Revisiting Domain-Adaptive 3D Object Detection by Reliable, Diverse and Class-balanced Pseudo-Labeling

Zhuoxiao Chen[1]   Yadan Luo[1]   Zheng Wang[2]   Mahsa Baktashmotlagh[1]   Zi Huang[1]

[1]The University of Queensland   [2]University of Electronic Science and Technology of China

{zhuoxiao.chen, y.luo, m.baktashmotlagh, helen.huang}@uq.edu.au, zh_wang@hotmail.com

Abstract

*Unsupervised domain adaptation (DA) with the aid of pseudo labeling techniques has emerged as a crucial approach for domain-adaptive 3D object detection. While effective, existing DA methods suffer from a substantial drop in performance when applied to a multi-class training setting, due to the co-existence of low-quality pseudo labels and class imbalance issues. In this paper, we address this challenge by proposing a novel **ReDB** framework tailored for learning to detect all classes at once. Our approach produces **R**eliable, **D**iverse, and class-**B**alanced pseudo 3D boxes to iteratively guide the self-training on a distributionally different target domain. To alleviate disruptions caused by the environmental discrepancy (e.g., beam numbers), the proposed cross-domain examination (CDE) assesses the correctness of pseudo labels by copy-pasting target instances into a source environment and measuring the prediction consistency. To reduce computational overhead and mitigate the object shift (e.g., scales and point densities), we design an overlapped boxes counting (OBC) metric that allows to uniformly downsample pseudo-labeled objects across different geometric characteristics. To confront the issue of inter-class imbalance, we progressively augment the target point clouds with a class-balanced set of pseudo-labeled target instances and source objects, which boosts recognition accuracies on both frequently appearing and rare classes. Experimental results on three benchmark datasets using both voxel-based (i.e., SECOND) and point-based 3D detectors (i.e., PointRCNN) demonstrate that our proposed ReDB approach outperforms existing 3D domain adaptation methods by a large margin, improving 23.15% mAP on the nuScenes → KITTI task. The code is available at https://github.com/zhuoxiao-chen/ReDB-DA-3Ddet.*

1. Introduction

As LiDAR-based 3D object detection continues to gain traction in various applications such as robotic systems [1, 40, 60, 56, 83, 74] and self-driving automobiles [8, 59,
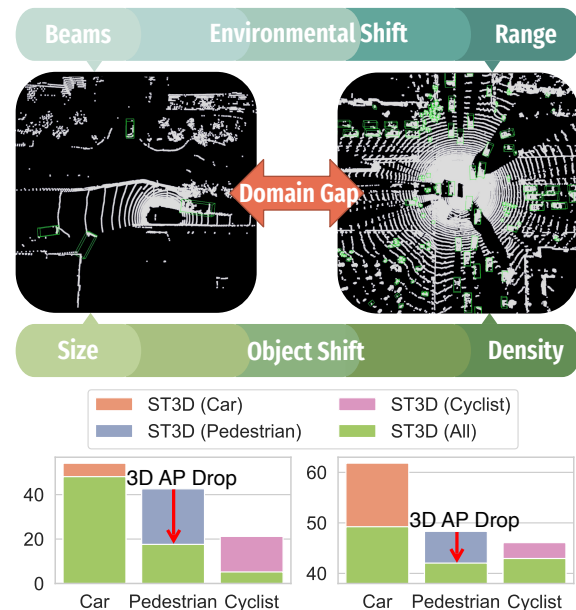

Figure 1: *Top*: An illustration of the domain gap in 3D point clouds. *Bottom*: Average Precision (AP) drop of ST3D [70] when applied to the multi-class adaptation from nuScenes to KITTI (*left*), and from Waymo to KITTI (*right*).

42, 61, 2, 78, 37, 39, 32, 25, 38], it becomes increasingly vital to address the challenges of deploying detectors in real-world scenes. The primary obstacles stem from the discrepancy between the training and test point cloud data, which are commonly curated from different scenes, locations, times, and sensor types, creating a *domain gap*. This domain gap mainly comes from the *object shift* and *environmental shift*, and can significantly degrade the prediction accuracy of 3D detectors. Object shift [62, 36] refers to the changes in the spatial distribution, point density, and scale of objects between the training and test domains. For instance, the average length of cars in the Waymo dataset [52] is significantly different from the one in the KITTI dataset [15] by around 0.91 meters [62]. Environmental shift, on the other hand, arises from composite differences in the sur-

rounding environment such as inconsistent beam numbers, angles, point cloud ranges, and data acquisition locations. For example, in Fig 1, Waymo generates 3D scenes by 64-beam LiDAR sensors, while nuScenes [3] comprises more sparse 32-beam environments with double large beam angles.

The co-existence of object shift and environmental shift poses a great challenge to the 3D detection models to be deployed in the wild. To combat such a problem, domain-adaptive 3D detection approaches [36, 71, 70, 79, 80] have been exploited to adapt the model trained on a labeled dataset (*i.e.*, source domain) to an unlabeled dataset from a different distribution (*i.e.*, target domain). In this area, pioneering research of ST3D [70] presents a self-training paradigm that generates pseudo-labeled bounding boxes to supervise the subsequent learning on the target point clouds. In the same vein, later studies [36, 79, 71, 76, 77] seek different solutions based on self-supervised techniques, such as mean-teacher framework [36] and contrastive learning [79] for stable optimization and obtaining better embeddings.

Revisiting Domain-adaptive 3D Detection Setup. The aforementioned domain-adaptive 3D detection approaches have typically followed a **single-class training** setting, where the models are trained to adapt to each class separately. While it is more practical and fairer to train the models with all classes, our empirical study has shown that the detection performance of prior works decreases considerably when switching to a multi-class setting (Fig 1). Such an average precision (AP) drop can be attributed to the poor quality of produced pseudo labels (*i.e.*, erroneous and redundant) and the lower recognition accuracy of rare classes (*e.g.*, 91 times fewer bicycles than cars in Waymo [52]).

In this work, we propose a novel REDB framework (Fig 2) that aims to generate Reliable, Diverse, class-Balanced pseudo labels for the domain-adaptive 3D detection task. Our approach addresses the challenge of multi-class learning by incorporating the following three mechanisms:

Reliability: Cross-domain Examination. To remove erroneous pseudo labels of high confidence and avoid error accumulation in self-training, we introduce a cross-domain examination (CDE) strategy to assess pseudo label reliability. After copying the pseudo-labeled target objects into a familiar source environment, the reliability is measured by the consistency *i.e.*, Intersection-over-Union (IoU) between two box predictions in the target domain and the source domain, respectively. Any objects with low IoUs will be considered unreliable and then discarded. To prevent point conflicts between the source and target point clouds, we remove the source points that fall within the region where pseudo-labeled objects will be copied to. The proposed CDE ensures that the accepted pseudo-labeled instances are domain-agnostic and less affected by environmental shifts.

Diversity: OBC-based Downsampling. To avoid redundant pseudo labels that are frequently and similarly scaled, it is important to prevent the trained detector from collapsing into a fixed pattern that may only detect certain types of objects (*e.g.*, small-sized cars), and miss other instances of unique styles (*e.g.*, buses and trucks). In order to enhance geometric diversity, we derive a metric called overlapped boxes counting (OBC) metric to uniformly downsample the pseudo labels. The metric design is motivated by the observation that 3D detector tends to predict more boxes for objects with uncommon geometries, as they are harder to localize with only a few tight boxes. We count the number of regressed boxes surrounding each detected object as OBC and use kernel density estimation (KDE) to estimate its empirical distribution. We then apply downsampling according to the inverse probability of KDE, effectively reducing the number of pseudo labels in the high-density OBC region, where objects have similar and frequent geometries. By learning from a diverse subset of pseudo labels, the 3D detector can better identify objects of various scales and point densities, potentially mitigating object shift.

