2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

WLST: Weak Labels Guided Self-training for Weakly-supervised Domain Adaptation on 3D Object Detection

Tsung-Lin Tsou[1], Tsung-Han Wu[1], and Winston H. Hsu[1,2]

***Abstract*— In the field of domain adaptation (DA) on 3D object detection, most of the work is dedicated to unsupervised domain adaptation (UDA). Yet, without any target annotations, the performance gap between the UDA approaches and the fully-supervised approach is still noticeable, which is impractical for real-world applications. On the other hand, weakly-supervised domain adaptation (WDA) is an underexplored yet practical task that only requires few labeling effort on the target domain. To improve the DA performance in a cost-effective way, we propose a general weak labels guided self-training framework, WLST, designed for WDA on 3D object detection. By incorporating autolabeler, which can generate 3D pseudo labels from 2D bounding boxes, into the existing self-training pipeline, our method is able to generate more robust and consistent pseudo labels that would benefit the training process on the target domain. Extensive experiments demonstrate the effectiveness, robustness, and detector-agnosticism of our WLST framework. Notably, it outperforms previous state-of-the-art methods on all evaluation tasks. Code and models are available at `https://github.com/jacky121298/WLST`. Note that the complete version with appendix is available on arXiv.**

I. Introduction

With the rapid development of 3D range sensors (*e.g.* LiDAR point clouds) and large-scale human-annotated datasets [1]–[3], 3D object detection in the field of autonomous driving has garnered great attention and obtained remarkable breakthroughs [4]–[10]. In order to deploy to real roads, 3D detectors must adapt to various real-world scenarios and perform robustly against numerous domain shifts arising from different settings of 3D range sensors, fickle weather conditions, miscellaneous objects in the driving scene, etc. However, existing 3D detectors are inadequate to tackle the domain gap realistically. Past work [11] has shown that the performance of a fully-supervised 3D detector trained on Waymo Open Dataset [3] dropped drastically when evaluated on KITTI Benchmark Dataset [2]. Therefore, developing an effective *domain adaptation (DA)* approach is needed.

In the field of DA on 3D object detection, most of the work [12]–[15] is dedicated to *unsupervised domain adaptation (UDA)*. Among them, the self-training approaches [13], [14] perform the best. They redesigned the naive self-training pipeline to improve the pseudo-label selection mechanism and utilize effective augmentation techniques in the model training process, which achieved state-of-the-art performance in many DA tasks. Yet, without any target annotations, the performance gap between the UDA approaches (*e.g.* 64.75 AP_{3D} on the Waymo $\rightarrow$ KITTI task) and the fully-supervised oracle approach (*e.g.* 83.00 AP_{3D} in the KITTI dataset) is still noticeable as shown in Tab. I. On the other hand, few work has been contributed to *weakly-supervised domain adaptation (WDA)*, among which SN [11] utilizes object size statistics of the target domain to mitigate the domain shifts. However, its effectiveness largely depends on object size distributions and performs even worse than the UDA approaches (*e.g.* 62.54 AP_{3D} on the Waymo $\rightarrow$ KITTI task). In summary, the above approaches are impractical for real-world applications.

To reduce such performance gaps in a cost-effective way, we propose a general weak labels guided self-training framework, WLST, designed for WDA on 3D object detection. Building upon the success of self-training UDA approaches [13], [14] and studies [16], [17] on autolabeler that can generate 3D pseudo labels from 2D bounding boxes, our WLST framework incorporates autolabeler into the existing self-training pipeline. Specifically, as shown in Fig. 1, a 3D detector and an autolabeler are first pre-trained on the labeled source domain. Then, pseudo labels would be generated by both models on the weak-labeled target domain. Finally, the 3D detector and autolabeler are iteratively improved by alternatively conducting pseudo-label generation and model re-training on these pseudo-labeled target data. Regarding annotation cost, statistics show that the time spent on annotating weak labels (*i.e.* 2D bounding boxes) can be approximately three to sixteen times less than strong labels (*i.e.* 3D bounding boxes) depending on the annotation tool used [18], rendering the cost affordable.

To further enhance the quality of pseudo labels generated by 3D detector and autolabeler, we design a pseudo-label selection mechanism to explore and leverage their distinct pros and cons. To elaborate, autolabeler has higher precision attributed to the fact that 2D bounding boxes help constrain the 3D search space for the pseudo labels as described in Fig. 2. Nevertheless, it works on the object level and couldn't learn the correlation between objects. On the other hand, 3D detector works on the scene level and has a larger Field of View (FoV), which enables a better understanding of the correlation between objects that leads to higher recall (see Fig. 3). Based on these observations, our proposed *consistency fusion strategy* leverages geometric consistency and cross-modality consistency of pseudo labels to retain high precision and high recall simultaneously (see Tab. III).

