn, {yanxiangchao, shibotian, liyikang, qiaoyu}@pjlab.org.cn

Abstract

Unsupervised Domain Adaptation (UDA) technique has been explored in 3D cross-domain tasks recently. Though preliminary progress has been made, the performance gap between the UDA-based 3D model and the supervised one trained with fully annotated target domain is still large. This motivates us to consider selecting partial-yet-important target data and labeling them at a minimum cost, to achieve a good trade-off between high performance and low annotation cost. To this end, we propose a Bi-domain active learning approach, namely Bi3D, to solve the cross-domain 3D object detection task. The Bi3D first develops a domainness-aware source sampling strategy, which identifies target-domain-like samples from the source domain to avoid the model being interfered by irrelevant source data. Then a diversity-based target sampling strategy is developed, which selects the most informative subset of target domain to improve the model adaptability to the target domain using as little annotation budget as possible. Experiments are conducted on typical cross-domain adaptation scenarios including cross-LiDAR-beam, cross-country, and cross-sensor, where Bi3D achieves a promising target-domain detection accuracy (89.63% on KITTI) compared with UDA-based work (84.29%), even surpassing the detector trained on the full set of the labeled target domain (88.98%). Our code is available at: <https://github.com/PJLab-ADG/3DTrans>.

1. Introduction

LiDAR-based 3D Object Detection (3DOD) [5, 13, 26, 28, 41] has advanced a lot recently. However, the generalization of a well-trained 3DOD model from a source point cloud dataset (domain) to another one, namely cross-

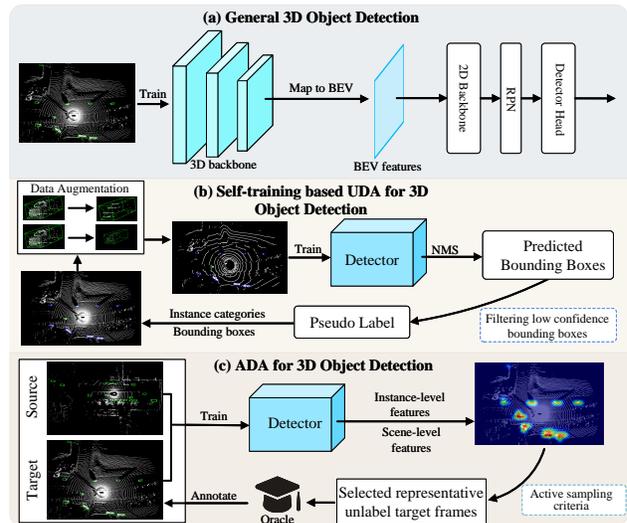


Figure 1. Comparisons among (a) The general 3DOD pipeline, (b) Self-training based Unsupervised Domain Adaptation 3DOD pipeline, and (c) Active Domain Adaptation 3DOD pipeline that selects representative target data, and then annotates them by an oracle (human expert) for subsequent model refinement.

domain 3DOD, is still under-explored. Such a task in fact is important in many real-world applications. For example, in the autonomous driving scenario, the target scene distribution frequently changes due to unforeseen differences in dynamically changing environments, making cross-domain 3DOD an urgent problem to be resolved.

Benefiting from the success of Unsupervised Domain Adaptation (UDA) technique in 2D cross-domain tasks [3, 7, 10, 14, 32, 46, 49], several attempts are made to apply UDA for tackling 3D cross-domain tasks [15, 20, 22, 37, 40, 43, 47]. ST3D [43] designs a self-training-based framework to adapt a pre-trained detector from the source domain to a new target domain. LiDAR distillation [37] exploits transferable knowledge learned from high-beam LiDAR data to the low one. Although these UDA 3D models have achieved significant performance gains for the cross-domain task, there is still a large performance gap between these UDA models

*This work was done when Jiakang Yuan was an intern at Shanghai AI Laboratory.

[†]Corresponding to: Tao Chen (eetchen@fudan.edu.cn), Bo Zhang (zhangbo@pjlab.org.cn)

and the supervised ones trained using a fully-annotated target domain. For example, ST3D [43] only achieves 72.94% AP_{3D} in nuScenes [1]-to-KITTI [8] cross-domain setting, yet the fully-supervised result using the same baseline detector can reach to 82.50% AP_{3D} on KITTI.

To further reduce the detection performance gap between UDA-based 3D models and the fully-supervised ones, an initial attempt is to leverage Active Domain Adaptation (ADA) technique [6, 17, 29, 38, 39], whose goal is to select a subset quota of all unlabeled samples from the target domain to perform the manual annotation for model training. Actually, the ADA task has been explored in 2D vision fields such as AADA [29], TQS [6], and CLUE [17], but its research on 3D point cloud data still remains blank. In order to verify the versatility of 2D image-based ADA methods towards 3D point cloud, we conduct extensive attempts by integrating the recently proposed ADA methods, *e.g.*, TQS [6] and CLUE [17], into several typical 3D baseline detectors, *e.g.*, PV-RCNN [26] and Voxel R-CNN [5]. Results show that these 2D ADA methods cannot obtain satisfactory detection accuracy under the 3D scene’s domain discrepancies. For example, PV-RCNN coupled with TQS only achieves 75.40% AP_{3D} , which largely falls behind the fully-supervised result 82.50% AP_{3D} .

As a result, directly selecting a subset of given 3D frames using 2D ADA methods to tackle 3D scene’s domain discrepancies is challenging, which can be attributed to the following reasons. (1) The **sparsity** of the 3D point clouds leads to huge inter-domain discrepancies that harm the discriminability of domain-related features. (2) The **intra-domain feature variations** are widespread within the source domain, which enlarges the differentiation between the selected target domain samples and the entire source domain samples, bringing negative transfer to the model adaptation on the target domain.

To this end, we propose a Bi-domain active learning (Bi3D) framework to conduct the active learning for the 3D point clouds. To tackle the problem of **sparsity**, we design a foreground region-aware discriminator, which exploits an RPN-based attention enhancement to derive a foreground-related domainness metric, that can be regarded as an important proxy for active sampling strategy. To address the problem of **intra-domain feature variations** within the source domain, we conceive a Bi-domain sampling approach, where Bi-domain means that data from both source and target domains are picked up for safe and robust model adaptation. Specifically, the Bi3D is composed of a domainness-aware source sampling strategy and a diversity-based target sampling strategy. The source sampling strategy aims to select target-domain-like samples from the source domain, by judging the corresponding domainness score of each given source sample. Then, the target sampling strategy is utilized to select diverse but representative

data from the target domain by dynamically maintaining a similarity bank. Finally, we employ the sampled data from both domains to adapt the source pre-trained detector on a new target domain at a low annotation cost.

The main contributions can be summarized as follows:

1. From a new perspective of chasing high performance at a low cost, we explore the possibilities of leveraging active learning to achieve effective 3D scene-level cross-domain object detection.
2. A Bi-domain active sampling approach is proposed, consisting of a domainness-aware source sampling strategy and a diversity-based target sampling strategy to identify the most informative samples from both source and target domains, boosting the model’s adaptation performance.
3. Experiments show that Bi3D outperforms state-of-the-art UDA works with only 1% target annotation budget for cross-domain 3DOD. Moreover, Bi3D achieves 89.63% AP_{BEV} in the nuScenes-to-KITTI scenario, surpassing the fully supervised result (88.98% AP_{BEV}) on the KITTI dataset.

2. Related Works

2.1. LiDAR-based General and UDA 3D Detection

LiDAR-based 3D object detection [2, 5, 13, 18, 26–28, 41, 42, 44, 48] has attracted increasing attention in real applications such as autonomous driving and robotics. Grid-based methods [5, 41] convert disordered point cloud data to regular grids and extract features by 2D/3D convolution. Inspired by PointNet [19], Point-based approaches [27, 44] use set abstraction to extract features and directly generate proposals from point cloud data. However, these general 3D detectors still face serious performance drops in cross-domain applications, *e.g.*, from Waymo or nuScenes to KITTI adaptation scenarios. UDA 3D object detection tackles the cross-domain distribution shift issue by various unsupervised methods. ST3D [43] proposes to use self-training and curriculum data augmentation to generate pseudo labels on a target domain to mitigate the large domain gap. LiDAR Distillation [37] proposes a distillation-based method, focusing on the knowledge transfer from high-beam data to low-beam data. However, there is still a large detection accuracy gap between these UDA methods [37, 43] and fully-supervised 3D detectors [26, 28, 41].

2.2. Active Domain Adaptation

Inspired by active learning methods [4, 11, 23–25, 34, 35, 45] which aim to achieve relatively high recognition accuracies only using a small portion of informative data, Active Domain Adaptation (ADA) [6, 16, 17, 29, 38] has emerged in 2D vision task, which selects the most informative target data for annotation and adapts the model to the target

domain by training on the selected data. CLUE [17] proposes to use an uncertainty-weighted clustering strategy to select informative target data. TQS [6] utilizes a hierarchical sampling strategy that performs active learning from multi-grained criteria such as transferable committee, transferable uncertainty, and transferable domainness.