Balance: Class-balanced Self-Training. Despite the fact that reliable and diverse pseudo labels can be selected by the previous two modules, there still exists a significant interclass imbalance. To achieve class-balanced self-training, we randomly inject pseudo-labeled objects into each target point cloud, with an equal number of samples from each category. By learning from such class-balanced target data, the model can better grasp the holistic semantics of the target labels. To enable a smooth transition of learning from the source data to the target data, we first augment the target data with labeled source objects in a class-balanced manner for the initial few steps. We then gradually reduce the ratio of source objects and increase the number of RED labels as the self-training proceeds. This progressive class-balanced self-training allows the model to adapt stably to the target domain, enhancing recognition for both frequently appearing and rare classes.

Our work rectifies the setup of domain-adaptive 3D detection to a multi-class scenario and proposes a novel REDB framework for pseudo labeling in domain-adaptive 3D detection. Extensive experiments on three large-scale testbeds evidence that the proposed REDB is of exceptional adaptability for both voxel-based and point-based contemporary 3D detectors in varying environments, improving 3D mAP of 20.66% and 23.15% over state-of-the-art methods on the nuScenes $\rightarrow$ KITTI tasks, respectively.

2. Related Work

3D Object Detection from Point Clouds. With the growing advancement of neural networks over the last decade, the 3D point clouds can be effectively encoded by deep models to either point-based representation or grid-based
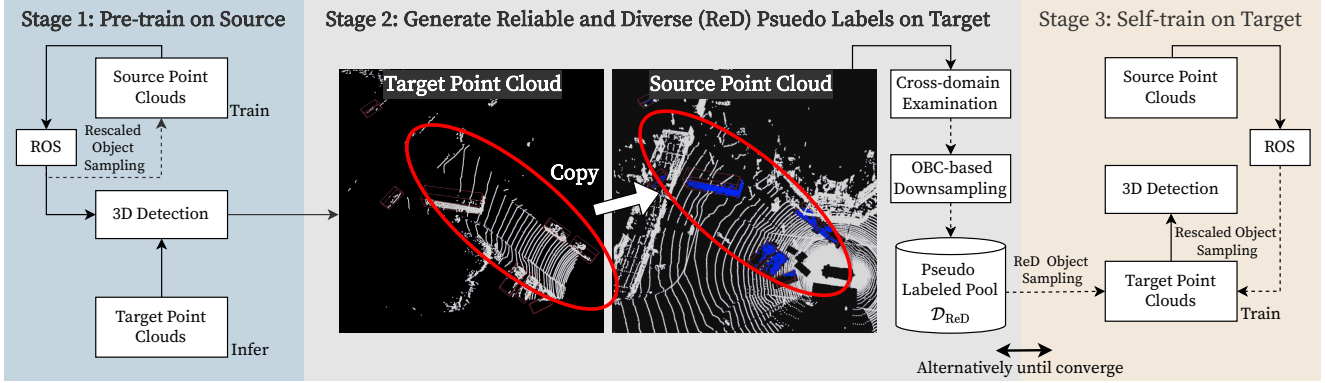
Figure 2: The overall framework of the proposed method: we firstly obtain high confident pseudo labels from Stage 1. In Stage 2, we check the reliability of pseudo labels by the cross-domain examination (CDE) and reduce geometrical redundancy by OBC-based downsampling. Finally, we progressively augment target point clouds by injecting RED labels and source labels for class-balanced self-training in Stage 3.

representation. The series of point-based detectors [48, 73, 72, 50, 41, 58, 81] directly encode 3D objects from raw points using PointNet encoder, then generate 3D proposals from point features. This type of encoding strategy can aid the model better preserve and learn the geometric properties of 3D objects. While, the grid-based detectors [11, 24, 69, 82, 68, 22, 63, 49, 9, 75, 47, 53, 12] rely on convolution neural networks for efficiently capturing regularly voxelized point clouds. Hybrid methods [18, 20, 71, 70] take advantages of both paradigms, yielding better results at the expense of processing speed. However, the preceding work overlooks the domain gap in different 3D scenes such that could scarcely be applied to unseen scenarios.

Domain Adaptation for 3D Detection. To surmount such difficulty, unsupervised domain adaptation (UDA) strives to transfer knowledge from a labelled source domain to an unlabeled target domain. There are primarily two types of strategies to eliminate the discrepancy between two domains. One is the adversarial-based approach [13, 14, 57, 44, 19, 10, 6, 34, 35, 33, 27, 21] that learns the domain-invariant features. The other is the statistical-based approach [31, 30, 29, 51, 5, 64, 65, 26, 28] that seeks to minimize the distribution distance between two domains. Despite the fact that UDA techniques have been extensively investigated in the last decade, few UDA approaches aim to address the domain shift existing in 3D object detection. The distribution shifts of 3D detection are identified as dissimilar object statistics [62], weather effects [67, 17], sensor difference [43, 16, 66] or synthesis-to-real gap [45, 7, 23]. To close the domain shift, an adversarial-based approach [80] was proposed to match features of different scales and ranges. Another streams of work focus on generating and enhancing 3D pseudo labels through memory bank [70, 71], mean-teacher paradigm [36] or contrastive learning [79]. However, existing methods necessitate either extra compu-

tational time or significant memory usage. Besides that, these methods generate low-quality pseudo-labels induced by environmental gaps and numerous analogous geometrical features, and overlook the class-imbalanced issue.

3. Methodology

3.1. Problem Definition

In this section, we mathematically formulate the problem of unsupervised domain adaptation for 3D object detection and set up the notations. The main objective is to adapt a 3D object detector trained on the labeled point clouds $\{(X_i^s, Y_i^s)\}_{i=1}^{N_s}$ from the source domain, to the unlabeled data $\{X_i^t\}_{i=1}^{N_t}$ in the target domain. N_s and N_t indicate the number of point clouds in the source and target domains, respectively. The 3D box annotation of the i-th source point cloud is denoted as $Y_i^s = \{b_j\}_{j=1}^{B_i}$ and $b_j = (x, y, z, w, l, h, \theta, c) \in \mathbb{R}^8$, where B_i is the total number of labeled boxes in X_i^s, (x, y, z) represents the box center location, (w, l, h) is the box dimension, θ is the box orientation, and $c \in \{1, ..., C\}$ is the object category.

3.2. Overall Framework

Our approach consists of three stages to facilitate knowledge transfer from the label-rich source domain to the label-scarce target domain, as illustrated in Fig 2. In Stage 1, the 3D detector (*e.g.*, SECOND [68] or PointRCNN [48]) is pre-trained on the source domain using the standard random object scaling (ROS) augmentation [70]. Upon reaching convergence of pre-training, the unlabeled point clouds are passed to the pre-trained detector to generate pseudo-labels $\{\widehat{Y}_i^t\}_{i=1}^{N_t}$ of high confidence for the target data $\{X_i^t\}_{i=1}^{N_t}$. Specifically, the produced pseudo labels will undergo scrutiny by the cross-domain examination (CDE)
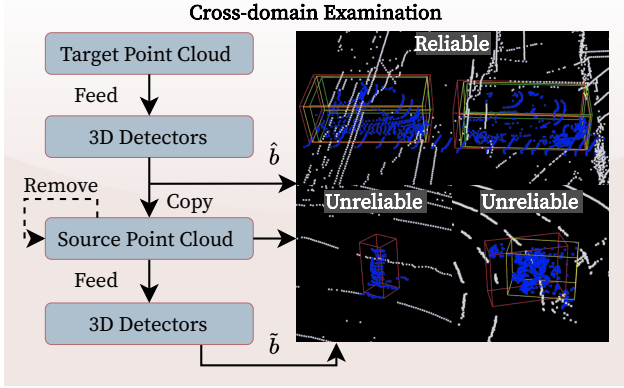
Cross-domain Examination

Figure 3: The proposed cross-domain examination (CDE) strategy involves copying pseudo target points X^{ps} in blue, with red and yellow boxes representing initial predictions in the target and source environments, respectively. Green boxes denote the ground truth for reference. If the intersection over union (IoU) between the target and source predictions is large, like in the first example, the reliability is accepted. However, in cases where $\hat{b}$ is not detected or the IoU is small, like in the second and third examples, these labels are considered unreliable due to inconsistencies caused by environmental shift.

strategy (Sec 3.3) and down-sampled by the OBC-based diversity module (Sec 3.4), forming a subset of reliable and diverse (RED) objects associated with pseudo labels. The details of stage 2 are illustrated in Fig 4. During model self-training on the target set in Stage 3, we randomly inject RED target objects and source objects into each target point clouds in a class-balanced manner (Sec 3.5), with the ratio of source samples gradually decreasing. The 3D detector is iteratively trained to adapt by alternating between Stage 2 and Stage 3.