Extensive experiments on three widely used 3D object detection datasets, nuSenses Dataset [1], KITTI Benchmark Dataset [2], and Waymo Open Dataset [3] demonstrate the effectiveness, robustness, and detector-agnosticism of our WLST framework. It can effectively close the performance

[1]National Taiwan University, [2]Mobile Drive Technology

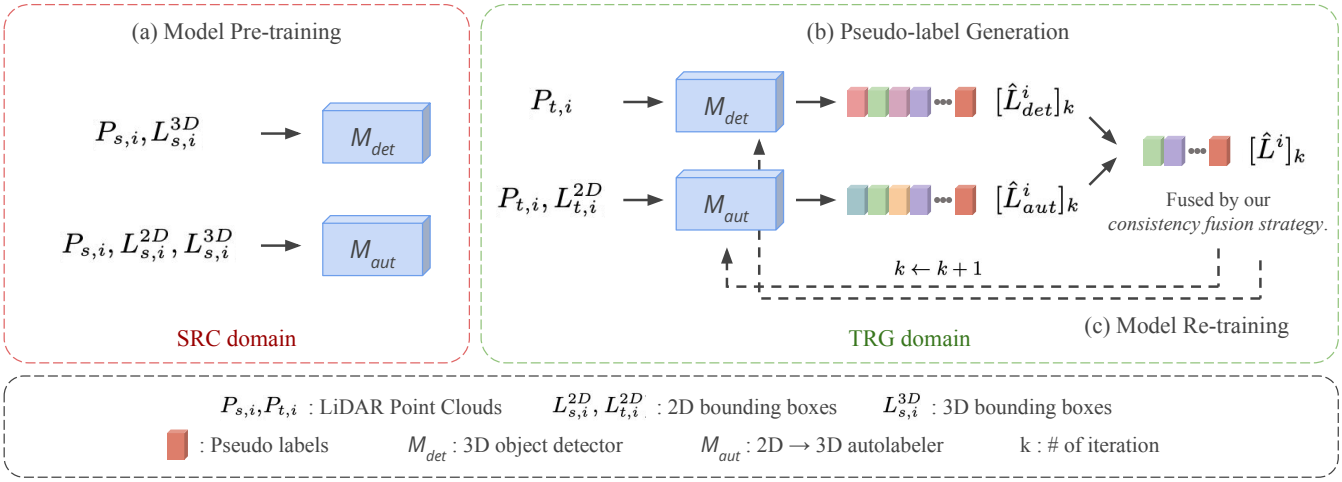

Fig. 1: **Our WLST framework** is composed of three stages. (a) Pre-train 3D detector and autolabeler on the source data. (see Sec. III-B.2) (b) Generate high-quality pseudo labels by our *consistency fusion strategy* on the target data. (see Sec. III-B.3) (c) Re-train 3D detector and autolabeler on the pseudo-labeled target data. (see Sec. III-B.4)

gap between source only approach and fully-supervised oracle approach by up to 87.26% in AP_{BEV} and up to 91.34% in AP_{3D} in Tab. I. Notably, we outperform previous state-of-the-art methods on all evaluation tasks. In summary, our main contributions are threefold:

- We formulate and investigate the problem of WDA on 3D object detection, an underexplored yet practical task that has the potential to improve the DA performance in a cost-effective way.
- We propose WLST, a general weak labels guided self-training framework, to obtain more robust and consistent pseudo labels. To the best of our knowledge, we are the first to incorporate autolabeler into the self-training pipeline.
- Our WLST framework is extensively evaluated on three widely used 3D object detection datasets and outperforms previous state-of-the-art methods on all evaluation tasks.

II. Related Work

LiDAR-based 3D Object Detection. Given the point clouds obtained from LiDAR sensors, 3D detectors aim to recognize and determine the 3D information of the objects, including location, dimension, orientation, and category. Based on data representations, 3D detectors can be divided into point-based, grid-based, and point-voxel-based. Point-based detectors [7], [19]–[21] first sample the point clouds and learn the features from gradually downsampled features. Grid-based detectors [9], [10], [22]–[25] first voxelize the point clouds into equally spaced grids and learn the features from these discrete grids. Point-voxel-based detectors [5], [6] utilize both points and voxels for 3D detection. In this work, we adopt PV-RCNN [5] as our 3D object detector.

Domain Adaptation for 3D Detection. Domain adaptation approaches aim to adapt the model trained on the source domain to the target domain. Wang *et al.* [11] identify the difference in object size statistics as the key factor of domain shifts and normalize the object size distribution of the source domain by using its statistics of the target domain to mitigate the domain shifts. However, its effectiveness largely depends on object size distributions. MLC-Net [12] leverages the teacher-student paradigm for pseudo-label generation via three levels of consistency to implement domain adaptation. Yang *et al.* [13] further conclude that the domain shifts arise not only from the object size statistics but also from the point cloud distribution. They propose a new self-training pipeline called ST3D and achieve state-of-the-art performance in many DA tasks. Yet, without any target annotations, the performance gap between the UDA approach and the fully-supervised oracle approach is still noticeable. Therefore, we propose WLST, a general weak labels guided self-training framework to obtain more consistent pseudo labels and improve the DA performance as illustrated in Sec. III-B.