Although the ADA technique has achieved great success in 2D image tasks, its exploration of 3D point cloud tasks is still insufficient. Furthermore, it is intractable to directly apply these 2D ADA methods to the 3D point cloud scenarios, since these 2D ADA works [6, 16, 17, 29] are not intended to tackle the distribution difference of point clouds with various spatial and geometric structures. Besides, previous ADA methods focus more on how to select samples from the target domain, ignoring that the source domain may contain many diverse samples and not all of them are beneficial for model adaptation to the target domain. In contrast, our Bi3D provides a new angle of view for achieving cross-domain generalization: a Bi-domain active learning strategy, which samples informative frames from both source and target domains.

3. Method

The overall Bi3D framework is shown in Fig. 2. To better illustrate the Bi3D principle, we first describe our problem definition and the selected baseline model. Next, we introduce the proposed Bi3D. Finally, we give the overall objectives and Bi-domain sampling and training strategies.

3.1. Preliminary

Problem Definition. Given a labeled source domain set $D_s = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{n_s}$, an unlabeled target domain set $D_t = \{\mathbf{x}_j^t\}_{j=1}^{n_t}$, and an annotation budget B , where $B \ll n_t$ and n_t denotes the total amount of target domain data. Following the standard ADA setting, a labeled target dataset \tilde{D}_t is constructed, which is initially empty and will be updated in R rounds of the sampling process. In the k -th sampling round where $k \leq R$, a subset ΔD_t^k is selected from D_t/\tilde{D}_t and labeled by an oracle (**human expert**). Then, \tilde{D}_t will be updated as $\tilde{D}_t \leftarrow \tilde{D}_t \cup \Delta D_t^k$. After R rounds of sampling, the number of data in \tilde{D}_t reaches the upper limit of annotation budget B , i.e., $|\tilde{D}_t| = B$. Note that different from previous ADA methods, in this work, we further construct a source subset \tilde{D}_s sampled from the original source domain D_s . The goal of the proposed Bi3D is to select both target-domain-like data from D_s and the most informative data from D_t , to constitute \tilde{D}_s and \tilde{D}_t , and make the 3D detector better adapt to target domain by jointly training on a mixture set from \tilde{D}_s and \tilde{D}_t .

Baseline Introduction. Following previous cross-domain studies [37, 43] for 3DOD, we use PV-RCNN [26] as our baseline model. PV-RCNN is a typical two-stage 3D detection framework that takes advantage of both the point-based network and 3D voxel-based CNN. The overall loss

function of PV-RCNN can be written as follows:

$$L_{det} = L_{rpn} + L_{rcnn} + L_{seg} \quad (1)$$

where L_{rpn} denotes the loss of Region Proposal Network (RPN), L_{rcnn} represents the proposal refinement loss and L_{seg} is the keypoint segmentation loss.

3.2. Bi3D: Bi-domain Active Learning for 3D Object Detection

To effectively measure the domainness of source and target samples, we first design a foreground region-aware discriminator. Then, based on the domain discriminator, we propose a Bi-domain sampling strategy to adapt a pre-trained 3D detector from its source domain to a new target domain.

Foreground Region-aware Discriminator. Considering that instance-level features lose the contextual relationship between the instance and its original scene, and meanwhile, a large number of negative anchors will greatly hinder domain discriminator learning, we thus generate scene-level representations by extracting from Bird-Eye-View (BEV) features. However, the BEV features extracted using 3D convolution are very sparse due to the sparse distribution of point cloud data, causing the traditional discriminator difficult to localize and learn on informative foreground regions, thus resulting in a biased domain representation learning.

To address this issue, we design a foreground region-aware discriminator, aiming at measuring the frame-level domainness score for both the source and target data by enhancing foreground-region features in the scene. Specifically, let \mathbf{x}_d^d denote the input point cloud data, where $d \in [s, t]$ means that the sample \mathbf{x} is from source domain s or target domain t . Next, the 3D feature volumes are first encoded by the 3D backbone F_{3D} and then converted into 2D BEV features $f_{bev} \in \mathbb{R}^{C \times H \times W}$, where C denotes the channel number, H and W are the height and width of the feature, respectively.

To make the domain discriminator pay more attention to foreground regions, we first obtain the objectness score $S_{obj} \in \mathbb{R}^{C' \times H \times W}$ by the RPN operation, where C' indicates the number of anchors per location. The objectness score represents the probability that a default anchor belongs to a foreground object. In order to quantitatively evaluate the prediction uncertainty of the detector for the current scene, inspired by previous methods [6, 9, 21] using entropy to measure uncertainty, we calculate the entropy score $S_{ent} \in \mathbb{R}^{C' \times H \times W}$ with the following formula:

$$S_{ent} = -S_{obj} \log S_{obj} - (1 - S_{obj}) \log(1 - S_{obj}), \quad (2)$$

where S_{ent} denotes the uncertainty of a spatial location being classified as an instance object. Based on Eq. 2, a scene-level attention map can be obtained by combining S_{obj} and

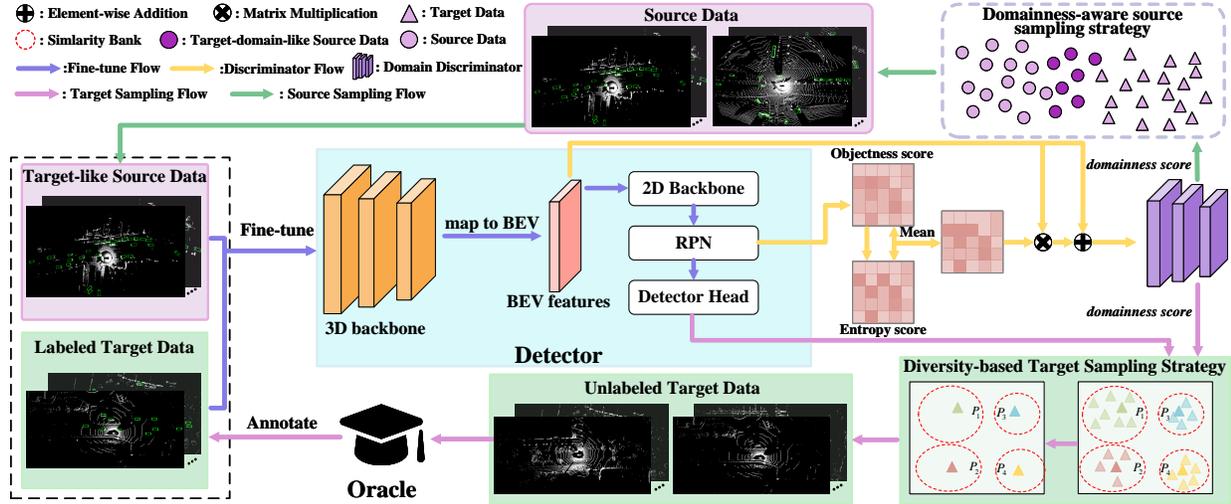


Figure 2. The overview of the proposed Bi3D, which employs PV-RCNN as our baseline and consists of domainness-aware source sampling strategy and diversity-based target sampling strategy. The target-domain-like source data are first selected by the learned domainness score, and then the detector is fine-tuned on the selected source domain data. Next, diverse and representative target data are selected using a similarity bank, and then annotated by an oracle. Finally, the detector is fine-tuned on both the selected source and target data.

S_{ent} , which can make the model pay more attention to foreground features. Thus, the foreground region-aware features can be calculated as follows:

$$\hat{f}_{bev} = (1 + (\hat{S}_{obj} + \hat{S}_{ent})/2) f_{bev}, \quad (3)$$

where \hat{f}_{bev} represents the foreground region-aware BEV features, and \hat{S}_{obj} and \hat{S}_{ent} are the maximum value of S_{obj} and S_{ent} along the channel dimension, respectively.

Based on the foreground region-aware feature maps \hat{f}_{bev} , a domain discriminator with a convolution block is utilized to classify whether the data is from the source domain or target domain. For a detailed structure of the discriminator, please refer to our supplementary material. The loss function of the domain discriminator can be written as follows:

$$L_{dom} = -\mathbb{E}_{x^s \sim D_s} [\log(1 - H(\hat{f}_{bev}^s))] - \mathbb{E}_{x^t \sim D_t} [\log(H(\hat{f}_{bev}^t))], \quad (4)$$

where L_{dom} is the domainness loss, and H denotes the domain discriminator, where we label the source domain and the target domain as '0' and '1', respectively.

