Syn-to-Real Unsupervised Domain Adaptation for Indoor 3D Object Detection

Yunsong Wang[1]
yunsong@comp.nus.edu.sg

Na Zhao[2]
na_zhao@sutd.edu.sg

Gim Hee Lee[1]
gimhee.lee@nus.edu.sg

[1] School of Computing
National University of Singapore
Singapore

[2] Information Systems Technology and Design
Singapore University of Technology and Design
Singapore

Abstract

The use of synthetic data in indoor 3D object detection offers the potential of greatly reducing the manual labor involved in 3D annotations and training effective zero-shot detectors. However, the complicated domain shifts across syn-to-real indoor datasets remains underexplored. In this paper, we propose a novel Object-wise Hierarchical Domain Alignment (OHDA) framework for syn-to-real unsupervised domain adaptation in indoor 3D object detection. Our approach includes an object-aware augmentation strategy to effectively diversify the source domain data, and we introduce a two-branch adaptation framework consisting of an adversarial training branch and a pseudo labeling branch, in order to simultaneously reach holistic-level and class-level domain alignment. The pseudo labeling is further refined through two proposed schemes specifically designed for indoor UDA. Our adaptation results from synthetic dataset 3D-FRONT to real-world datasets ScanNetV2 and SUN RGB-D demonstrate remarkable mAP_{25} improvements of 9.7% and 9.1% over Source-Only baselines, respectively, and consistently outperform the methods adapted from 2D and 3D outdoor scenarios. The code is available at **https://github.com/wangys16/OHDA**.

1 Introduction

3D object detection is a fundamental yet challenging task that plays a crucial role in numerous applications such as autonomous driving and augmented reality. In recent years, several deep learning approaches [15, 20, 26, 42, 43] developed for 3D object detection have achieved promising performance. However, the performance heavily relies on large amounts of well-annotated training data, and an in-distribution assumption between training and testing data. When the testing environment (target domain) follows a different distribution from the training environment (source domain), the performance would suffer from a significant drop, as shown by the prediction from the source-only model in Figure 1. This out-of-distribution phenomenon is caused by the domain shift due to a variety of reasons such as different sensor configurations [55] or geographical locations [59] during training and testing data collection.

To deal with the domain shift between the training and testing environments, unsupervised domain adaption (UDA) is proposed by learning domain-invariant knowledge from both the labeled source domain data and the unlabeled target domain data. Recently, a few domain adaptive 3D object detection methods [6, 31, 39, 41] have been devised to avoid requiring heavy 3D annotations on the target domains. Nonetheless, these works predominantly focus on outdoor 3D object detection. Directly applying these techniques to indoor settings, as referenced in Table 1 and Table 2, often yields suboptimal outcomes. The root of such underperformance mainly lies in the distinct domain shifts characterizing outdoor and indoor contexts. Outdoor datasets normally focus on detecting cars and the major domain shift is the average sizes of cars [39]. In contrast, indoor environments present intricate domain gaps across diverse object categories, and demand further investigations.

To this end, we focus on unsupervised domain adaptation for indoor 3D object detection. Compared to outdoor scenarios, different indoor datasets can have much larger and more complicated cross-domain gaps due to different capturing patterns (single-view *vs* multi-view) and diverse object categories with various shape distributions, which lead to the difficulty of using real-world datasets as source domain. In contrary, synthetic data offers large quantity of complete scene-level point clouds with labor-free 3D labels, covering common object categories with diverse shapes, which is suitable for the challenge of indoor UDA and has the potential of training an effective zero-shot 3D detectors without real-world 3D labels. Therefore, in this work we study the Synthetic-to-Real Unsupervised Domain Adaptation (SR-UDA) for indoor 3D object detection.