Weakly-supervised 3D Detection. Weakly-supervised learning is a promising approach to utilize noisy, limited, or imprecise data to provide supervision signals and lessen the annotation cost. For 3D detection, weakly-supervised 3D approaches aim to obtain an autolabeler to enhance the weak labels into stronger forms (*e.g.* from 2D bounding boxes to 3D boxes). Then, a 3D detector would be trained on these 3D pseudo labels. For example, Wei *et al.* [17] propose a non-training frustum-aware geometric reasoning framework (FGR) to generate 3D pseudo labels from the frustum point clouds based on a 2D bounding box. Meng *et al.* [26] develop a quick BEV center click annotation strategy and generate 3D pseudo labels from these BEV center click annotations. Liu *et al.* [16] introduce a trainable model called MAP-Gen, which leverages dense image information to tackle the sparsity issue of 3D point clouds and generates high-quality 3D pseudo labels from 2D bounding boxes. In spite of the state-of-the-art performance of MAP-Gen [16], it still needs a small amount of ground truth 3D labels to train its autolabeler. Despite the promising results obtained from the above methods, they fail to consider cross-domain

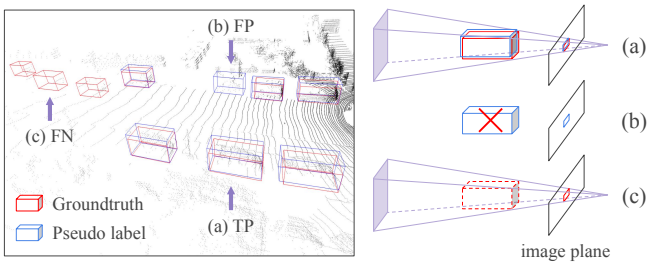

Fig. 2: **Left:** Visualization of false positive (FP), false negative (FN), and true positive (TP) boxes of the pseudo labels. **Right:** According to the projective geometry, frustums can be generated by utilizing their 2D bounding boxes as the projection source and they define the 3D search space for pseudo labels, which manifests that an object should be located in the frustum corresponding to its 2D bounding box. In other words, when we re-project the pseudo labels into 2D image plane, (a) A TP box tends to have a higher IoU with its corresponding 2D bounding box. (b) A FP box does not have corresponding 2D bounding box and it is less likely to have a decent IoU with any 2D bounding box. (c) We can also learn that an object should exist in the frustum corresponding to a FN box.

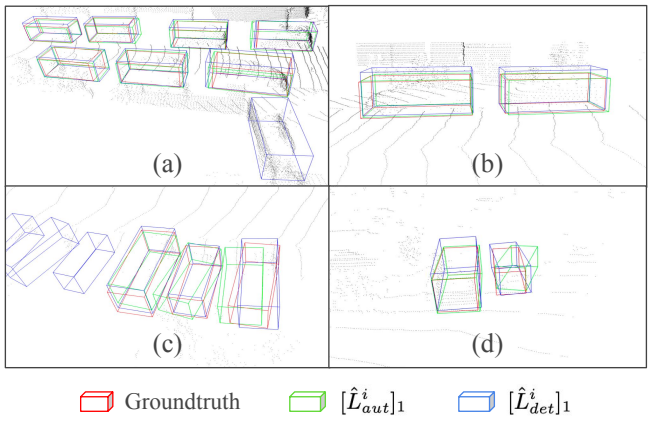

Fig. 3: Visualization of pseudo labels $[\hat{L}^i_{det}]_1$ and $[\hat{L}^i_{aut}]_1$ generated by 3D detector and autolabeler respectively. We observed that **Top:** $[\hat{L}^i_{aut}]_1$ has higher precision. (a) It is less likely to predict extra FP boxes. (b) It is able to predict the heights of objects more precisely. **Bottom:** $[\hat{L}^i_{det}]_1$ has higher recall. (c, d) It has a better understanding of the correlation between objects, *e.g.* a line of vehicles.

scenarios. Hence, we propose an autolabeler designed for DA as illustrated in Sec. III-B.1.

III. Methodology

We formulate the problem of weakly-supervised domain adaptation (WDA) on 3D object detection in Sec. III-A and present our weak labels guided self-training framework, WLST, in Sec. III-B.

A. Problem Formulation

Under the weakly-supervised domain adaptation (WDA) on 3D object detection setting, the goal is to adapt a 3D object detector from the labeled source domain $D_s = \{(P_{s,i}, L^{2D}_{s,i}, L^{3D}_{s,i})\}^{n_s}_{i=1}$ to the weak-labeled target domain $D_t = \{(P_{t,i}, L^{2D}_{t,i})\}^{n_t}_{i=1}$, where n_s and n_t denote the number of samples from the source and target domain respectively. Here, $P_{s,i}$, $L^{2D}_{s,i}$, and $L^{3D}_{s,i}$ represent LiDAR point clouds, 2D bounding boxes (*i.e.* weak labels), and 3D bounding boxes (*i.e.* strong labels) from the i-th source domain sample. The 2D bounding boxes are parameterized by their coordinates of the top-left and bottom-right corners in the image plane. (Different datasets might parameterize 2D bounding boxes differently.) The 3D bounding boxes are parameterized by their center location (c_x, c_y, c_z), dimension (l, w, h), orientation θ, and category. Similarly, $P_{t,i}$ and $L^{2D}_{t,i}$ represent LiDAR point clouds and 2D bounding boxes from the i-th target domain sample.