Domainness-aware Source Sampling Strategy. Previous DA works mainly focus on how to fully exploit representative data from the target domain, which actually ignores that there are a certain number of source samples interfering with the target domain representation learning. Thus, we propose a simple but effective domainness-aware source sampling strategy, aiming at selecting target-domain-like samples from the source domain to initially strengthen the model adaptability. In particular, we first calculate scene-level domainness score s_i^s of all source data using the aforementioned domain discriminator, where $s_i^s = H(\hat{f}_{bev}^s)$ and

s_i^s can be regarded as a similarity metric between source data and target data. s_i^s with a relatively high value indicates that the i -th frame data from the source domain complies with the data distribution of the target domain. To select the source data with a high domainness score, we **simply sort s_i^s in descending order**, thus \tilde{D}_s can be built by sampling the sorted data with a proportion or threshold. Please refer to our supplementary material for the study of the number of selected source data. Note that there is a smaller domain gap between \tilde{D}_s and D_t and therefore by fine-tuning the detector on \tilde{D}_s , the performance of the model on the target domain will be improved. As a result, the detector can extract more accurate instance-level features, benefiting to select more informative target data.

Diversity-based Target Sampling Strategy. To make the detector better adapt to the target domain, we first fine-tune the detector on \tilde{D}_s , and select representative data from the target domain. However, since the adjacent frames in scenes like autonomous driving are usually similar, traditional active learning methods (*e.g.*, Query-by-Committee [23], Query-by-Uncertainty [34]) encounter a major challenge that they often select the samples with a small between-class difference, causing redundant frame annotation operations. Thus, we design a diversity-based target sampling strategy to select diverse-and-representative target domain data.

Given ROI features $\mathbf{I}_j = [I_j^1, I_j^2, \dots, I_j^k]$ from the j -th target frame, the corresponding confidence scores $\mathbf{d}_j = [d_j^1, d_j^2, \dots, d_j^k]$ can be easily obtained by the baseline detector with the standard post-processing process, *i.e.*, Non-Maximum Suppression (NMS). We first use confidence scores to re-weight all ROI instance-level features to obtain more accurate instance descriptions \hat{I}_j in the current

Algorithm 1 Diversity-based Target Sampling Strategy

Input: The j -th unlabeled frame x_j^t , where $x_j^t \in D_t/\tilde{D}_t$, and the k -th round of sampling budget b_k

Output: The selected target set ΔD_t^k

- 1: Calculate the re-weighted ROI features \hat{I}_j with the obtained domainness score s_j^t from the unlabeled frame x_j^t , by $s_j^t = H(\hat{f}_{bev}^t)$.
 - 2: Initialize the similarity bank $\mathbf{P} := \emptyset$ and budget prototypes $\mathbf{c} := \emptyset$
 - 3: **for** x_j^t in D_t/\tilde{D}_t **do**
 - 4: **if** $|\Delta D_t^k| < b_k$ **then**
 - 5: Update $\mathbf{P} := \mathbf{P} \cup x_j^t$, $\mathbf{c} := \mathbf{c} \cup \hat{I}_j$
 - 6: **else**
 - 7: Calculate the similarity $\alpha_{\hat{I}_j, \mathbf{c}}$ between \hat{I}_j and \mathbf{c}
 - 8: Calculate the similarity $\alpha_{\mathbf{c}}$ of prototypes in \mathbf{c}
 - 9: **if** $\max(\alpha_{\hat{I}_j, \mathbf{c}}) < \min(\alpha_{\mathbf{c}})$ **then**
 - 10: Merge the most similar banks P_m and P_n
 - 11: and the corresponding prototypes c_m and c_n
 - 12: using Eq. 5
 - 13: Update $\mathbf{P} := \mathbf{P} \cup x_j^t$, $\mathbf{c} := \mathbf{c} \cup \hat{I}_j$
 - 14: **else**
 - 15: Merge \hat{I}_j into P_m , where the corresponding
 - 16: prototype c_m is most similar to \hat{I}_j
 - 17: Select data in each bank by s_j^t and fill the ΔD_t^k
 - 18: **return** Selected target subset ΔD_t^k
-

frame x_j^t , where $\hat{I}_j = \mathbf{I}_j^T \mathbf{d}_j$, and the domainness score of target domain s_j^t can be calculated by the designed domain discriminator H described above. As summarized in Algorithm 1, the basic idea of the diversity-based target sampling strategy is to maintain a similarity bank, where all unlabeled target data are clustered based on pairwise similarity of re-weighted ROI features to ensure the diversity of selected target data. In particular, we use cosine distance to measure the similarity α and dynamically update the prototypes of candidate ROI features using the following formula:

$$\hat{c}(P_m, P_n) = \frac{\text{num}(P_m) \times c_m + \text{num}(P_n) \times c_n}{\text{num}(P_m) + \text{num}(P_n)}, \quad (5)$$

where c_m , c_n are the m -th and n -th prototypes assigned according to the preset budget, meaning that each budget is represented by one prototype. P_m and P_n denote the similarity bank of the above m -th and n -th budget-wise prototypes, which are used to buffer unlabeled frames, and $\text{num}(\cdot)$ denotes the number of unlabeled frames in the buffer. After Algorithm 1 is finished, to sample more diverse and representative frames from the target domain, we select one unlabeled frame x_j^t with the top-1 domainness score s_j^t from each updated bank \mathbf{P} , to form the full set of all data for manual annotation.

3.3. Overall Objective and Bi-domain Sampling and Training Strategy

Overall Objectives. The overall objective can be formulated as follows:

$$L_{det} = \mathbb{E}_{x \sim \tilde{D}_s \cup \tilde{D}_t} [L_{rpn} + L_{rcnn} + L_{seg}], \quad (6)$$

where the definition of L_{rpn} , L_{rcnn} , L_{seg} follows Eq. 1.

Bi-domain Sampling and Training Strategy. To adapt the detector from the source domain to the target domain, our method includes four steps. 1) *Pre-training on source domain:* The detector is firstly pre-trained on D_s using Eq. 1 to ensure that the detector can learn sufficient knowledge for model transfer. 2) *Training the domain discriminator:* We freeze the parameters from the baseline detector while training the designed domain discriminator using L_{dom} in Eq. 4. 3) *Active sampling source domain:* In this step, we select target-domain-like source data and fine-tune the detector on \tilde{D}_s to reduce the domain gap. 4) *Active sampling target domain:* Based on the selected source data and the fine-tuned detector, we further sample the most informative target data and re-train the detector on both \tilde{D}_s and \tilde{D}_t .

4. Experiments

4.1. Experimental Setup

Datasets. We conduct experiments on four popular autonomous driving datasets: KITTI [8], Waymo [30], nuScenes [1] and Lyft [12]. We consider four cross-domain settings including cross-LiDAR-beam (*i.e.* Waymo-to-nuScenes, nuScenes-to-KITTI), cross-country (*i.e.* Waymo-to-KITTI), and cross-sensor scenarios (*i.e.* Waymo-to-Lyft). Following previous domain adaptation works [37, 43], we use the KITTI evaluation metric to perform all experiments on Car (Vehicle in Waymo) category.

Implementation Details. We evaluate the proposed Bi3D on two widely-used detectors: PV-RCNN [26] and Voxel R-CNN [5]. Following [37, 43], we only use the coordinate encoding (x, y, z) of raw point cloud as the detector input, and set the voxel size of both PV-RCNN and Voxel R-CNN to $(0.1m, 0.1m, 0.15m)$ on all datasets. **In the stage of active sampling source domain,** we first select the target-domain-like data from the source domain in the initial training epoch and fine-tune the detector for the following 15 epochs. **In the stage of active sampling target domain,** we mainly consider the situation that the annotation budget B is equal to 1% and 5%, respectively, which follows the standard experimental setting in the ADA task. Our method is implemented using OpenPCDet [31].

4.2. Comparison Baselines

To verify the effectiveness of the proposed Bi3D, we design several baseline methods including both active learning and active domain adaptation based methods.