3.3. Reliability: Cross-domain Examination

The proposed cross-domain examination (CDE) strategy aims to identify high-quality pseudo labels that are robust to the environmental shift. First, given a 3D detector $F(\cdot; \theta^s)$ pre-trained on the source domain, we can obtain the initial pseudo labels by inferring a target point cloud X_i^t as:

$$\widehat{Y}_i^t = \{\hat{b}_j\}_{j=1}^{\widehat{B}_i} = F(X_i^t; \theta^s), \qquad (1)$$

where θ^s is the pre-trained weights, and $\widehat{Y}^t \in \mathbb{R}^{\widehat{B} \times 8}$ denotes the pseudo labels with the model confidence greater than δ_{pos}. $\widehat{B}_i$ is the number of obtained pseudo labels. Below, we omit the subscript i for simplicity. To attain a real source environment, we randomly sample a source point cloud $X^s \sim rand(\{(X_i^s)\}_{i=1}^{N_s})$ and copy-paste the target point clouds $X^{ps} \subset X^t$ that fall within $\widehat{Y}^t$, into the sampled source point cloud as shown in Fig 3. To pre-

vent conflicts between existing points and the pseudo points to be pasted, we apply the object points removal operator $\mathbb{R}(\cdot)$ to remove source points residing in the area overlapped with the copied target samples X^{ps}. Now, we can safely move the target points to the source point clouds as $\mathbb{R}(X^s) \oplus X^{ps}$, where $\oplus$ represents point concatenation. We generate new predictions $\widetilde{Y}^t$ of high confidence for the concatenated source point clouds as:

$$\widetilde{Y}_i^t = \{\tilde{b}_j\}_{j=1}^{\widetilde{B}_i} = F(\mathbb{R}(X^s) \oplus X^{ps}); \theta^s), \qquad (2)$$

where $\widetilde{B}_i$ represents the number of boxes in the updated source clouds. To examine the resiliency of the pseudo-labeled object $\hat{b} \in \hat{Y}^t$ to the environmental shift, we compute the intersection over union (IoU) between the initial predicted box $\hat{b}$ and its corresponding prediction $\tilde{b} \in \tilde{Y}^t$ in the source environment. Through comparing the IoU to a predefined IoU threshold δ_{cde}, we can effectively discriminate reliable pseudo labels from unreliable ones by using the following criterion:

$$f_{\text{CDE}}(\hat{b}; \tilde{b}) = \begin{cases} \hat{b} & \text{IoU}(\hat{b}, \tilde{b}) \geq \delta_{\text{cde}}, \\ \emptyset & \text{otherwise.} \end{cases} \qquad (3)$$

We apply the CDE only at the first round of pseudo labelling because the pre-trained detector has no knowledge about target environments. The process of proposed CDE strategy is presented in Fig 3 with three examples for illustration.

3.4. Diversity: OBC-based Downsampling

Although we have already removed unreliable pseudo labels, the remaining label set may contain objects with similar patterns. This hinders the model from learning a more comprehensive distribution of the target domain. To ensure that the pseudo label set is representative of the broad spectrum of geometric characteristics present in target data, we investigate ways to properly measure object diversity. Our empirical study shows that 3D detectors generate more box predictions around unfamiliar objects, as visualized in Fig 5. It is observed that, the box predictions (in yellow) before non-maximum suppression (NMS) have obvious overlaps (i.e., IoU $> \delta_{\text{obc}}$) with the corresponding final predicted box (in red). Besides, predicted boxes for low-density objects typically orientate differently, while the regressed box scales for large objects have a high variance.

This observation motivates us to derive the concept of overlapped boxes counting (OBC), which counts the number of predicted boxes that surround the same object with the IoU $> \delta_{\text{obc}}$. The OBC can comprehensively reflect the uniqueness in geometric characteristics of objects, including scales, locations, point densities. A qualitative analysis of OBC is provided in Sec 4.3, which evidence that pseudo-labeled target objects with high OBC values tend to
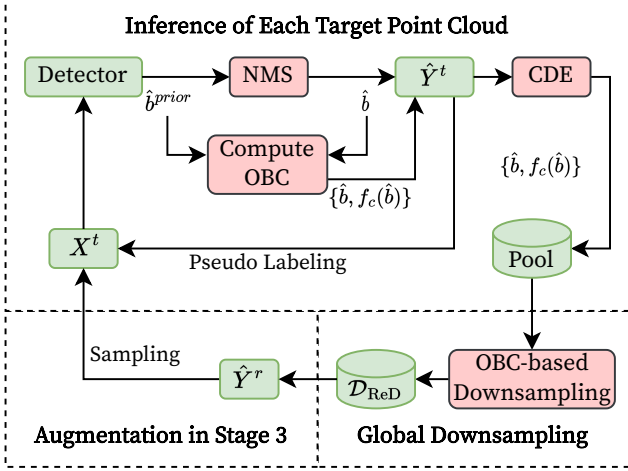
Figure 4: In Stage 2 of our approach, we input each target point cloud X^t to the pre-trained detector and obtain 3D box predictions. By applying NMS, we select confident and non-overlapped boxes $\hat{b}$ as pseudo labels. Simultaneously, we compute OBC score for each pseudo-labeled $\hat{b}$, based on prior-NMS boxes $\hat{b}^{prior}$. Next, we use CDE to add qualified $\hat{b}$ to a pool. After inferring all target data, we globally downsample the pseudo-labeled boxes in the pool using OBC, resulting in $\mathcal{D}_{\text{ReD}}$. In Stage 3, we augment each target point cloud by sampling high-quality objects from $\mathcal{D}_{\text{ReD}}$ (Eq 6) in a class-balanced manner during self-training. Stage 2 and 3 are alternated until convergence.

be uncommon in different geometric aspects. To estimate the probability density function (PDF) of OBC values, we use bar plots to show the frequency of different OBC values, and fit it with Kernel Density Estimation (KDE) (as shown in the blue curve of Fig 5). It can be seen from the plot that the distribution of OBC is very skewed, implying that the majority of geometric features are quite analogous since fell at high frequency interval (*i.e.*, from 3 to 12). Conversely, objects with rare geometrics fell on the long distribution tail (*i.e.*, from 12 to 30). To get a diverse subset and exclude too many objects in high density regions, we propose to leverage the inverse KDE function (shown in red curve) to assign a low sampling weight for pseudo-labeled objects fell in dense regions, while high sampling probability for objects in tail regions. In particular, we firstly collect the set $\mathcal{O}$ of all OBC values with the cardinality $\hat{B} = \sum_{i=1}^{N_t} \hat{B}_i$:

$$\mathcal{O} = \{f_c(\hat{b}_j) | b_j \in \hat{Y}_i^t\}_{i=1}^{N_t}, \qquad (4)$$

where $f_c(\cdot)$ is the function to count OBC value for each pseudo-labeled box $\hat{b}$. The KDE with the Gaussian kernel of the random variable O can be calculated with a finite set of $\mathcal{O}$:

$$f_{\text{KDE}}(O) = \frac{1}{\hat{B}\sigma\sqrt{2\pi}} \sum_{j=1}^{\hat{B}} e^{-\frac{1}{2}(\frac{f_c(\hat{b}_j)-O}{\sigma})^2}, \qquad (5)$$
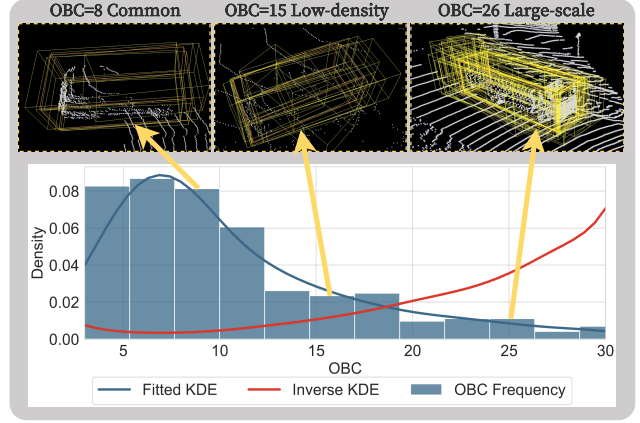

Figure 5: An illustration of overlapped boxes counting (OBC). The upper part shows box predictions before NMS generated around three positively predicted objects associated with different OBC values. The bottom part is the plot demonstrating the distributions of OBC values of all detected objects, with the fitted KDE (blue) and inverse KDE (red) for diverse sampling.

where σ is the standard deviation. The sampled subset $\mathcal{D}_{\text{ReD}}$ of size $\hat{B}/d$ are uniformly downsampled from the inverse KDE, where $d > 1$ is the down sampling rate. Finally, the sampled subset $\mathcal{D}_{\text{ReD}}$ contains pseudo-labelled objects with more uniformly distributed OBC values, indicating less common patterns and more diverse geometrics (*e.g.*, object scales, distances and densities).