B. Weak Labels Guided Self-training Framework

In this section, we present WLST, a general weak labels guided self-training framework that adapts the 3D detector trained on the labeled source domain to the weak-labeled target domain with the guidance of weak labels, which is shown in Fig. 1. Our framework is composed of three stages: (1) Pre-train a 3D detector and an autolabeler on the source data (see Sec. III-B.2). (2) Generate high-quality pseudo labels by our *consistency fusion strategy* on the target data (see Sec. III-B.3). (3) Re-train the 3D detector and autolabeler on the pseudo-labeled target data (see Sec. III-B.4).

1) Autolabeler: An autolabeler aims to generate 3D pseudo labels from weak labels (*i.e.* 2D bounding boxes). Despite the promising results obtained from [16], [17], they fail to consider cross-domain scenarios. Hence, we propose an autolabeler designed for DA as illustrated in Fig. 4. Inspired by Frustum PointNets [27] and Cascade-RCNN [28], we adopt coordinate transformations (*e.g.* frustum coordinate, mask coordinate) to canonicalize the point cloud for more effective learning and utilize cascaded box regression networks to fine-tune the pseudo boxes iteratively.

Specifically, we first extract the frustum points from a given 2D bounding box in the camera coordinate shown in Fig. 4 (a). With the known camera intrinsic and extrinsic matrices, the frustum can be generated by utilizing a 2D bounding box as the projection source and it defines a 3D search space for the pseudo label. We then gather the point clouds within the frustum to form frustum points as the input of autolabeler.

To make the distribution of frustum points more aligned across objects, we transform their coordination to orthogonalize the +X axis of the frustum to the image plane shown in Fig. 4 (b). Then, the frustum points are passed to a PointNet [19] based foreground segmentation network M_{seg} to extract foreground points. Furthermore, in order to perform residual-based 3D localization, foreground points are transformed to the mask coordinate by translating their centroid to the origin shown in Fig. 4 (c).

Subsequently, to perform robustly against numerous domain shifts, we utilize two cascaded PointNet [19] based box regression networks M_{reg} and M'_{reg} to regress the 3D pseudo label. To be specific, we utilize the first network M_{reg} to predict the initial 3D bounding box in the first place. Then, foreground points are transformed to the box coordinate, in which the box's orientation is parallel to the +X axis

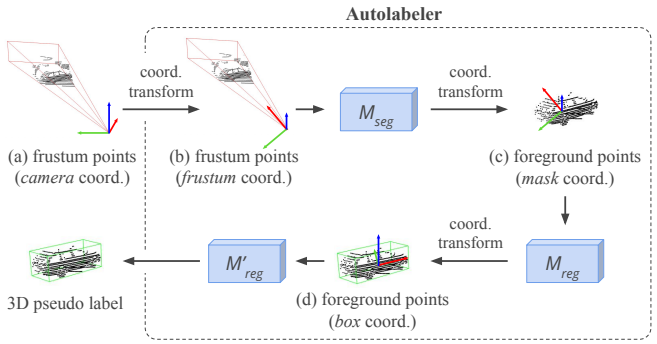

Fig. 4: **Our proposed autolabeler designed for DA.** The model takes the frustum points in the camera coordinate as input and outputs a 3D pseudo label. (M_{seg} denotes foreground segmentation network, and M_{reg} denotes box regression network.)

and the box's center is at the origin shown in Fig. 4 (d), to perform residual-based 3D localization and orientation. Ultimately, we generate the final 3D pseudo label from the second network M'_{reg}.

2) Model Pre-training: Our WLST framework starts from pre-training a 3D detector and an autolabeler on the labeled source domain $D_s = \{(P_{s,i}, L^{2D}_{s,i}, L^{3D}_{s,i})\}^{n_s}_{i=1}$. Apart from valuable knowledge, the pre-trained models also learn the biases from the source domain due to inevitable domain shifts. According to recent works [11]–[13] which aim to mitigate the domain shifts, they unanimously agreed that the bias in object size statistics has harmful impacts on 3D object detection and leads to inaccurate size prediction of 3D bounding boxes on the target domain. To alleviate this problem, we randomly rescale the object size similar to [13] in the pre-training process to simulate the diverse object size distribution on the target domain, which makes the 3D detector and autolabeler more robust against object size bias.

3) Pseudo-label Generation: With the pre-trained 3D detector and autolabeler, the next step is to generate pseudo labels on the weak-labeled target domain. For clarity, we refer to $[\hat{L}^i_{det}]_k$ as initial pseudo labels generated by 3D detector M_{det} at the k-th iteration and $[\hat{L}^i_{aut}]_k$ as initial pseudo labels generated by autolabeler M_{aut} at the k-th iteration. Note that non-maximum suppression (NMS) was conducted for $[\hat{L}^i_{det}]_k$ to get rid of the redundant boxes.