Task	Method	PV-RCNN		Voxel R-CNN	
		AP _{BEV} / AP _{3D}	Closed Gap	AP _{BEV} / AP _{3D}	Closed Gap
Waymo→KITTI	Source Only	61.18 / 22.01	-	64.87 / 19.90	-
	ST3D [43]	84.10 / 64.78	+82.45% / +70.71%	65.67 / 20.14	+03.26% / +00.38%
	Ours (1%)	85.13 / 71.36	+86.15% / +81.58%	86.35 / 72.70	+87.42% / +83.36%
	SN [36]	79.78 / 63.60	+66.91% / +68.76%	71.65 / 61.63	+27.55% / +65.88%
	ST3D (w/ SN) [43]	86.65 / 76.86	+91.62% / +90.68%	80.23 / 68.98	+62.52% / +77.49%
	CLUE (w/ SN, 1%) [17]	82.13 / 73.14	+75.36% / +84.53%	81.93 / 70.89	+69.43% / +80.50%
	TQS (w/ SN, 1%) [6]	82.00 / 72.04	+74.89% / +82.77%	78.26 / 67.11	+54.50% / +74.53%
	Ours (w/ SN, 1%)	87.12 / 78.03	+93.31% / +92.61%	88.09 / 79.14	+94.51% / +93.53%
	Ours (w/ SN, 5%)	89.53 / 81.32	+102.64% / +97.39%	90.18 / 81.34	+103.01% / +97.00%
	Oracle	88.98 / 82.50	-	89.44 / 83.24	-
Waymo→Lyft	Source Only	75.49 / 58.53	-	70.52 / 53.48	-
	ST3D [43]	77.68 / 60.53	+19.96% / +15.20%	72.27 / 54.94	+15.97% / +21.22%
	Ours (1%)	79.06 / 63.70	+32.54% / +39.29%	78.39 / 64.50	+71.81% / +160.17%
	SN [36]	72.82 / 56.64	-24.34% / -14.36%	68.77 / 52.67	-15.97% / -11.77%
	ST3D (w/ SN) [43]	74.95 / 58.54	-04.92% / +00.08%	69.91 / 54.23	-05.57% / +10.90%
	CLUE (w/ SN, 1%) [17]	75.23 / 62.17	-02.37% / +27.66%	75.61 / 59.34	+46.44% / +85.17%
	TQS (w/ SN, 1%) [6]	70.87 / 55.25	-42.11% / -24.92%	71.11 / 56.28	+05.38% / +40.70%
	Ours (w/ SN, 1%)	79.07 / 63.74	+32.63% / +39.59%	77.00 / 61.23	+59.12% / +112.65%
	Ours (w/ SN, 5%)	80.12 / 65.54	+42.21% / +53.27%	79.15 / 65.26	+78.74% / +171.22%
	Oracle	86.46 / 71.69	-	81.48 / 60.36	-
Waymo→nuScenes	Source Only	34.50 / 21.47	-	32.58 / 16.53	-
	ST3D [43]	36.42 / 22.99	+10.32% / +08.89%	34.68 / 17.17	+12.40% / +03.33%
	LiDAR Distill [37]	43.31 / 25.63	+47.34% / +24.34%	-	-
	Ours (1%)	45.52 / 30.75	+59.22% / +54.30%	44.86 / 29.52	+72.45% / +67.63%
	SN [36]	34.22 / 22.29	-01.50% / +04.80%	29.43 / 19.21	-18.60% / +13.95%
	ST3D (w/ SN) [43]	36.62 / 23.67	+11.39% / +12.87%	32.77 / 22.21	+01.12% / +29.57%
	CLUE (w/ SN, 1%) [17]	38.18 / 26.96	+19.77% / +32.12%	37.27 / 25.12	+39.49% / +44.72%
	TQS (w/ SN, 1%) [6]	35.47 / 25.00	+05.01% / +20.66%	36.38 / 24.18	+22.43% / +39.82%
	Ours (w/ SN, 1%)	45.00 / 30.81	+56.42% / +54.65%	45.29 / 29.70	+75.03% / +68.56%
	Ours (w/ SN, 5%)	48.03 / 32.02	+72.70% / +61.73%	47.02 / 31.23	+85.24% / +76.52%
Oracle	53.11 / 38.56	-	49.52 / 35.74	-	
nuScenes→KITTI	Source Only	68.15 / 37.17	-	67.27 / 30.54	-
	ST3D [43]	78.36 / 70.85	+49.02% / +74.30%	74.16 / 35.55	+31.08% / +09.51%
	Ours (1%)	84.91 / 71.56	+80.64% / +75.87%	86.10 / 72.75	+84.93% / +80.08%
	SN [36]	60.48 / 49.47	-36.82% / +27.13%	44.00 / 25.20	-104.96% / -10.13%
	ST3D (w/ SN) [43]	84.29 / 72.94	+77.48% / +78.91%	52.44 / 20.99	-66.89% / -18.12%
	CLUE (w/ SN, 1%) [17]	74.77 / 64.43	+37.18% / +60.14%	79.12 / 68.02	+53.45% / +71.12%
	TQS (w/ SN, 1%) [6]	84.66 / 75.40	+79.26% / +84.34%	77.98 / 66.02	+48.31% / +67.32%
	Ours (w/ SN, 1%)	87.00 / 77.55	+90.49% / +89.08%	87.33 / 77.24	+90.48% / +88.61%
	Ours (w/ SN, 5%)	89.63 / 81.02	+103.12% / +96.73%	88.15 / 79.06	+94.18% / +92.07%
	Oracle	88.98 / 82.50	-	89.44 / 83.24	-

Table 1. Results on different adaptation scenarios under 1% and 5% annotation budget. Following [37, 43], we report AP_{BEV} and AP_{3D} over 40 positions’ recall for the car category at IoU = 0.7. **Source Only** denotes that the pre-trained detector is directly evaluated on the target domain, and **Oracle** represents the detection results obtained using the fully-annotated target domain. Closed Gap denotes the performance gap closed by various methods along Source Only and Oracle results. The best adaptation results are marked in **bold**.

1) Random: We randomly select the target domain data for performing the manual annotation.

2) Entropy [34]: By measuring the entropy of samples from the target domain, we select the samples with relatively high entropy scores, which can represent the sample-level uncertainty predicted by a detector.

3) Committee [23]: By using multiple classifiers to predict the categories of target samples, the samples with inconsistent prediction scores along with all classifiers are selected.

4) CLUE [17]: CLUE is a representative work under ADA setting, which proposes Clustering Uncertainty-weighted Embeddings in order to select informative-and-diverse tar-

get data by means of a re-weighted uncertainty clustering.

5) TQS [6]: TQS is a prior work to explore the transferable criteria which are specially designed to mitigate the domain gap. TQS picks up data by combining a series of transferable sampling strategies (such as committee, uncertainty, and domainness) to reduce the sampling uncertainty.

4.3. Main Results

Comparison with 2D ADA works. To verify the effectiveness of our Bi3D and ensure the fairness of experiments, we first compare our method with two widely used 2D ADA methods (*i.e.* CLUE and TQS) under the same cross-domain setting. As shown in Table 1, it can be seen that compared with 2D ADA methods, our Bi3D achieves better results by a large margin on all cross-domain scenarios, demonstrating the method’s scalability for 3D point cloud detection tasks. Meanwhile, we can observe that 2D ADA methods cannot achieve satisfactory results, and even fall behind UDA methods (*i.e.* Waymo→KITTI on PV-RCNN). A detailed analysis is described in Section 4.4.

Comparison with 3D UDA works. We deeply review the cross-domain 3D object detection works [37, 43], and find that previous works mainly focus on the study of UDA 3D detection. To show the effectiveness of active learning, we compare our Bi3D with these cross-domain 3D detection works. For example, ST3D [43] uses self-training to iteratively improve the performance on the target domain, LiDAR Distillation [37] generates the low-beam pseudo point cloud and distills the knowledge from high-beam data, which achieves state-of-the-art results on high-to-low beam adaptation scenario. It can be seen from Table 1 that, the Bi3D greatly reduces the performance gap between different domains, surpassing all state-of-the-art UDA 3DOD methods. Note that the Bi3D largely improves the performance on the difficult Waymo→nuScenes setting (AP_{BEV} : 36.42% → 45.52% compared to ST3D, and 43.31% → 45.52% compared to LiDAR Distillation, AP_{3D} : 22.99% → 30.75% compared to ST3D, and 25.63% → 30.75% compared to LiDAR Distillation). Besides, our experiments are conducted under 1% target annotation budget, demonstrating that Bi3D can largely improve the cross-domain detection performance at a low annotation cost.

Comparison with 3D weakly-supervised DA works. SN [36] is a typical weakly-supervised DA method, which uses statistic-level normalization to reduce the domain difference caused by source-to-target object size variances. We conduct experiments combining our method with SN. The results are reported in Table 1, which show that our method outperforms all methods with SN operation. We find that the result can be further improved especially on cross-country adaptation setting (*i.e.* 71.56% → 77.55% on nuScenes→KITTI and 71.36% → 78.03% on Waymo→KITTI). This is mainly because SN can reduce the domain shift caused by object size variations and is ben-

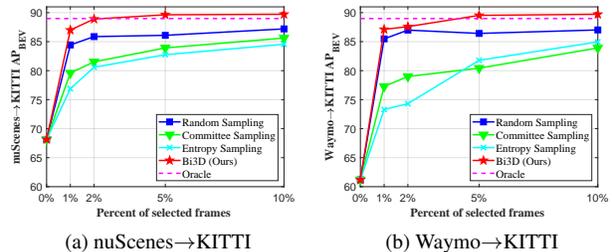


Figure 3. Results of various target annotation budgets.

eficial to pick up more target-domain-like source data.

4.4. Insightful Analyses

Results of Changing Target Domain Annotation Budget. In this part, we compare our Bi3D with several typical active learning methods (*i.e.* query-by-committee [23] and query-by-uncertainty [34]), and their results demonstrate that Bi3D can consistently outperform all these methods. We also conduct experiments on nuScenes→KITTI and Waymo→KITTI by changing the annotation budget. As illustrated in Fig. 3, we plot the trend of AP_{BEV} at different manual annotation budgets. It can be seen that our Bi3D achieves a promising detection accuracy gain, even outperforming many active learning methods. Besides, with the increase of the number of manually annotated target frames, the model detection accuracy is constantly improved. Furthermore, when the manually labeled target data reaches 5% of the total number of unlabeled frames, Bi3D can greatly improve the cross-domain detection accuracy of the baseline detector, even surpassing the fully-supervised results with 100% labeled target data.