3.5. Balance: Class-balanced Self-Training

Regardless OBC-based downsampling creates a pool of reliable and intra-class diverse (RED) pseudo-labels, but the inter-class imbalance issue still poses a challenge for multi-class adaptation. We aim to leverage a class-balanced paradigm to inject the generated pseudo labels to each target point cloud, during the self training stage:

$$\widehat{Y}^r = \{\{\hat{b}_j^1\}_{j=1}^{S^r} \oplus \cdots \oplus \{\hat{b}_j^C\}_{j=1}^{S^r}\} \sim \mathcal{D}_{\text{ReD}}, \qquad (6)$$

where $\oplus$ represents the point cloud concatenation and the superscript $1, \cdots, C$ of $\hat{b}$ indicates the predicted class of the pseudo label, and S^r is the sampling number of target RED labels for each class. To moderate and stable the adaptation process with the huge domain gap, we augment the target data by injecting source ground truth (GT) labels with ROS augmentation:

$$
\begin{aligned}
P^g &= \{Y_1^s \oplus \cdots \oplus Y_{N_s}^s\}, \\
Y^g &= \{\{b_j^1\}_{j=1}^{S^g} \oplus \cdots \oplus \{b_j^C\}_{j=1}^{S^g}\} \sim P^g,
\end{aligned}
\qquad (7)
$$

where S^g is the GT sampling number for each class, P^g is the source GT pool containing labelled boxes from all

source point clouds. We gradually *increase* S^r and *reduce* S^g during the self training, that enables 3D detector to smoothly generalize to the target domain. Finally, we concatenate the sampled target pseudo-labeled RED boxes and source GT boxes to the initial pseudo labels, including the points X^r and X^g inside target RED boxes and source GT boxes, respectively:

$$\begin{aligned} \widehat{Y}^t &\leftarrow \widehat{Y}^t \oplus \widehat{Y}^r \oplus Y^g, \\ X^t &\leftarrow X^t \oplus X^r \oplus X^g. \end{aligned} \tag{8}$$

Finally, we have each target point cloud augmented with class-balanced REDB pseudo labels as $(X_i^t, \widehat{Y}_i^t)$, and train the detector as:

$$F(\,\cdot\,;\theta^t) \leftarrow \{(X_i^t, \widehat{Y}_i^t)\}_{i=1}^{N_t}. \tag{9}$$

We alternatively generate target pseudo labels (Stage 2) and self-train the model (Stage 3), until the 3D detector fully is optimized, as illustrated in Fig 2. The detailed Algorithm is summarized in the supplementary materials.

4. Experiments

4.1. Experimental Setup

Datasets. Experiments are carried out on three widely used LiDAR 3D object detection datasets: KITTI [15], Waymo [52], and nuScenes [3]. We follow [70, 71] to address different realistic scenarios of 3D domain adaptation: (i) Adaptation across domains with object shifts (*e.g.*, scale, point density), (ii) Adaptation across domains with environmental shift (*e.g.*, data generation regions, LiDAR beams, angels and range).

Baseline Methods. We compare the proposed method with baselines using voxel-based backbone (*e.g.*, SECOND). (i) SOURCE ONLY refers to the pre-trained model from the source domain evaluated directly on the target domain without adaptations; (ii) SN [62] takes size statistics of target objects and rescale source objects during pre-training; (iii) ST3D [70] is the pioneering pseudo labelling-based method that does not require any knowledge about backbone models and dataset statistics; (iv) ST3D++ [71] is the extended version of ST3D by domain-specific batch normalization [4], achieving the SOTA performance. (v) ORACLE means a fully supervised model trained on the target domain. We also compare with baselines that only work for the point-based backbone (*e.g.*, PointRCNN). (vi) SF-UDA$^{3\text{D}}$ [46] seeks the best scale for pseudo-labeling from consecutive frames by estimating motion coherence. (vii) MLC-NET [36] is a mean-teacher [54] framework aiming to stable optimize the self-training.

Implementation Details. Unlike the previous approach, we pre-trained and self-trained a single 3D detection backbone for all categories, simultaneously, rather than unfairly

train a single-class model. Our code is built on the point cloud detection codebase OpenPCDet [55] and runs on two NVIDIA Tesla V100 GPUs. For both backbones, we set the batch size to 4 for each GPU. We set hyperparameter $\delta_{\text{cde}} = 0.6$, $\delta_{\text{pos}} = 0.6$, $\delta_{\text{obc}} = 0.3$, $d = 5$ and epochs to 90 for all tasks. Regarding the class-balanced sampling, we set $S^r = 5$ and $S^g = 10$, then increase S^r and decrease S^g by 2, at each round of pseudo label generation. The implementation code will be available. To fairly evaluate baselines and the proposed method, we utilize the KITTI evaluation metric for evaluation on the three classes: car (equivalent to the vehicle for similar classes in the Waymo), pedestrians and cyclists (equivalent to bicyclist and motorcyclist in nuScenes). Based on the official KITTI evaluation metric, we report the average precision for each class, in both 3D (*i.e.*, AP$_{3\text{D}}$) and the bird's eye view (*i.e.*, AP$_{\text{BEV}}$) over 40 recall positions, with IoU threshold 0.7 for cars and 0.5 for pedestrians and cyclists. We also calculate the mean AP for all classes in 3D and BEV, denoted as mAP$_{3\text{D}}$ and mAP$_{\text{BEV}}$, respectively.

4.2. Main Results and Analysis

We conducted comprehensive experiments on three different 3D adaptation tasks under a multi-class training setting, as summarized and reported in Tab 1. We can clearly observe that REDB consistently improves the performance on Waymo $\rightarrow$ KITTI and nuScenes $\rightarrow$ KITTI by a large margin of 21.27% and 17.13% in terms of mAP$_{3\text{D}}$, which largely close the performance gap between SOURCE ONLY and ORACLE. When comparing with the latest SOTA method ST3D++, our REDB exhibits superior performance in terms of mAP$_{\text{BEV}}$ for all three adaptation tasks, with improvements of 2.9%, 12.5%, and 12.6%, respectively. It is worth noting that the proposed REDB gains significantly more performance for the last two tasks (*i.e.*, Waymo $\rightarrow$ nuScenes and nuScenes $\rightarrow$ KITTI) than the first task (*i.e.*, Waymo $\rightarrow$ KITTI), indicating the REDB is more effective for adapting to 3D scenes with larger environmental gaps. It is further evident that the REDB method stands out for its well-balanced performance across all categories, while all baselines are biased towards the most frequently appearing class (*i.e.* car) and underperformed in rare classes (*i.e.* pedestrian and cyclist). Overall, the proposed REDB excels all baselines on both mAP$_{3\text{D}}$ and mAP$_{\text{BEV}}$ across all scenarios of 3D adaptation tasks.