Consistency Fusion Strategy. We propose *consistency fusion strategy* to effectively select high-quality pseudo labels $[\hat{L}^i]_k$ from $[\hat{L}^i_{det}]_k$ and $[\hat{L}^i_{aut}]_k$ in accordance with *geometric consistency* and *cross-modality consistency* of these pseudo labels.

For *geometric consistency*, we propose a 2D Intersection over Union (IoU) based criterion to assess the existence probability of pseudo labels. Specifically, as illustrated in Fig. 2, objects should be located in the frustums corresponding to their 2D bounding boxes. In other words, when we re-project the pseudo labels into 2D image plane with the known camera projection matrix, a TP box tends to have a higher IoU with its corresponding 2D bounding box, whereas a FP box is less likely to have a decent IoU with any 2D bounding box. To be more specific, for the pseudo label in $[\hat{L}^i_{aut}]_k$ which has an exact corresponding 2D bounding box, we calculate the 2D IoU between the convex hull of the re-projected pseudo label's corners and its corresponding 2D bounding box in the image plane as the existence probability of this pseudo label, which is denoted as $prob$. For the pseudo label in $[\hat{L}^i_{det}]_k$ which has no exact corresponding 2D bounding box, we first calculate the 2D IoU matrix $E = e_w \in \mathbb{R}^{n_w}$ between the convex hull of the re-projected pseudo label's corners and n_w 2D bounding boxes in $L^{2D}_{t,i}$. Then, we take the maximum value in E as the existence probability of this pseudo label. To the best of our knowledge, we are the first to demonstrate that it can be a good criterion to assess the existence probability of the pseudo labels.

For *cross-modality consistency*, we match and fuse the pseudo labels generated by different modalities (*i.e.* 3D detector and autolabeler) that have similar locations, dimensions, and orientations. To be more specific, we calculate the pair-wise 3D IoU matrix $I = i_{j,j'} \in \mathbb{R}^{n_u \times n_v}$ between each box in $[\hat{L}^i_{det}]_k$ and each box in $[\hat{L}^i_{aut}]_k$. Here, we assume that $[\hat{L}^i_{det}]_k$ contains n_u boxes and is denoted as $[\hat{L}^i_{det}]_k = \{(box_u, s_u, prob_u)^k_j\}^{n_u}_{j=1}$, which are box parameters, predicted confidence score, and the existence probability of this box respectively. Similarly, we assume $[\hat{L}^i_{aut}]_k$ contains n_v boxes and is denoted as $[\hat{L}^i_{aut}]_k = \{(box_v, s_v, prob_v)^k_{j'}\}^{n_v}_{j'=1}$. For all pair-wise boxes $(box_u, s_u, prob_u)^k_j$ and $(box_v, s_v, prob_v)^k_{j'}$, they are successfully matched if they achieve both *geometric consistency* and *cross-modality consistency* as

$$\begin{cases} \max(prob_u, prob_v) \geq T_{exist}, \\ i_{j,j'} > 0.1, \end{cases} \tag{1}$$

Note that we set $T_{exist} = 0.7$ refer to KITTI 2D object detection benchmark [2]. Later, they are further fused by only keeping the box with a higher confidence score and then cached into the $[\hat{L}^i]_k$ as $(box, s, prob)^k =$

$$\begin{cases} (box_u, s_u, prob_u)^k_j & , \text{ if } s_u > s_v, \\ (box_v, s_v, prob_v)^k_{j'} & , \text{ otherwise}, \end{cases} \tag{2}$$

For other unmatched boxes in $[\hat{L}^i_{det}]_k$ and $[\hat{L}^i_{aut}]_k$, they fail to achieve either *geometric consistency* or *cross-modality consistency*. Hence, we lower their confidence scores by their existence probability due to their higher uncertainty as $s = s \times prob$ and then cached into the $[\hat{L}^i]_k$. Eventually, we filter out the ambiguous boxes whose confidence scores are lower than a threshold T. (We set $T = 0.6$ in practice.) Benefited from our *consistency fusion strategy*, we could generate more robust and consistent pseudo boxes to improve the process of model re-training.

4) Model Re-training: With the high-quality pseudo labels $[\hat{L}^i]_k$ generated by our *consistency fusion strategy*, we re-train the 3D detector and autolabeler on $\{(P_{t,i}, L^{2D}_{t,i}, [\hat{L}^i]_k\}$. Moreover, we use curriculum data augmentation (CDA) technique proposed by [13] in the model re-training process to gradually generate more challenging cases for the benefit of training process.