As described above, we found that when 2D ADA works (as shown in Table 1) and 2D active learning works (as illustrated in Fig. 3) are deployed to 3D cross-domain scenarios, their results are unsatisfactory. Here, we analyze why these methods are not applicable to the cross-domain 3DOD task. We attribute the reason to the following two aspects. 1) *2D Density vs. 3D Sparsity:* Compared with 2D images, 3D point cloud is extremely sparse, which makes Global Average Pooling (GAP) based feature extractor not suitable for 3D scenes. As a result, directly leveraging CNN on highly sparse feature maps cannot extract informative features. 2) *2D Diversity vs. 3D Correlation:* Unlike 2D natural images that have more diverse appearances, the point cloud objects in autonomous driving are closely related, especially between adjacent frames in the same sequence. Thus, simply applying 2D ADA methods to 3DOD will yield similar importance metric scores of candidate data, resulting in labeling redundancy.

Ablation Studies. The effectiveness of two key components, including domainness-aware source sampling strategy and diversity-based target sampling strategy, is verified

Source	Target	SN	nuScenes→KITTI	Waymo→KITTI
			AP _{BEV} / AP _{3D}	AP _{BEV} / AP _{3D}
-	-	-	68.15 / 37.17	61.18 / 22.01
Ran.	-	✗	58.02 / 31.09	57.49 / 8.78
Act.	-	✗	73.90 / 43.02	68.27 / 28.53
Ran.	-	✓	70.13 / 58.80	71.23 / 56.20
Act.	-	✓	81.84 / 65.40	81.53 / 67.41
Ran.	Ran.	✓	84.42 / 75.12	85.48 / 75.89
Act.	Ran.	✓	85.02 / 75.43	85.70 / 76.12
Ran.	Act.	✓	86.53 / 76.54	86.12 / 76.92
Act.	Act.	✗	84.91 / 71.56	85.13 / 71.36
Act.	Act.	✓	87.00 / 77.55	87.12 / 78.03

Table 2. Component-level ablation studies. Ran. represents random sampling and Act. represents active sampling using our method. The ablation studies are conducted under 1% target annotation budget on PV-RCNN.

Score	Entropy	Dom.	nuScenes→KITTI	Waymo→KITTI
			AP _{BEV} / AP _{3D}	AP _{BEV} / AP _{3D}
-	-	S	79.65 / 58.74	77.71 / 55.52
✓	-	S	77.93 / 63.52	82.29 / 54.74
-	✓	S	78.90 / 60.02	79.79 / 58.43
✓	✓	S	81.84 / 65.40	81.53 / 67.41
-	-	S+T	86.13 / 76.65	85.58 / 76.69
✓	-	S+T	86.09 / 77.21	86.71 / 77.91
-	✓	S+T	86.39 / 76.82	85.91 / 77.96
✓	✓	S+T	87.00 / 77.55	87.12 / 78.03

Table 3. Ablation study of foreground region-aware discriminator under 1% target annotation budget. Score and Entropy in this table denote that we employ the objectness and entropy evaluation metrics, respectively. S denotes that only the source domain is used for the designed active sampling strategy and S+T represents our Bi3D).

on nuScenes→KITTI and Waymo→KITTI settings. On one hand, Table 2 indicates that the sampling strategy designed for the source domain has a better performance than the random sampling strategy. This is mainly due to that we pick up a portion of frames whose distribution characteristics are similar to the target domain. On the one hand, the sampled source data is also beneficial to improve the detector’s adaptability, further helping to select more representative samples from the target domain, as verified by comparing Source+Act and Source+Ran in Table 2. Moreover, Table 2 also shows that the designed diversity-based target domain sampling strategy also can significantly boost the model transferability between domains.

As mentioned in Section 3.2, we enhance the scene-level foreground features by combining objectness and entropy scores. In order to verify the necessity of such a design, we conduct experiments by changing the input features of the domain discriminator. It can be observed from Table 3 that, combining objectness and entropy scores achieves the best target-domain accuracy in both cases of only sampling

Method	Source	Target	Avg.
	AP _{BEV} / AP _{3D}	AP _{BEV} / AP _{3D}	AP _{BEV} / AP _{3D}
Source Only	47.57 / 32.43	68.15 / 38.17	57.86 / 35.30
ST3D [43]	28.57 / 19.88	78.36 / 70.85	53.46 / 45.36
Ours	40.71 / 22.89	84.91 / 71.56	62.81 / 47.22

Table 4. Generalization ability of Bi3D. We report the results tested on both source and target domains. Avg. denotes the average accuracy across two domains.

Method	nuScenes→KITTI	Waymo→KITTI
	AP _{BEV} / AP _{3D}	AP _{BEV} / AP _{3D}
ST3D [43]	84.29 / 72.94	86.65 / 76.86
Ours	87.00 / 77.55	87.12 / 78.03
Ours+ST3D [43]	89.28 / 79.69	87.83 / 81.23
Oracle	88.98 / 82.50	88.98 / 82.50

Table 5. The studies of combining Bi3D with UDA method under 1% target annotation budget on PV-RCNN.

source data, and sampling source and target data. This shows that the objectness score and entropy score can provide the location information of the foreground and make the model ignore the noisy background.

Bi3D for Enhancing 3DOD Generalization. Although domain adaptation tasks including UDA and ADA can achieve higher detection accuracy on a new target domain, their detection accuracy on the original source domain usually degrades after the domain adaptation process is finished. It can be seen from Table 4 that, the source-domain performance achieved by an adapted ST3D will cause a serious performance drop. In contrast, since our Bi3D utilizes a Bi-domain active sampling strategy to pick up both source and target samples, the adapted detector can maintain a certain degree of transferability toward the original source domain. From Table 4, we can observe that compared with ST3D, our model can obtain a better source domain performance. This shows that our Bi3D can enhance the detector’s dataset-level generalization ability for both source and target domains.

Combining with UDA Method. Current UDA works [43] mainly leverage self-training to perform the pseudo-labeling on the unlabeled target domain, which is orthogonal to our Bi3D. Therefore, we conduct the experiments of combining our Bi3D and ST3D [43]. In particular, we employ Bi3D to actively sample 1% target data to perform the manual annotation, and utilize ST3D to pseudo-label the remaining unlabeled target data. Then, we fine-tune the detector using both annotated data and pseudo-labeled data. It can be seen from Table 5 that, our Bi3D can be flexibly combined with ST3D, significantly outperforming both the Bi3D and ST3D methods.

5. Conclusion

In this work, for the first time, we presented a Bi3D framework, which develops a Bi-domain active sampling approach to dynamically select important frames from both source and target domains, achieving domain transfer at a low data cost. Experimentally, Bi3D achieves consistent accuracy gains on many cross-domain settings, e.g., for Waymo-to-KITTI setting, Bi3D re-trained on only 5% target domain data (KITTI) outperforms the corresponding baseline model trained using 100% labeled KITTI data.

Acknowledgement

This work is supported by National Natural Science Foundation of China (No. U1909207 and 62071127), Zhejiang Lab Project (No. 2021KH0AB05) and Science and Technology Commission of Shanghai Municipality (grant No. 22DZ1100102).

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. nuscnets: A multi-modal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2, 5, 12, 13
- [2] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017. 2
- [3] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018. 1
- [4] Sanjoy Dasgupta and Daniel Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th international conference on Machine learning*, pages 208–215, 2008. 2
- [5] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li. Voxel r-cnn: Towards high performance voxel-based 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1201–1209, 2021. 1, 2, 5
- [6] Bo Fu, Zhangjie Cao, Jianmin Wang, and Mingsheng Long. Transferable query selection for active domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7272–7281, 2021. 2, 3, 6, 7, 13, 15
- [7] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 1
- [8] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 2, 5, 12, 13
- [9] Alex Holub, Pietro Perona, and Michael C Burl. Entropy-based active learning for object recognition. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE, 2008. 3
- [10] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5001–5009, 2018. 1
- [11] Ajay J Joshi, Fatih Porikli, and Nikolaos P Papanikolopoulos. Scalable active learning for multiclass image classification. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2259–2273, 2012. 2
- [12] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platinisky, W. Jiang, and V. Shet. Level 5 perception dataset 2020. <https://level-5.global/level5/data/>, 2019. 5, 12, 13
- [13] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019. 1, 2
- [14] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. *Advances in neural information processing systems*, 29, 2016. 1
- [15] Zhipeng Luo, Zhongang Cai, Changqing Zhou, Gongjie Zhang, Haiyu Zhao, Shuai Yi, Shijian Lu, Hongsheng Li, Shanghang Zhang, and Ziwei Liu. Unsupervised domain adaptive 3d detection with multi-level consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8866–8875, 2021. 1
- [16] Munan Ning, Donghuan Lu, Dong Wei, Cheng Bian, Chenglang Yuan, Shuang Yu, Kai Ma, and Yefeng Zheng. Multi-anchor active domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9112–9122, 2021. 2, 3
- [17] Viraj Prabhu, Arjun Chandrasekaran, Kate Saenko, and Judy Hoffman. Active domain adaptation via clustering uncertainty-weighted embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8505–8514, 2021. 2, 3, 6, 13, 15
- [18] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 918–927, 2018. 2
- [19] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 2
- [20] Can Qin, Haoxuan You, Lichen Wang, C-C Jay Kuo, and Yun Fu. Pointdan: A multi-scale 3d domain adaption network for point cloud representation. *Advances in Neural Information Processing Systems*, 32, 2019. 1