Performance Analysis under Difficult Conditions. To corroborate the effectiveness of the REDB approach against the baseline methods, we present additional experimental results in Tab 2, using a more challenging evaluation metric: the average precision (AP) at **Hard** difficulty level defined by KITTI dataset. It is observed that the proposed REDB outperforms the SOTA baseline ST3D++ by 5.81% of mAP$_{3\text{D}}$ for adapting from Waymo to KITTI. Concern-

Table 1: Experiment results of three adaptation tasks are presented, with the reported average precision (AP) for bird's-eye view (AP$_{BEV}$) / 3D (AP$_{3D}$) of car, pedestrian, and cyclist with IoU threshold set to 0.7, 0.5, and 0.5 respectively. When the KITTI is the target domain, we report the AP at **Moderate** difficulty. The last column shows the mean AP for all classes. We indicate the best adaptation result by **bold** and we highlight the row representing our method.

TASK	METHOD	CAR	PEDESTRIAN	CYCLIST	MEAN AP
Waymo → KITTI	SOURCE ONLY	51.48 / 19.78	40.80 / 31.26	47.63 / 35.45	46.64 / 28.83
	SN	76.61 / **54.14**	**52.48 / 48.20**	34.56 / 32.74	54.55 / 45.03
	ST3D	77.62 / 49.24	44.45 / 42.04	47.74 / 42.95	56.60 / 44.70
	ST3D++	77.68 / 50.03	49.09 / 46.19	51.50 / 47.70	59.42 / 47.97
	REDB	**80.37** / 54.12	51.01 / **48.20**	**52.05 / 47.97**	**61.14 / 50.10**
	ORACLE	83.29 / 73.45	46.64 / 41.33	62.92 / 60.32	64.28 / 58.37
Waymo → nuScenes	SOURCE ONLY	**30.64** / 17.18	1.81 / 0.58	0.97 / 0.88	11.14 / 6.22
	SN	29.56 / 17.78	1.33 / 1.07	2.92 / 2.61	11.27 / 7.15
	ST3D	28.42 / 17.83	1.64 / 1.39	4.01 / 3.54	11.36 / 7.59
	ST3D++	28.87 / **19.15**	1.82 / 1.58	4.09 / 3.74	11.60 / 8.16
	REDB	30.12 / 18.56	**2.47 / 2.14**	**6.56 / 5.19**	**13.05 / 8.63**
	ORACLE	51.88 / 34.87	25.24 / 18.92	15.06 / 11.73	30.73 / 21.84
nuScenes → KITTI	SOURCE ONLY	39.15 / 7.65	21.54 / 16.87	6.31 / 2.44	22.33 / 8.98
	SN	56.08 / 28.67	23.05 / 16.84	5.11 / 2.32	28.08 / 15.94
	ST3D	71.50 / 48.09	22.64 / 17.61	7.86 / 5.20	34.00 / 23.64
	ST3D++	69.90 / 44.62	24.11 / 18.20	10.14 / 6.39	33.75 / 21.64
	REDB	**74.23 / 51.31**	**25.95 / 18.38**	**13.82 / 8.64**	**38.00 / 26.11**
	ORACLE	83.29 / 73.45	46.64 / 41.33	62.92 / 60.32	64.28 / 58.37

Table 2: Experiment results at **Hard** difficulty of adapting to the KITTI dataset. We report average precision (AP) for bird's-eye view (AP$_{BEV}$) / 3D (AP$_{3D}$) of car, pedestrian, and cyclist with IoU threshold set to 0.7, 0.5, and 0.5 respectively. The last column shows the mean AP for all classes. We indicate the best adaptation result by **bold**, and we highlight the row representing our method.

TASK	METHOD	CAR	PEDESTRIAN	CYCLIST	MEAN AP
Waymo → KITTI	SN	75.57 / **53.87**	**49.15 / 45.58**	33.09 / 31.19	52.60 / 43.55
	ST3D	75.31 / 48.01	41.42 / 43.02	45.76 / 40.76	54.16 / 43.93
	ST3D++	75.54 / 48.03	45.63 / 42.00	47.12 / 43.70	56.10 / 44.58
	REDB	**78.24** / 52.56	46.01 / 43.22	**49.72 / 45.73**	**57.99 / 47.17**
nuScenes → KITTI	SN	54.29 / 25.81	21.43 / 15.09	5.13 / 2.31	26.95 / 14.40
	ST3D	68.21 / 43.44	21.30 / 15.96	7.33 / 5.08	32.28 / 21.49
	ST3D++	68.29 / 40.85	22.16 / 16.31	7.32 / 3.91	32.59 / 20.36
	REDB	**69.76 / 46.17**	**25.01 / 18.88**	**9.54 / 6.15**	**34.77 / 23.73**

ing a more challenging adaptation task (*i.e.*, nuScenes → KITTI) involving significant environmental shifts in beam numbers, angles and the point cloud range, the proposed ReDB exhibits superior performance over all other baseline models across all categories: 16.55% of mAP$_{3D}$ than the SOTA baseline. Thus, our approach surpasses the baselines by a large margin when assessed with KITTI metric at **Hard** difficulty, indicating the efficacy of the ReDB in facilitating effective generalization of 3D detectors to difficult target objects.

4.3. Component Analysis

Ablation Study. To investigate the impact of the derived CDE-based reliable pseudo label generation, OBC-based downsampling and class-balanced self-training, we compare five variants of the ReDB model in the task of adapting Waymo to KITTI, shown in Tab 4. GT indicates the source ground truth sampling and PS indicates the target pseudo label sampling. For fairness, we keep the same hyperparameters for all ablation variants. To explore the effectiveness of the proposed CDE and OBC modules, we remove each of

Table 3: Performance comparisons (AP_{BEV} and AP_{3D}) with PointRCNN backbone on nuScenes → KITTI task. ‡ indicates the results reported in the original paper, where only a single category was adapted.

	Method	CAR EASY	MOD.	HARD	PEDESTRIAN EASY	MOD.	HARD	CYCLIST EASY	MOD.	HARD	AVERAGE EASY	MOD.	HARD
3D	SOURCE ONLY	42.77	32.11	28.75	43.26	37.29	33.16	3.28	4.09	4.18	29.77	24.60	22.03
	SN	66.56	50.32	45.92	42.96	37.15	32.45	9.07	7.57	7.42	39.53	31.68	28.60
	ST3D	48.85	41.90	38.92	45.67	38.71	33.09	26.50	19.35	18.38	40.34	33.32	30.13
	ST3D++	60.45	49.36	45.88	50.77	42.43	36.64	27.20	17.94	16.52	46.14	36.58	33.01
	SF-UDA^{3D} ‡	68.80	49.80	45.00	-	-	-	-	-	-	-	-	-
	MLC-NET ‡	71.30	55.40	49.00	-	-	-	-	-	-	-	-	-
	REDB	**71.45**	**57.9**	**53.91**	**52.32**	**44.33**	**37.95**	**45.13**	**32.93**	**31.05**	**56.30**	**45.05**	**40.97**
BEV	SOURCE ONLY	80.91	60.75	56.00	45.84	39.78	35.45	3.38	4.45	4.57	43.37	34.99	32.00
	SN	81.17	63.34	56.80	44.80	38.73	34.39	9.65	9.39	8.93	45.21	37.15	33.67
	ST3D	85.61	69.48	64.52	47.31	39.96	35.05	32.60	23.06	21.73	55.17	44.17	40.43
	ST3D++	85.81	69.17	64.74	52.68	45.10	39.30	31.30	20.29	18.42	56.59	44.85	40.82
	REDB	**91.50**	**76.01**	**71.42**	**54.83**	**45.87**	**39.34**	**48.44**	**35.49**	**33.23**	**64.93**	**52.46**	**48.00**

Table 4: Ablative study of different modules on adapting Waymo to KITTI task.