Task	Setting	Method	AP_{BEV}	Closed Gap	AP_{3D}	Closed Gap
Waymo → KITTI	-	Source Only	60.32	-	21.66	-
	UDA	ST3D [13]	83.37	+75.50%	64.75	+70.25%
		ST3D++ [14]†	84.59	+79.50%	67.73	+75.11%
	WDA	SN [11]	78.24	+58.70%	62.54	+66.64%
		ST3D (w/ SN) [13]	86.53	+85.85%	76.85	+89.97%
		ST3D++ (w/ SN) [14]†	86.92	+87.13%	77.36	+90.81%
		WLST (Ours)	**86.96**	**+87.26%**	**77.69**	**+91.34%**
	-	Oracle	90.85	-	83.00	-
Waymo → nuScenes	-	Source Only	34.51	-	21.44	-
	UDA	ST3D [13]	36.38	+9.99%	22.99	+9.03%
	WDA	SN [11]	34.95	+0.02%	22.19	+4.37%
		ST3D (w/ SN) [13]	36.65	+11.43%	23.66	+12.93%
		WLST (Ours)	**39.54**	**+26.87%**	**24.46**	**+17.59%**
	-	Oracle	53.23	-	38.61	-
nuScenes → KITTI	-	Source Only	69.26	-	39.17	-
	UDA	ST3D [13]	77.38	+37.61%	70.86	+72.30%
	WDA	SN [11]	60.12	-42.33%	46.23	+16.11%
		ST3D (w/ SN) [13]	83.84	+67.53%	72.91	+76.98%
		WLST (Ours)	**87.16**	**+82.91%**	**77.73**	**+87.98%**
	-	Oracle	90.85	-	83.00	-

TABLE I: **Experiment results on three DA tasks.** Our WLST adopts PV-RCNN [5] as 3D detector and outperforms all existing methods on AP_{BEV} and AP_{3D} of the car category at IoU = 0.7. The reported AP are the results on the moderate case when KITTI is regarded as the target domain and are the overall results for other DA tasks. We also report the closed gap to assess how much the performance gap between Source Only and Oracle is closed. † refers to the results reported by [14].

IV. Experiments

A. Experiment Settings

Datasets. Our experiments are conducted on three widely used 3D object detection datasets, nuSenses Dataset [1], KITTI Benchmark Dataset [2], and Waymo Open Dataset [3], and focus on three DA tasks: (i) Waymo → KITTI, (ii) Waymo → nuScenes, and (iii) nuScenes → KITTI.

Method Comparison. We compare our WLST with other unsupervised approaches (*i.e.* Source Only, ST3D [13], ST3D++ [14]), weakly-supervised approaches (*i.e.* SN [11], ST3D (w/ SN) [13], ST3D++ (w/ SN) [14]), and fully-supervised approach (*i.e.* Oracle). (1) *Source Only* directly evaluates the source domain pre-trained model on the target domain. (2) *SN* [11] is a baseline WDA method that leverages object size statistics of the target domain. (3) *ST3D* and *ST3D++* are the state-of-the-art UDA approaches. (4) *ST3D (w/ SN)* and *ST3D++ (w/ SN)* are the state-of-the-art WDA approaches which are equipped with the SN. (5) *Oracle* evaluates the fully-supervised model trained on the target domain.

Evaluation Metric. We follow [13] and adopt the KITTI evaluation metric on the car category, which is known as the vehicle in the Waymo Open Dataset. In addition, we evaluate objects in the ring view except KITTI dataset as it only provides 2D and 3D bounding box annotations for objects within the Field of View (FoV) of the front camera. We report the Average Precision (AP) over 40 recall positions, and set the IoU thresholds as 0.7 for both the bird's eye view (BEV) IoU and 3D IoU. We also use the closed gap evaluation metric proposed by [13] to assess how much the performance gap between Source Only and Oracle is closed,

Method	Fusion	AP_{BEV} / AP_{3D}
3D detector only		80.97 / 64.53
Autolabeler only		83.36 / 71.22
Non-Maximum Suppression (NMS)	✓	86.49 / 76.89
Bayesian Fusion [29]	✓	86.29 / 76.39
CLOCs3D [30]	✓	85.75 / 75.92
Consistency Fusion Strategy (Ours)	✓	**89.14 / 77.69**

TABLE II: **Fusion Strategy Analysis.** We compare our fusion strategy to other fusion strategies and report AP_{BEV} and AP_{3D} of the car category at IoU = 0.7 on the Waymo → KITTI task. The results suggest that either directly using pseudo labels from 3D detector or autolabeler is suboptimal. In contrast, our proposed *consistency fusion strategy* obtains the best outcome on both AP_{BEV} and AP_{3D}.

which is **closed gap** $= \frac{AP_{Method} - AP_{Source\ Only}}{AP_{Oracle} - AP_{Source\ Only}} \times 100\%$.

B. Experiment Results

We analyze the experiment results in terms of three DA scenarios: (i) Waymo → KITTI: domains with a larger difference in object size statistics, (ii) Waymo → nuScenes: domains with a larger difference in point cloud distribution, and (iii) nuScenes → KITTI: domains with a larger difference in object size statistics as well as in point cloud distribution.