- [21] Zhicong Qiu, David J Miller, and George Kesidis. A maximum entropy framework for semisupervised and active learning with unknown and label-scarce classes. *IEEE transactions on neural networks and learning systems*, 28(4):917–933, 2016. **3**
- [22] Cristiano Saltori, Stéphane Lathuilière, Nicu Sebe, Elisa Ricci, and Fabio Galasso. Sf-uda 3d: Source-free unsupervised domain adaptation for lidar-based 3d object detection. In *2020 International Conference on 3D Vision (3DV)*, pages 771–780. IEEE, 2020. **1**
- [23] Greg Schohn and David A. Cohn. Less is more: Active learning with support vector machines. In *International Conference on Machine Learning*, 2000. **2, 4, 6, 7, 13**
- [24] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. **2**
- [25] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294, 1992. **2**
- [26] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10529–10538, 2020. **1, 2, 3, 5**
- [27] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 770–779, 2019. **2**
- [28] Shaoshuai Shi, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2647–2664, 2020. **1, 2**
- [29] Jong-Chyi Su, Yi-Hsuan Tsai, Kihyuk Sohn, Buyu Liu, Subhansu Maji, and Manmohan Chandraker. Active adversarial domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 739–748, 2020. **2, 3**
- [30] Pei Sun, Henrik Kretschmar, 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. **5, 12, 13**
- [31] OpenPCDet Development Team. Openpcdet: An open-source toolbox for 3d object detection from point clouds. <https://github.com/open-mmlab/OpenPCDet>, 2020. **5**
- [32] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. **1**
- [33] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. **14**
- [34] Dan Wang and Yi Shang. A new active labeling method for deep learning. In *2014 International joint conference on neural networks (IJCNN)*, pages 112–119. IEEE, 2014. **2, 4, 6, 7, 13**
- [35] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600, 2016. **2**
- [36] Yan Wang, Xiangyu Chen, Yurong You, Li Erran Li, Bharath Hariharan, Mark Campbell, Kilian Q Weinberger, and Weiliun Chao. 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*, pages 11713–11723, 2020. **6, 7, 15**
- [37] Yi Wei, Zibu Wei, Yongming Rao, Jiaxin Li, Jie Zhou, and Jiwen Lu. Lidar distillation: Bridging the beam-induced domain gap for 3d object detection. *arXiv preprint arXiv:2203.14956*, 2022. **1, 2, 3, 5, 6, 7, 13, 15**
- [38] Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu, Xinjing Cheng, and Guoren Wang. Active learning for domain adaptation: An energy-based approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8708–8716, 2022. **2**
- [39] Ming Xie, Yuxi Li, Yabiao Wang, Zekun Luo, Zhenye Gan, Zhongyi Sun, Mingmin Chi, Chengjie Wang, and Pei Wang. Learning distinctive margin toward active domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7993–8002, 2022. **2**
- [40] Qiangeng Xu, Yin Zhou, Weiyue Wang, Charles R Qi, and Dragomir Anguelov. Spg: Unsupervised domain adaptation for 3d object detection via semantic point generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15446–15456, 2021. **1**
- [41] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. **1, 2, 13**
- [42] Bin Yang, Wenjie Luo, and Raquel Urtasun. Pixor: Real-time 3d object detection from point clouds. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7652–7660, 2018. **2**
- [43] Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. St3d: Self-training for unsupervised domain adaptation on 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10368–10378, 2021. **1, 2, 3, 5, 6, 7, 8, 12, 13, 15**
- [44] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Std: Sparse-to-dense 3d object detector for point cloud. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1951–1960, 2019. **2**
- [45] Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. Multiple instance active learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5330–5339, 2021. **2**
- [46] Bo Zhang, Tao Chen, Bin Wang, and Ruoyao Li. Joint distribution alignment via adversarial learning for domain adap-

- tive object detection. *IEEE Transactions on Multimedia*, 2021. [1](#)
- [47] Weichen Zhang, Wen Li, and Dong Xu. Srdan: Scale-aware and range-aware domain adaptation network for cross-dataset 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6769–6779, 2021. [1](#)
- [48] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018. [2](#)
- [49] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. [1](#)

Outline

Due to the eight-page limitation of the submission paper, we provide more details and visualizations from the following aspects:

- Sec. **A**: More details of the proposed Bi3D.
- Sec. **B**: Details of datasets.
- Sec. **C**: Implementation details for Bi3D and 2D image-related ADA works.
- Sec. **D**: More experimental results.
- Sec. **E**: Qualitative results.

A. More details of Bi3D

A.1. Detailed Structure of Domain Discriminator

As mentioned in the Method Section in our main text, we use a conventional convolution-based discriminator to measure the domainness score of samples from source and target domains. In particular, given a BEV feature map $\hat{f}_{bev}^d \in \mathbb{R}^{C \times H \times W}$ enhanced by objectness and entropy scores from source or target domain, where $d \in [s, t]$ denotes source or target domain and H, W, C represent the channel, height and width of the feature map, respectively. As shown in Fig. 4, we first use 5 layers convolution followed by LeakyReLU to extract scene-level features and the features are down-sampled by 32 times in this process. Then, the Global Average Pooling (GAP) is utilized to extract the scene-level vectors. Further, we use a single layer Multi-Layer Perception (MLP) to obtain the domainness score of each candidate frame.

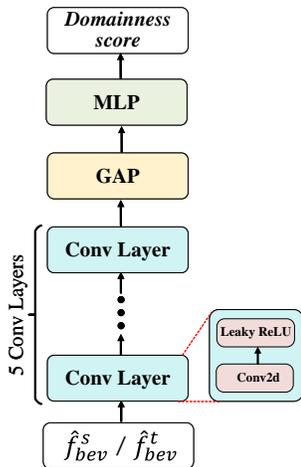


Figure 4. Detailed structure of domain discriminator. \hat{f}_{bev}^s and \hat{f}_{bev}^t denote enhanced foreground features from source domain and target domain.

A.2. Algorithm of the Proposed Bi3D

In this Section, we show the detail training procedure of the Bi3D in Algorithm 2.

Algorithm 2 Bi3D

- 1: Train the detector on train dataset D_s .
 - 2: Train the domain discriminator on D_s and D_t .
 - 3: Select target-domain-like source data \tilde{D}_s by domainness-aware source sampling strategy.
 - 4: Fine-tune the detector on \tilde{D}_s .
 - 5: **for** current epoch < max epochs **do**
 - 6: **if** current epoch in sample epochs **then**
 - 7: Select diverse and representative target data
 - 8: ΔD_t using diversity-based target sampling
 - 9: strategy and update \tilde{D}_t .
 - 10: Train detector on \tilde{D}_t and \tilde{D}_s .
-

B. Datasets

KITTI. KITTI dataset [8] is one of the most popular datasets for autonomous driving, which 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. Due to only annotations of the field of view of the front (FOV) camera is provided, we remove the points outside of the range in the test phase.

Waymo. Waymo Open Dataset [30] is a large-scale autonomous driving dataset which is composed of 1000 sequences and divided into a train set with 798 sequences (~ 1.5 million samples) and a validation set with 202 sequences (~ 4 million samples). The Waymo dataset is gathered in the USA by a 64-beam LiDAR with annotations in full 360°. We use one fifth data of Waymo train set when Waymo is regarded to the source domain.

nuScenes. nuScenes 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 time in a day and different weathers.

Lyft. Lyft level 5 dataset [12] is composed of 18900 training samples and 3780 validation samples, which are collected in the USA. As mentioned in ST3D [43], the Lyft dataset does not annotate objects on both sides of the road which is different from other datasets and will cause the detection accuracy degradation.

C. Implementation Details

C.1. More Bi3D Implementation Details

As shown in Table 6, the point cloud range varies in different datasets, followed by ST3D [43], we align

Dataset	Point Cloud Range
KITTI [8]	[0, -40, -3, 70.4, 40, 1]
Waymo [30]	[-75.2, -75.2, -2, 75.2, 75.2, 4]
nuScenes [1]	[-51.2, -51.2, -5.0, 51.2, 51.2, 3.0]
Lyft [12]	[-80.0, -80.0, -5.0, 80.0, 80.0, 3.0]

Table 6. Point cloud range of different datasets.