GT	PS	CDE	OBC	3D DETECTION mAP EASY	MODERATE	HARD	BEV DETECTION mAP EASY	MODERATE	HARD
-	-	-	-	48.41	42.26	40.18	60.68	53.52	51.01
√	-	-	-	54.10	45.21	42.42	63.13	56.73	54.00
-	√	-	-	51.92	46.04	43.68	62.57	54.57	51.32
-	√	√	-	54.04	47.07	44.90	64.70	58.15	55.69
-	√	-	√	53.99	47.66	44.65	65.72	58.51	55.36
√	√	√	√	**56.52**	**50.10**	**47.17**	**67.36**	**61.14**	**57.99**

them and observe a dramatic mAP_{3D} drop of 4.9% (row-5) and 5.34% (row-4) compared to the full model (row-6), respectively. To validate the proposed class-balanced self-training, we find when comparing to the non-sampling baseline (row-1), the source ground truth sampling (row-2) and the target pseudo label sampling (row-3) increase mAP_{3D} of 7% and 8.9%, respectively. It is noticeable that, source ground truth sampling (row-2) produces remarkable mAP_{3D} gains at easy level (5.69%) but very limited gains at moderate and hard levels (2.95% and 2.24%), caused by a lack of knowledge about diverse geometrical features in the target domain. Therefore, removing any module from the proposed REDB induces a clear drop in 3D / BEV mAP scores, confirming the importance of each component that contributes to multi-class 3D domain adaptation.

Sensitivity to Detector Architecture. To validate the sensitivity of performance to choices of voxel-based and point-based detectors, we plug the proposed REDB paradigm into PointRCNN. As reported in Tab 3, our approach consistently outperforms the latest SOTA method ST3D++ by 23.32% and 24.11% regarding mAP_{3D} at moderate and hard levels, respectively. Notably, our multi-class method has even achieved superior performance (10.02% and 19.8%)

compared to MLC-NET and SF-UDA^{3D}, both of which were under the unfair single-class training setting and are exclusively designed for the point-based 3D detector. The outstanding results of the proposed method using a variety of 3D detectors, demonstrate the REDB is a model-agnostic and insensitive to detector architecture.

Sensitivity to the hyperparameters δ_{cde} and d. We show the sensitivity of our approach to varying the CDE IoU threshold δ_{cde} and diversity downsampling rate d in Fig 7. We vary the value of δ_{cde} and d, from $[0.3, 0.6]$ and $[2, 10]$, respectively, while fixing the other configurations to the default setting. We can observe that, when δ_{cde} varies, there is only 2.69% and 2.07% fluctuation on mAP_{3D} and mAP_{BEV} scores, respectively. Regarding the variation of d, we notice when d is set as a large value (e.g., $d = 10$), the performance drops within an acceptable range: 1.37% in mAP_{3D} and 2.57% in mAP_{BEV}. Such a small drop could be caused by the insufficient scale of the down-sampled RED pool. Hence, selecting a proper value of downsampling rate (e.g., $d = 5$), reaches the best results. Little fluctuation caused by varying δ_{cde} and d evidence that, the proposed method is robust to different choices of hyperparameters.
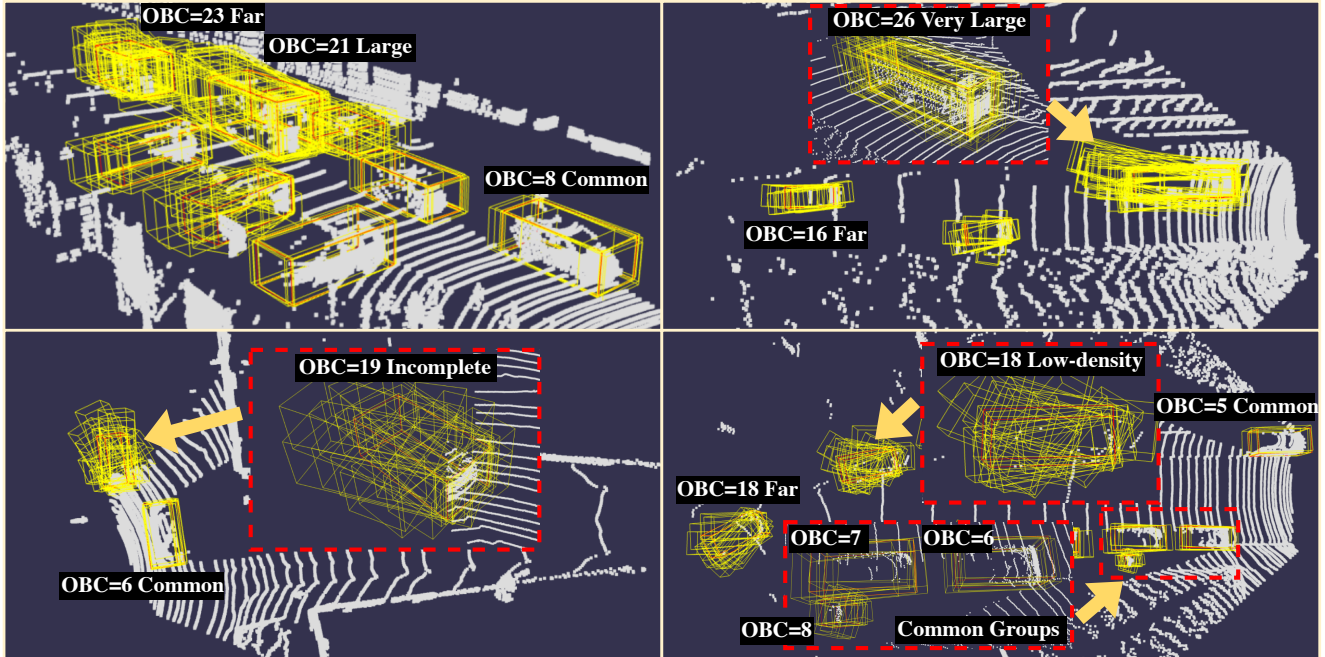
Figure 6: Case studies of the overlapped boxes counting. We visualize the predicted boxes before NMS (in yellow) and its count, with the corresponding geometric description. By inverse frequency sampling of OBCs, geometrical diverse objects are added to the RED Pool.
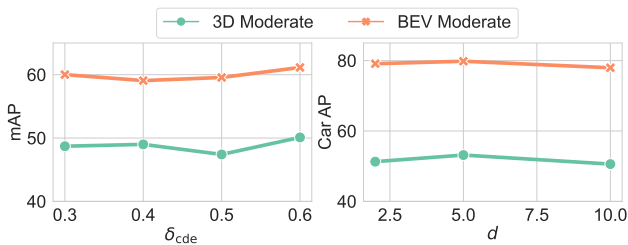

Figure 7: Hyperparameter sensitivity of δ_{cde} and d.

Visualization of OBC-based Diversity. To intuitively understand the merits of our proposed diversity module, Fig 6 visualizes the counted predicted boxes before NMS for different pseudo-labeled target objects. We can see that most objects with small OBC values (*e.g.*, between 5 and 8) are, (1) typically closer to the LiDAR sensor, (2) with a complete object shape, (3) and commonly small-sized objects. These objects usually have highly similar and complete geometric features, making up the majority of the dataset. In contrast, objects with high OBCs are usually diverse in one or more aspects of geometrical representations. Regarding the object size, large-sized objects tend to yield high OBC scores (*e.g.*, 21 and 26). In addition to object volume, we can also find objects that are obviously far away from the LiDAR center can produce higher OBC values (from 16 to 23), and low-density and heavily occluded objects also

have high OBC values of 18 and 19, respectively. More statistical analysis can be found in supplementary materials. Therefore, the proposed OBC metric can effectively quantify the diversity of pseudo-labels in multiple dimensions of geometric features, aiding the 3D detector to learn a more diverse distribution of target objects, thereby alleviating multidimensional object shift.

5. Conclusion

This paper revisits the setting of unsupervised domain-adaptive 3D detection and investigates the generation of reliable, diverse, and class-balanced pseudo labels for effective multi-class adaptation. The quantitative and qualitative analysis demonstrates the effectiveness of the derived cross-domain examination and OBC-based downsampling strategies, allowing pseudo labels that are robust to environmental and object shifts to be preserved. The class-balanced self-training helps detectors to learn frequently appearing and rare classes simultaneously. The proposed approach is proven to be versatile and can accommodate voxel-based and point-based 3D detectors.