For the first scenario (*i.e.* Waymo → KITTI), we found that it is a relatively simple task due to the fact that both domains have dense point cloud distribution by utilizing 64-beam LiDAR. Any method on this task can effectively close the performance gap between Source Only and Oracle. Yet, our method still outperforms all UDA methods by a large margin (around $\sim$2% in AP_{BEV}, $\sim$10% in AP_{3D}) and better than the WDA methods by around $\sim$0.04% in AP_{BEV} and around $\sim$0.3% in AP_{3D}. These encouraging results validate

Pseudo Labels	Recall 0.7	Precision 0.7
$[\hat{L}^i_{det}]_k$	**50.38**	69.11
$[\hat{L}^i_{aut}]_k$	45.54	72.48
$[\hat{L}^i]_k$	48.01	**78.20**

TABLE III: **Qualitative analysis on pseudo labels.** We evaluate the quality of pseudo labels on the Waymo $\rightarrow$ KITTI task by Recall with IoU $>$ 0.7 and Precision with IoU $>$ 0.7. The pseudo labels $[\hat{L}^i_{det}]_k$, $[\hat{L}^i_{aut}]_k$, and $[\hat{L}^i]_k$ are generated by 3D detector, autolabeler, and later fused by our *consistency fusion strategy* respectively. Our fusion strategy effectively eliminates the redundant FP boxes to obtain high precision and retain high recall simultaneously.

Components			Recall 0.7	Precision 0.7
frustum coord. transform	mask coord. transform	cascaded networks		
			10.26	46.87
✓			33.60	58.02
✓	✓		34.07	60.84
✓	✓	✓	**45.54**	**72.48**

TABLE IV: **Component Analysis in Autolabeler.** We evaluate the quality of pseudo labels generated by autolabeler on the Waymo $\rightarrow$ KITTI task by Recall with IoU $>$ 0.7 and Precision with IoU $>$ 0.7. The results validate the effectiveness of the coordinate transformation components and the design of cascaded networks.

that our method can effectively close the performance gap by 87.26% in AP_{BEV} and 91.34% in AP_{3D}.

For the second scenario (*i.e.* Waymo $\rightarrow$ nuScenes), we found that it is a relatively difficult task when we adapt detectors from the domain with denser point cloud distribution (*e.g.* 64-beam LiDAR) to the domain with sparser point cloud distribution (*e.g.* 32-beam LiDAR). We observed that the baseline method SN only has minor performance gain when the domain shifts in object size statistics is subtle. However, our method also attains a considerable performance gain and outperforms all existing methods.

For the third scenario (*i.e.* nuScenes $\rightarrow$ KITTI), despite its larger difference in point cloud distribution, we found it relatively easy to adapt detectors when the target domain has denser point cloud distribution. That is, it manifests that the point density of the target domain is more crucial on DA tasks than the point density of the source domain. We can obtain comparable performance on KITTI dataset regardless of the point density of the source domain (*e.g.* Waymo $\rightarrow$ KITTI task, nuScenes $\rightarrow$ KITTI task). Furthermore, our method outperforms current state-of-the-art WDA method by a large margin (around $\sim$3% in AP_{BEV}, $\sim$5% in AP_{3D}).

These promising results validate that our method can effectively adapt the 3D object detector trained on the source domain to the target domain and perform robustly against numerous domain shifts.

C. Ablation Studies

Fusion Strategy Analysis. As demonstrated in Tab. II, we conduct fusion strategy analysis on the Waymo $\rightarrow$ KITTI task. Apart from our *consistency fusion strategy*, we also study other fusion strategies like Non-Maximum Suppression (NMS), Bayesian Fusion [29], and CLOCs3D. Bayesian Fusion [29] is a non-learning based fusion strategy derived from the Bayes' rule that assumes conditional independence across modalities. CLOCs3D is extended from CLOCs [30] and we modified the feature tensor in [30] as $T_{i,j} = \{IoU^{3D}_{i,j}, s^{3D}_i, s^{3D}_j, prob_i, prob_j\}$ where $IoU^{3D}_{i,j}$ denotes 3D IoU between pseudo labels, s denotes the predicted confidence score, and $prob$ denotes the existence probability of the pseudo label. Surprisingly, we found that the baseline strategy NMS performs well enough by only selecting boxes with higher confidence scores. Yet, the learning-based fusion strategy CLOCs3D does not perform well possibly because the large difference in the input data distribution (*i.e.* feature tensor) from different domains affects its efficacy. In contrast, our *consistency fusion strategy* performs the best as it leverages geometric consistency and cross-modality consistency to obtain more robust and consistent pseudo labels.

To further validate the effectiveness of our *consistency fusion strategy*, we also conduct qualitative analysis on the pseudo labels $[\hat{L}^i_{det}]_k$, $[\hat{L}^i_{aut}]_k$, and $[\hat{L}^k]_k$ which are generated by 3D detector, autolabeler, and later fused by our *consistency fusion strategy* respectively. According to the statistical results in Tab. III, we found that $[\hat{L}^i_{aut}]_k$ has higher precision attributed to the fact that 2D bounding boxes help constrain the 3D search space for the pseudo labels as described in Fig. 2. $[\hat{L}^i_{det}]_k$ has higher recall as it has a larger Field of View (FoV), which enables a better understanding of the correlation between objects. Nevertheless, our *consistency fusion strategy* effectively eliminates the redundant FP boxes to obtain high precision and retain high recall simultaneously.