	Budget	KITTI	nuScenes	Lyft
number	1%	18	140	94
epochs		[0, 5]	[0, 5]	[0, 5]
number	5%	37	280	188
epochs		[0, 2, 4, 6, 8]	[0, 2, 4, 6, 8]	[0, 2, 4, 6, 8]

Table 7. Details of target sample budget and sample epochs. Number denotes the sampling number per epoch.

the point cloud range with the Waymo dataset (*i.e.*, [-75.2, -75.2, -2, 75.2, 75.2, 4]). In active sampling target domain stage, when annotation budget B of the target domain is equal to 1%, we perform a two-round sampling process. Besides, the sampling process will be performed 5 rounds, when annotation budget B is set to 5%. The sample number and sample epochs of the active sampling target data is shown in Table 7. The widely-used data augmentation methods (*e.g.*, random world flip, random world rotation, random world scaling) are also used in our experiments.

C.2. Details about Reproduced 2D ADA Methods

TQS [6]. TQS consists of transferable committee, transferable uncertainty and transferable domainness which is based on image-level features. In our reproduction, we extract the scene-level feature from BEV features using CNN and use three classifier heads to construct the committee. Following TQS [6], different from previous works using entropy to evaluate the uncertainty, we calculate the the margin of objectness score and 0.5 which is also different from the original TQS. This is because we only focus on a single category and cannot use margin of the highest score and the second-highest score. In addition, we calculate domainness score by a domain discriminator and set μ as 0.75 and σ as 0.4, which is the same as TQS. In order to keep consistency with TQS, we use source data and selected target data to fine-tune the detector trained on the source domain.

CLUE [17]. CLUE uses uncertainty-weighted cluster to select target data. Following CLUE, we first obtain the uncertainty of each frame by calculating the predictive entropy of the proposals after NMS and then use the mean value of the entropy to represent the frame-level uncertainty. Further, weighted K-Means which is proposed by CLUE with b_k centroids is utilized to cluster, where b_k denotes the annotation budget of the current sampling epoch.

The sampling number and sampling epochs are consistent with our Bi3D, as shown in Table 7.

C.3. Details about reproduced 2D AL Methods

Query by Committee [23]. Committee-based methods often use multiple classifier heads with different initialization to keep diversity from each other. We use two classifiers with different initialization and calculate the distance of output logits of two classifier heads and select the target data which the two classifier heads most disagree.

Query by Uncertainty [34]. The most common method to measure the uncertainty is to calculate the entropy. Here, we calculate the entropy of the instance-level classifier score of the proposals after NMS and use the mean value of the entropy to represent the frame-level uncertainty.

The sampling number and sampling epochs are consistent with our Bi3D, as shown in Table 7.

D. Experimental Results

D.1. Second-IOU Results

SECOND [41] is a widely used detector that greatly improves the efficiency of the model by using sparse convolution. Followed by [37, 43], we also conduct the experiment on SECOND-IoU, where an extra IoU head is added to SECOND. As shown in Table 8, the proposed Bi3D can also achieve the best result compared to UDA methods by only sampling 1% data from nuScenes. This shows our approach is applicable to multiple detectors.

Task	Method	AP _{BEV} / AP _{3D}
Waymo→nuScenes	Source Only	32.91 / 17.24
	SN	33.23 / 18.57
	ST3D	35.92 / 20.19
	Lidar Distill	40.66 / 22.86
	Ours	42.15 / 26.24

Table 8. Results of Waymo→nuScenes adaptation task using SECOND-IoU. Source Only means that the model trained on the Waymo dataset is directly tested on nuScenes. We report AP_{BEV} and AP_{3D} over 40 recall positions of the car category at IoU = 0.7.

D.2. More Ablation Studies

Experiments on number of selected source data. In our main text, the experimental results have shown the effectiveness of our proposed domainness-aware source sampling strategy. To fully explore the influence of the number of selected source data, we conduct further experiments. As shown in Table 9, selecting more source domain data degrades performance (*e.g.*, select 25% source data) as the domain gap between \tilde{D}_s and D_t becomes larger, where \tilde{D}_s denotes the set of selected target-domain-like source data and D_t represents target dataset.

Number of Source		AP_{BEV} / AP_{3D}
Proportion	1%	86.95 / 77.86
	2%	86.55 / 77.55
	5%	86.66 / 77.59
	10%	85.89 / 76.78
	25%	85.24 / 75.68
Threshold	> 0	87.00 / 77.58

Table 9. Experiments on number of selected source data. We conduct the experiments on Waymo→nuScenes adaptation task using PV-RCNN. Number of Source indicates the number of selected source data using domainness-aware source sample strategy. We report AP_{BEV} and AP_{3D} over 40 recall positions of the car category at IoU = 0.7.

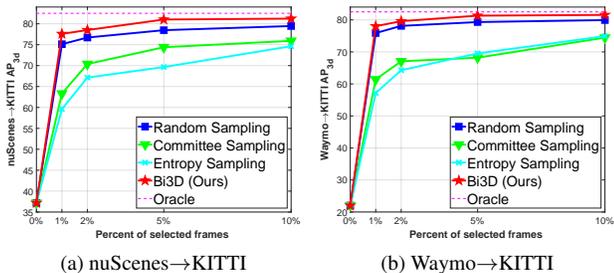


Figure 5. Results of various target annotation budgets.

Varying Annotation Budget. We show the line chart of AP_{BEV} at different target annotation budgets in our main text. Here we show the trend of AP_{3D} . As shown in Fig. 5, similar to the results of AP_{BEV} , the proposed Bi3D achieves a promising detection accuracy gain, even outperforming many active learning methods. In addition, with the increase of the number of manual annotated target frames, the model detection accuracy is constantly improved.

D.3. Results with IoU=0.5

In this section, we show AP_{BEV} and AP_{3D} results with the IoU threshold 0.5. The results are shown in Table 10, we can observe that our Bi3D can also achieve the best performance on various cross-domain 3DOD tasks.

E. Qualitative Results

E.1. t-SNE results

To illustrate that Bi3D effectively samples target-like source data, we first visualize the scene-level features of source and target domains using t-SNE [33]. As shown in Fig. 6, the selected source data distribute around the domain boundaries between the source and target domains, meaning that we sample target-domain-like source data. Besides, the visualization of instance-level features from the target domain is shown in Fig. 7, and we can observe that the diverse target data are selected using Bi3D.

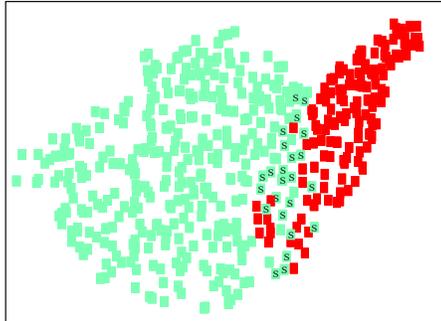


Figure 6. t-SNE result of domainness-aware source sampling strategy. The green squares represent data from source domain and red squares represent target domain. We mark the selected source data with 's'.

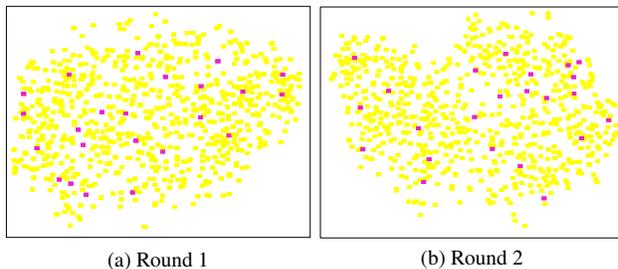


Figure 7. t-SNE result of diversity-based target sampling strategy. The purple squares represent selected target data. As mentioned in C.1, we perform a two round sampling strategy. (a) and (b) are the t-SNE results of each sampling round.

E.2. Visualization

To better verify the effectiveness of our Bi3D, we finally provide some visualizations. Fig. 8 and Fig. 9 show qualitative results of Waymo-to-KITTI cross-domain scenario equipped with PV-RCNN. Fig. 10 shows the qualitative results of Waymo-to-nuScenes cross-domain scenario. It can be seen that our method can predict high-quality 3D bounding boxes. Besides, due to the differences in the taxonomy of different datasets (*e.g.*, in the waymo dataset, cars, trucks and buses are annotated as 'Vehicle' and it is quite different from nuScenes dataset, which only annotated cars as 'car'), we can observe that the model detects trucks and buses in nuScenes, which will reduce the detection accuracy.