6. Acknowledgments

References

[1] Syeda Mariam Ahmed, Yan Zhi Tan, Chee-Meng Chew, Abdullah Al Mamun, and Fook Seng Wong. Edge and corner detection for unorganized 3d point clouds with application to robotic welding. In *Proc. International Conference on Intelligent Robots and Systems (IROS)*, pages 7350–7355, 2018. 1

[2] Eduardo Arnold, Omar Y. Al-Jarrah, Mehrdad Dianati, Saber Fallah, David Oxtoby, and Alex Mouzakitis. A survey on 3d object detection methods for autonomous driving applications. *IEEE Transactions on Intelligent Transportation Systems*, 20(10):3782–3795, 2019. 1

[3] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11618–11628, 2020. 2, 6

[4] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7354–7362, 2019. 6

[5] Chao Chen, Zhihang Fu, Zhihong Chen, Sheng Jin, Zhaowei Cheng, Xinyu Jin, and Xian-Sheng Hua. Homm: Higher-order moment matching for unsupervised domain adaptation. In *Proc. Conference on Artificial Intelligence (AAAI)*, pages 3422–3429, 2020. 3

[6] Zhuoxiao Chen, Yadan Luo, and Mahsa Baktashmotlagh. Conditional extreme value theory for open set video domain adaptation. In *Proc. ACM International Conference on Multimedia in Asia (MM Asia)*, 2022. 3

[7] Robert DeBortoli, Fuxin Li, Ashish Kapoor, and Geoffrey A. Hollinger. Adversarial training on point clouds for sim-to-real 3d object detection. *IEEE Robotics and Automation Letters*, 6(4):6662–6669, 2021. 3

[8] Boyang Deng, Charles R. Qi, Mahyar Najibi, Thomas A. Funkhouser, Yin Zhou, and Dragomir Anguelov. Revisiting 3d object detection from an egocentric perspective. In *Proc. Annual Conference on Neural Information Processing (NeurIPS)*, pages 26066–26079, 2021. 1

[9] 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 *Proc. Conference on Artificial Intelligence (AAAI)*, pages 1201–1209, 2021. 3

[10] Zhijie Deng, Yucen Luo, and Jun Zhu. Cluster alignment with a teacher for unsupervised domain adaptation. In *Proc. International Conference on Computer Vision (ICCV)*, pages 9943–9952, 2019. 3

[11] Martin Engelcke, Dushyant Rao, Dominic Zeng Wang, Chi Hay Tong, and Ingmar Posner. Vote3deep: Fast object detection in 3d point clouds using efficient convolutional neural networks. In *Proc. International Conference on Robotics and Automation (ICRA)*, pages 1355–1361, 2017. 3

[12] Lue Fan, Ziqi Pang, Tianyuan Zhang, Yu-Xiong Wang, Hang Zhao, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. Embracing single stride 3d object detector with sparse transformer. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8448–8458, 2022. 3

[13] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proc. International Conference on Machine Learning (ICML)*, volume 37, pages 1180–1189, 2015. 3

[14] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17:59:1–59:35, 2016. 3

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012. 1, 6

[16] Qiqi Gu, Qianyu Zhou, Minghao Xu, Zhengyang Feng, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. PIT: position-invariant transform for cross-fov domain adaptation. In *Proc. International Conference on Computer Vision (ICCV)*, pages 8741–8750, 2021. 3

[17] Martin Hahner, Christos Sakaridis, Mario Bijelic, Felix Heide, Fisher Yu, Dengxin Dai, and Luc Van Gool. Lidar snowfall simulation for robust 3d object detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16343–16353, 2022. 3

[18] Chenhang He, Ruihuang Li, Shuai Li, and Lei Zhang. Voxel set transformer: A set-to-set approach to 3d object detection from point clouds. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8407–8417, 2022. 3

[19] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *Proc. International Conference on Machine Learning (ICML)*, volume 80, pages 1994–2003, 2018. 3

[20] Jordan S. K. Hu, Tianshu Kuai, and Steven L. Waslander. Point density-aware voxels for lidar 3d object detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8459–8468, 2022. 3

[21] Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Re-energizing domain discriminator with sample relabeling for adversarial domain adaptation. In *Proc. International Conference on Computer Vision (ICCV)*, pages 9154–9163, 2021. 3

[22] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12697–12705, 2019. 3

[23] Alexander Lehner, Stefano Gasperini, Alvaro Marcos-Ramiro, Michael Schmidt, Mohammad-Ali Nikouei Mahani, Nassir Navab, Benjamin Busam, and Federico Tombari. 3d-vfield: Adversarial augmentation of point clouds for domain generalization in 3d object detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17274–17283, 2022. 3

[24] Bo Li. 3d fully convolutional network for vehicle detection in point cloud. In *Proc. International Conference on Intelligent Robots and Systems (IROS)*, pages 1513–1518, 2017. 3

[25] Buyu Li, Wanli Ouyang, Lu Sheng, Xingyu Zeng, and Xiaogang Wang. GS3D: an efficient 3d object detection framework for autonomous driving. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1019–1028, 2019. 1

[26] Shuang Li, Binhui Xie, Qiuxia Lin, Chi Harold Liu, Gao Huang, and Guoren Wang. Generalized domain conditioned adaptation network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4093–4109, 2022. 3

[27] Xiaofeng Liu, Zhenhua Guo, Site Li, Fangxu Xing, Jane You, C.-C. Jay Kuo, Georges El Fakhri, and Jonghye Woo. Adversarial unsupervised domain adaptation with conditional and label shift: Infer, align and iterate. In *Proc. International Conference on Computer Vision (ICCV)*, pages 10347–10356, 2021. 3

[28] Mingsheng Long, Yue Cao, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Transferable representation learning with deep adaptation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12):3071–3085, 2019. 3

[29] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *Proc. International Conference on Machine Learning (ICML)*, volume 37, pages 97–105, 2015. 3

[30] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Unsupervised domain adaptation with residual transfer networks. In *Proc. Annual Conference on Neural Information Processing (NeurIPS)*, pages 136–144, 2016. 3

[31] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks. In *Proc. International Conference on Machine Learning (ICML)*, volume 70, pages 2208–2217, 2017. 3

[32] Yadan Luo, Zhuoxiao Chen, Zijian Wang, Xin Yu, Zi Huang, and Mahsa Baktashmotlagh. Exploring active 3d object detection from a generalization perspective. In *Proc. International Conference on Learning Representations (ICLR)*, 2023. 1

[33] Yadan Luo, Zi Huang, Zijian Wang, Zheng Zhang, and Mahsa Baktashmotlagh. Adversarial bipartite graph learning for video domain adaptation. In *Proc. ACM International Conference on Multimedia (MM)*, pages 19–27, 2020. 3

[34] Yadan Luo, Zijian Wang, Zhuoxiao Chen, Zi Huang, and Mahsa Baktashmotlagh. Source-free progressive graph learning for open-set domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11240–11255, 2023. 3

[35] Yadan Luo, Zijian Wang, Zi Huang, and Mahsa Baktashmotlagh. Progressive graph learning for open-set domain adaptation. In *Proc. International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 6468–6478, 2020. 3

[36] 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 *Proc. International Conference on Computer Vision (ICCV)*, pages 8846–8855, 2021. 1, 2, 3, 6

[37] Jiageng Mao, Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. 3d object detection for autonomous driving: A review and new outlooks. *CoRR*, abs/2206.09474, 2022. 1

[38] Scott McCrae and Avideh Zakhor. 3d object detection for autonomous driving using temporal lidar data. In *Proc. International Conference on Image Processing (ICIP)*, pages 2661–2665, 2020. 1

[39] Gregory P. Meyer, Ankit Laddha, Eric Kee, Carlos Vallespi-Gonzalez, and Carl K. Wellington. Lasernet: An efficient probabilistic 3d object detector for autonomous driving. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12677–12686, 2019. 1

[40] Hector A. Montes, Justin Le Louedec, Grzegorz Cielniak, and Tom Duckett. Real-time detection of broccoli crops in 3d point clouds for autonomous robotic harvesting. In *Proc. International Conference on Intelligent Robots and Systems (IROS)*, pages 10483–10488, 2020. 1

[41] Xuran Pan, Zhuofan Xia, Shiji Song, Li Erran Li, and Gao Huang. 3d object detection with pointformer. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7463–7472, 2021. 3