Component Analysis in Autolabeler. We propose an autolabeler designed for DA as shown in Fig. 4. Inspired by Frustum PointNets [27] and Cascade-RCNN [28], we adopt coordinate transformations (*e.g.* frustum coordinate, mask coordinate) to canonicalize the point cloud for more effective learning and utilize cascaded box regression networks to fine-tune the pseudo boxes iteratively. As illustrated in Tab. IV, we see that both coordinate transformation components effectively make the distribution of points more aligned across objects and render the autolabeler converge easier. Moreover, the design of cascaded network further make the autolabeler perform robustly against domain shifts.

V. Conclusion

We propose a general weak labels guided self-training framework, WLST, designed for WDA on 3D object detection. By incorporating autolabeler into the existing self-training pipeline, our method is able to generate more robust and consistent pseudo labels. Extensive experiments demonstrate the effectiveness of our framework.

VI. Acknowledgement

This work was supported in part by National Science and Technology Council, Taiwan, under Grant NSTC 112-2634-F-002-006 and by Qualcomm through a Taiwan University Research Collaboration Project. We are grateful to Mobile Drive Technology Co., Ltd (MobileDrive) and the National Center for High-performance Computing.

REFERENCES

[1] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[2] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[3] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. 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*, 2020, pp. 2446–2454.

[4] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 697–12 705.

[5] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. 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*, 2020, pp. 10 529–10 538.

[6] S. Shi, L. Jiang, J. Deng, Z. Wang, C. Guo, J. Shi, X. Wang, and H. Li, "Pv-rcnn++: Point-voxel feature set abstraction with local vector representation for 3d object detection," *International Journal of Computer Vision*, pp. 1–21, 2022.

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

[8] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

[9] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 784–11 793.

[10] Y. Zhou and O. 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*, 2018, pp. 4490–4499.

[11] Y. Wang, X. Chen, Y. You, L. E. Li, B. Hariharan, M. Campbell, K. Q. Weinberger, and W.-L. 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*, 2020, pp. 11 713–11 723.

[12] Z. Luo, Z. Cai, C. Zhou, G. Zhang, H. Zhao, S. Yi, S. Lu, H. Li, S. Zhang, and Z. Liu, "Unsupervised domain adaptive 3d detection with multi-level consistency," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8866–8875.

[13] J. Yang, S. Shi, Z. Wang, H. Li, and X. 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*, 2021, pp. 10 368–10 378.

[14] ——, "St3d++: denoised self-training for unsupervised domain adaptation on 3d object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[15] Y. You, C. A. Diaz-Ruiz, Y. Wang, W.-L. Chao, B. Hariharan, M. Campbell, and K. Q. Weinbergert, "Exploiting playbacks in unsupervised domain adaptation for 3d object detection in self-driving cars," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 5070–5077.

[16] C. Liu, X. Qian, X. Qi, E. Y. Lam, S.-C. Tan, and N. Wong, "Mapgen: An automated 3d-box annotation flow with multimodal attention point generator," in *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022, pp. 1148–1155.

[17] Y. Wei, S. Su, J. Lu, and J. Zhou, "Fgr: Frustum-aware geometric reasoning for weakly supervised 3d vehicle detection," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 4348–4354.

[18] Y. S. Tang and G. H. Lee, "Transferable semi-supervised 3d object detection from rgb-d data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1931–1940.

[19] C. R. Qi, H. Su, K. Mo, and L. 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*, 2017, pp. 652–660.

[20] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

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

[22] J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li, "Voxel r-cnn: Towards high performance voxel-based 3d object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1201–1209.

[23] R. Ge, Z. Ding, Y. Hu, Y. Wang, S. Chen, L. Huang, and Y. Li, "Afdet: Anchor free one stage 3d object detection," *arXiv preprint arXiv:2006.12671*, 2020.

[24] S. Shi, Z. Wang, J. Shi, X. Wang, and H. 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*, vol. 43, no. 8, pp. 2647–2664, 2020.

[25] W. Zheng, W. Tang, S. Chen, L. Jiang, and C.-W. Fu, "Cia-ssd: Confident iou-aware single-stage object detector from point cloud," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 4, 2021, pp. 3555–3562.

[26] Q. Meng, W. Wang, T. Zhou, J. Shen, Y. Jia, and L. Van Gool, "Towards a weakly supervised framework for 3d point cloud object detection and annotation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 8, pp. 4454–4468, 2021.

[27] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 918–927.

[28] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.

[29] Y.-T. Chen, J. Shi, Z. Ye, C. Mertz, D. Ramanan, and S. Kong, "Multimodal object detection via probabilistic ensembling," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*. Springer, 2022, pp. 139–158.

[30] S. Pang, D. Morris, and H. Radha, "Clocs: Camera-lidar object candidates fusion for 3d object detection," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 386–10 393.