Task	Method	PV-RCNN	Voxel R-CNN
		AP _{BEV} / AP _{3D}	AP _{BEV} / AP _{3D}
Waymo→KITTI	Source Only	88.33 / 87.17	89.49 / 87.41
	ST3D [43]	92.40 / 92.18	92.34 / 84.92
	Ours	93.88 / 92.18	92.76 / 92.08
	SN [36]	86.32 / 85.72	82.00 / 81.57
	ST3D (w/ SN) [43]	91.49 / 90.77	86.76 / 86.40
	CLUE (w/ SN) [17]	90.42 / 88.87	88.13 / 88.00
	TQS (w/ SN) [6]	88.17 / 89.87	84.77 / 83.00
	Ours (w/ SN)	92.48 / 92.31	94.31 / 94.16
	Ours (w/ SN, 5%)	93.31 / 93.26	94.27 / 94.17
	Oracle	94.08 / 92.28	95.54 / 95.50
Waymo→Lyft	Source Only	82.38 / 80.45	77.09 / 75.16
	ST3D [43]	84.52 / 82.61	78.43 / 76.68
	Ours	86.03 / 83.90	85.04 / 83.04
	SN [36]	80.12 / 78.09	76.26 / 73.54
	ST3D (w/ SN) [43]	82.21 / 81.70	77.46 / 75.08
	CLUE (w/ SN) [17]	84.00 / 81.96	83.74 / 81.77
	TQS (w/ SN) [6]	75.60 / 73.45	77.64 / 75.67
	Ours (w/ SN)	86.03 / 83.86	86.19 / 83.70
	Ours (w/ SN, 5%)	86.70 / 84.26	86.84 / 83.58
	Oracle	92.38 / 91.87	90.19 / 87.18
Waymo→nuScenes	Source Only	40.48 / 36.95	31.62 / 28.25
	ST3D [43]	40.90 / 38.67	44.06 / 34.62
	Ours	52.99 / 49.17	52.39 / 48.83
	SN [36]	40.27 / 36.59	33.99 / 31.23
	ST3D (w/ SN) [43]	41.42 / 38.99	38.27 / 34.31
	CLUE (w/ SN) [17]	43.79 / 40.80	42.67 / 39.87
	TQS (w/ SN) [6]	41.10 / 38.01	38.87 / 36.38
	Ours (w/ SN)	52.65 / 48.01	52.73 / 49.04
	Ours (w/ SN, 5%)	55.63 / 51.78	53.01 / 49.63
	Oracle	61.52 / 58.04	58.33 / 54.61
nuScenes→KITTI	Source Only	80.88 / 78.47	85.81 / 81.76
	ST3D [43]	83.75 / 83.64	92.33 / 82.93
	Ours	92.54 / 92.36	94.86 / 93.28
	SN [36]	66.22 / 65.82	48.59 / 47.49
	ST3D (w/ SN) [43]	90.47 / 90.25	80.08 / 78.51
	CLUE (w/ SN) [17]	82.04 / 80.59	85.97 / 82.86
	TQS (w/ SN) [6]	91.90 / 90.37	85.88 / 84.39
	Ours (w/ SN)	92.93 / 92.74	93.47 / 93.32
	Ours (w/ SN, 5%)	94.70 / 93.44	93.65 / 92.46
	Oracle	94.97 / 94.85	95.54 / 95.50

Table 10. Results of different adaptation scenarios under 1% annotation budget. Following [37, 43], we report AP_{BEV} and AP_{3D} over recall 40 positions of the car category at IoU = 0.5. Source Only denotes that the pre-trained detector is directly evaluated on the target domain, and Oracle represents that the detection results using the fully-annotated target domain. Closed Gap denotes the performance gap closed by various approaches along Source Only and Oracle results. The best adaptation results are marked in **bold**.

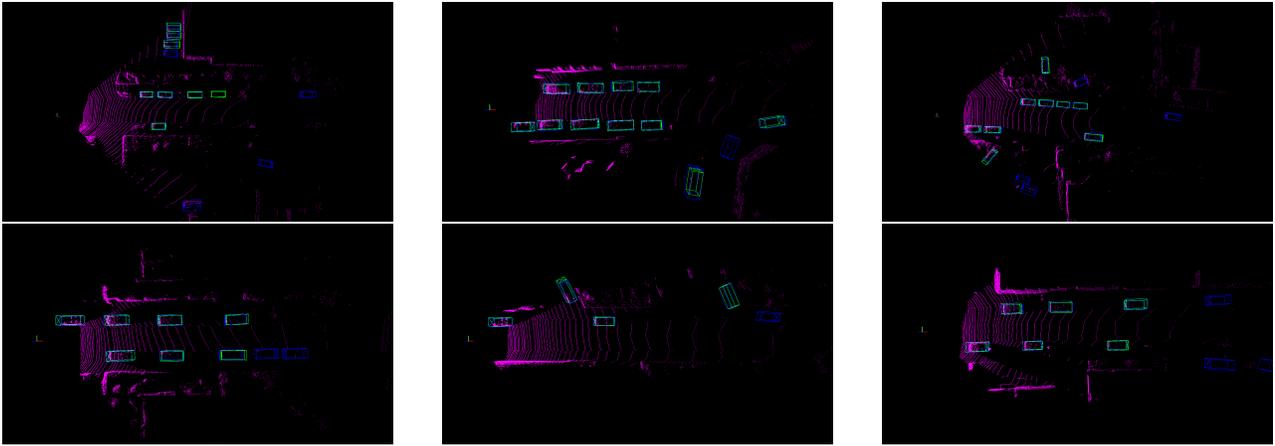


Figure 8. Qualitative results of Waymo-to-KITTI cross-domain scenario. The green and blue bounding boxes represent groundtruths and detector predictions respectively.

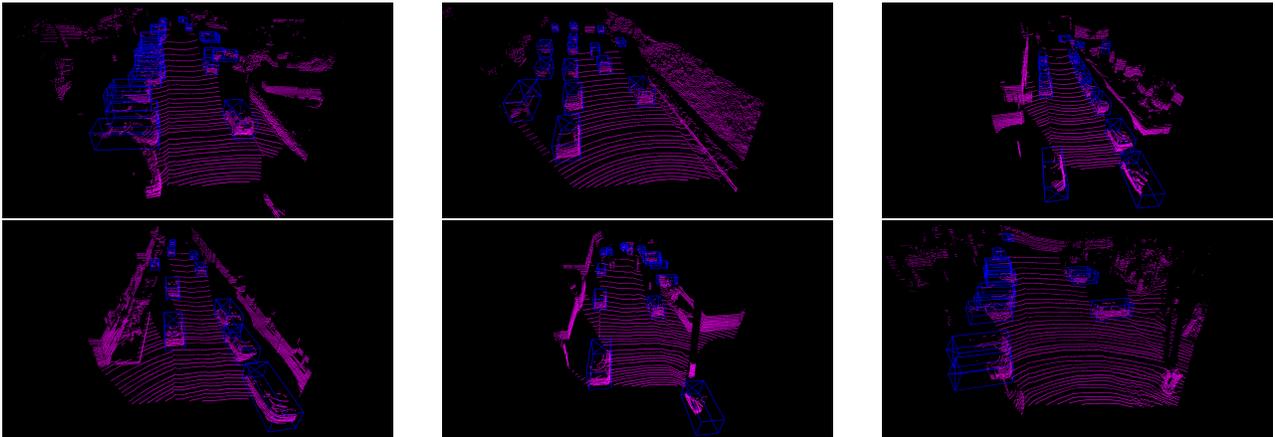


Figure 9. Qualitative results of Waymo-to-KITTI cross-domain scenario. We visualize the detection results in the target domain (KITTI).

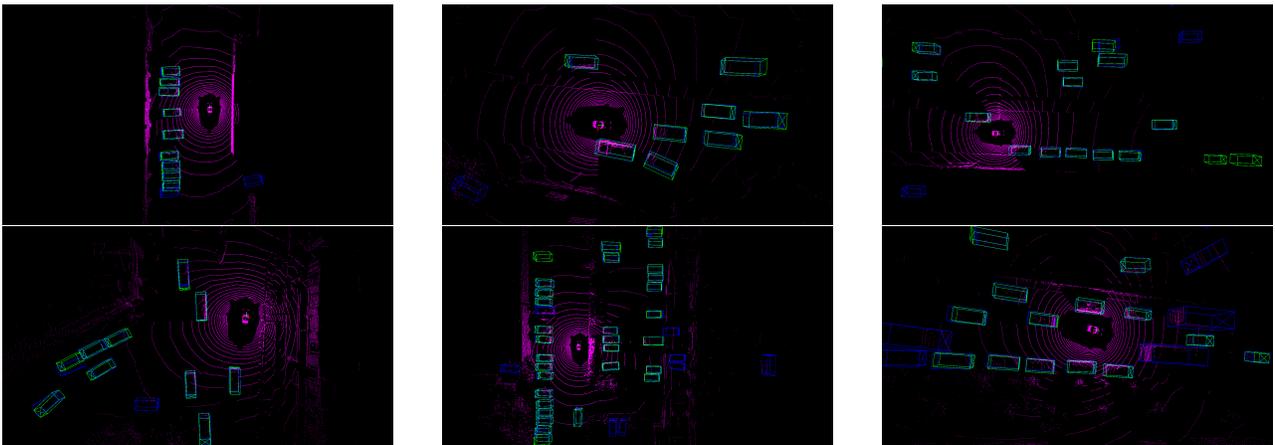


Figure 10. Qualitative results of Waymo-to-nuScenes cross-domain scenario. We visualize the detection results in the target domain (nuScenes).