[42] Rui Qian, Xin Lai, and Xirong Li. 3d object detection for autonomous driving: A survey. *Pattern Recognition*, 130:108796, 2022. 1

[43] Christoph B Rist, Markus Enzweiler, and Dariu M Gavrila. Cross-sensor deep domain adaptation for lidar detection and segmentation. In *Proc. Intelligent Vehicles Symposium, (IV)*, pages 1535–1542, 2019. 3

[44] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3723–3732, 2018. 3

[45] Khaled Saleh, Ahmed Abobakr, Mohammed Hassan Attia, Julie Iskander, Darius Nahavandi, Mohammed Hossny, and Saeid Nahavandi. Domain adaptation for vehicle detection from bird's eye view lidar point cloud data. In *Proc. International Conference on Computer Vision*, pages 3235–3242, 2019. 3

[46] 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 *Proc. International Conference on 3D Vision (3DV)*, pages 771–780, 2020. 6

[47] Guangsheng Shi, Ruifeng Li, and Chao Ma. Pillarnet: Real-time and high-performance pillar-based 3d object detection. In *Proc. European Conference on Computer Vision (ECCV)*, volume 13670, pages 35–52, 2022. 3

[48] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointr-cnn: 3d object proposal generation and detection from point cloud. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–779, 2019. 3

[49] 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, 2021. 3

[50] Weijing Shi and Raj Rajkumar. Point-gnn: Graph neural network for 3d object detection in a point cloud. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1708–1716, 2020. 3

[51] Baochen Sun and Kate Saenko. Deep CORAL: correlation alignment for deep domain adaptation. In *Proc. European Conference on Computer Vision Workshops (ECCV Workshops)*, volume 9915 of *Lecture Notes in Computer Science*, pages 443–450, 2016. 3

[52] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2443–2451, 2020. 1, 2, 6

[53] Pei Sun, Mingxing Tan, Weiyue Wang, Chenxi Liu, Fei Xia, Zhaoqi Leng, and Dragomir Anguelov. Swformer: Sparse window transformer for 3d object detection in point clouds. In *Proc. European Conference on Computer Vision (ECCV)*, volume 13670, pages 426–442, 2022. 3

[54] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proc. Annual Conference on Neural Information Processing (NeurIPS)*, pages 1195–1204, 2017. 6

[55] OpenPCDet Development Team. Openpcdet: An open-source toolbox for 3d object detection from point clouds. https://github.com/open-mmlab/OpenPCDet, 2020. 6

[56] Jonathan Tremblay, Thang To, and Stan Birchfield. Falling things: A synthetic dataset for 3d object detection and pose estimation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, pages 2038–2041, 2018. 1

[57] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2962–2971, 2017. 3

[58] Haiyang Wang, Shaoshuai Shi, Ze Yang, Rongyao Fang, Qi Qian, Hongsheng Li, Bernt Schiele, and Liwei Wang. Rbgnet: Ray-based grouping for 3d object detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1100–1109, 2022. 3

[59] Jun Wang, Shiyi Lan, Mingfei Gao, and Larry S. Davis. Infofocus: 3d object detection for autonomous driving with dynamic information modeling. In *Proc. European Conference on Computer Vision (ECCV)*, volume 12355, pages 405–420. Springer, 2020. 1

[60] Li Wang, Ruifeng Li, Jingwen Sun, Xingxing Liu, Lijun Zhao, Hock Soon Seah, Chee Kwang Quah, and Budianto Tandianus. Multi-view fusion-based 3d object detection for robot indoor scene perception. *Sensors*, 19(19):4092, 2019. 1

[61] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark E. Campbell, and Kilian Q. Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8445–8453, 2019. 1

[62] Yan Wang, Xiangyu Chen, Yurong You, Li Erran Li, Bharath Hariharan, Mark E. Campbell, Kilian Q. Weinberger, and Wei-Lun Chao. Train in germany, test in the USA: making 3d object detectors generalize. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11710–11720, 2020. 1, 3, 6

[63] Yue Wang, Alireza Fathi, Abhijit Kundu, David A. Ross, Caroline Pantofaru, Thomas A. Funkhouser, and Justin Solomon. Pillar-based object detection for autonomous driving. In *Proc. European Conference on Computer Vision (ECCV)*, volume 12367 of *Lecture Notes in Computer Science*, pages 18–34, 2020. 3

[64] Zijian Wang, Yadan Luo, Zi Huang, and Mahsa Baktashmotlagh. Prototype-matching graph network for heterogeneous domain adaptation. In *Proc. ACM International Conference on Multimedia (MM)*, pages 2104–2112, 2020. 3

[65] Zixin Wang, Yadan Luo, Peng-Fei Zhang, Sen Wang, and Zi Huang. Discovering domain disentanglement for generalized multi-source domain adaptation. In *Proc. International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2022. 3

[66] 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. In *Proc. European Conference on Computer Vision (ECCV)*, volume 13699 of *Lecture Notes in Computer Science*, pages 179–195, 2022. 3

[67] 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 *Proc. International Conference on Computer Vision (ICCV)*, pages 15426–15436, 2021. 3

[68] Yan Yan, Yuxing Mao, and Bo Li. SECOND: sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 3

[69] Bin Yang, Wenjie Luo, and Raquel Urtasun. PIXOR: real-time 3d object detection from point clouds. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7652–7660, 2018. 3

[70] Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. ST3D: self-training for unsupervised domain adaptation on 3d object detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10368–10378, 2021. 1, 2, 3, 6

[71] Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. St3d++: denoised self-training for unsupervised domain adaptation on 3d object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2, 3, 6

[72] Zetong Yang, Yanan Sun, Shu Liu, and Jiaya Jia. 3dssd: Point-based 3d single stage object detector. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11037–11045, 2020. 3

[73] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. STD: sparse-to-dense 3d object detector for point cloud. In *Proc. International Conference on Computer Vision (ICCV)*, pages 1951–1960, 2019. 3

[74] Cang Ye and Xiangfei Qian. 3-d object recognition of a robotic navigation aid for the visually impaired. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(2):441–450, 2018. 1

[75] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3d object detection and tracking. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11784–11793, 2021. 3

[76] Yurong You, Katie Luo, Cheng Perng Phoo, Wei-Lun Chao, Wen Sun, Bharath Hariharan, Mark E. Campbell, and Kilian Q. Weinberger. Learning to detect mobile objects from lidar scans without labels. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1120–1130, 2022. 2

[77] Yurong You, Cheng Perng Phoo, Katie Luo, Travis Zhang, Wei-Lun Chao, Bharath Hariharan, Mark Campbell, and Kilian Q. Weinberger. Unsupervised adaptation from repeated traversals for autonomous driving. In *Proc. Annual Conference on Neural Information Processing (NeurIPS)*, 2022. 2

[78] Yurong You, Yan Wang, Wei-Lun Chao, Divyansh Garg, Geoff Pleiss, Bharath Hariharan, Mark E. Campbell, and Kilian Q. Weinberger. Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. In *Proc. International Conference on Learning Representations (ICLR)*, 2020. 1

[79] Yihan Zeng, Chunwei Wang, Yunbo Wang, Hang Xu, Chaoqiang Ye, Zhen Yang, and Chao Ma. Learning transferable features for point cloud detection via 3d contrastive co-training. In *Proc. Annual Conference on Neural Information Processing (NeurIPS)*, pages 21493–21504, 2021. 2, 3

[80] Weichen Zhang, Wen Li, and Dong Xu. SRDAN: scale-aware and range-aware domain adaptation network for cross-dataset 3d object detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6769–6779, 2021. 2, 3

[81] Yifan Zhang, Qingyong Hu, Guoquan Xu, Yanxin Ma, Jianwei Wan, and Yulan Guo. Not all points are equal: Learning highly efficient point-based detectors for 3d lidar point clouds. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18931–18940, 2022. 3

[82] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4490–4499, 2018. 3

[83] Zhengxue Zhou, Leihui Li, Alexander Fürsterling, Hjalte Joshua Durocher, Jesper Mouridsen, and Xuping Zhang. Learning-based object detection and localization for a mobile robot manipulator in SME production. *Robotics and Computer-Integrated Manufacturing*, 73:102229, 2022. 1