Figure 1: **Pipelines of Source-Only and OHDA.**

We adopt the synthetic datatset 3D-FRONT [9] as the source domain, which contains 18,968 rooms furnished by 13,151 furniture objects with high-quality textures. To narrow down the syn-to-real domain gap, we introduce an object-aware augmentation strategy, which effectively diversifies the synthetic scenes and is seamlessly integrated with a virtual scan simulation (VSS) technique [8] for data augmentation. To adapt the knowledge across syn-to-real domain shift, we propose an Object-wise Hierarchical Domain Alignment (OHDA) framework, which comprises of an proposal-level domain adversarial training branch and a pseudo-labeling branch, which aligns the cross-domain proposal features and performs more effective joint training on cross-domain data. We further refine the pseudo labeling through two proposed modules: a simple-yet-effective adaptive thresholding strategy to resist the class-specific domain gaps, and a reweighting scheme on pseudo labels based to their consistency towards model perturbation. We assess our methodology using two benchmark datasets for indoor 3D object detection, namely ScanNetV2 [5] and SUN RGB-D [32]. In these evaluations, our OHDA method consistently showcases significant improvements. In summary, our key **contributions** include:
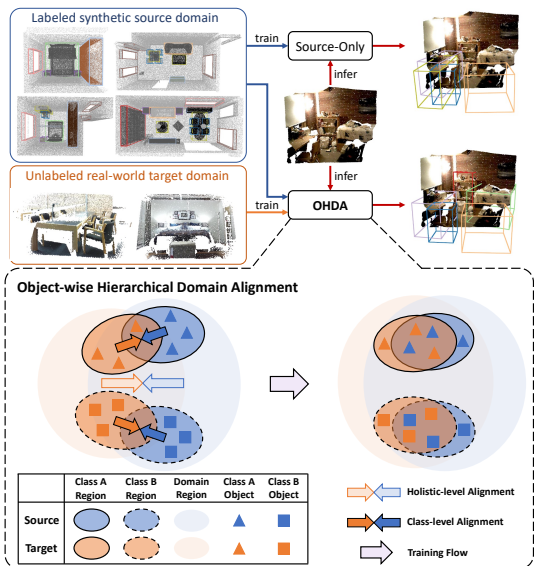
- To the best of our knowledge, this is the first study that delves into UDA for indoor 3D object detection using a syn-to-real approach, which eliminates the need for large-scale 3D labels that are notably costly and challenging to acquire for 3D object detection.

- We propose OHDA as the first solution to the SR-UDA task, which benefits from the two-branch structure that leads to holistic-level and class-level domain alignment. We further propose specific strategies to adapt OHDA to the SR-UDA setting, including object-aware augmentation and two pseudo label refinement modules.

- We establish new benchmarks for 3D-FRONT → ScanNetV2 and 3D-FRONT → SUN RGB-D, improving mAP$_{25}$ by 9.7% and 9.1% over the Source-Only baseline, respectively, consistently surpassing existing UDA methods adapted from different contexts.

2 Related Work

Indoor 3D Object Detection. Benefiting from the well preserved spatial information in point clouds, methods that directly process on point clouds have become the majority in state-of-the-art indoor 3D object detection. Early methods transform the raw 3D point clouds into voxels [24, 44] or use the 2D priors [18, 25] to tackle the difficulties of working on 3D data. Inspired by Hough Voting in 2D object detection [10, 28], VoteNet [26] proposes to predict the offsets from sampled points to their corresponding bounding box centers, which is followed by clustering and PointNet [27]-based grouping. Following VoteNet, several recent methods further incorporate 3D primitives [42], back-tracing strategy [4], or object-level attention [20].

Unsupervised Domain Adaptation. Given the labeled source domain data and the unlabeled target domain data, unsupervised domain adaptation (UDA) aims to effectively adapt knowledge from the source to target domain. This task has been largely explored in 2D, where the discrepancy-based methods [21, 22, 23, 30] learn to minimize the domain discrepancy, the domain adversarial training methods [11, 12, 13, 17, 36] address it by jointly training a domain classifier with min-max optimization, and self-training (*i.e.* pseudo labeling) methods [14, 16, 19, 38, 45] iteratively train the model using both the labeled source and pseudo-labeled target domain data. Most existing methods focus on unsupervised domain adaptation for image classification and semantic segmentation tasks, which are less challenging than the object detection task that requires the precise regression of the bounding boxes.

Domain Adaptive Object Detection. For domain adaptive 2D object detection, most of previous work leverage domain adversarial training to align the cross-domain global feature or proposals [2, 3, 33, 40]. Recently, several approaches also convert this task to the semi-supervised learning problem and utilize advanced pseudo labeling techniques such as Mean Teacher [1, 7, 37] to address the domain shift problem. For domain adaptive 3D object detection, existing works mainly address it under the outdoor scenarios, through bounding box size adaptation [39], pseudo label time-consistency [31], Mean Teacher [41] and global-level domain adversarial training [6]. In contrast, the UDA of 3D object detection in indoor scenarios, which consist of more diverse object categories with class-specific domain gaps across datasets, is more challenging and pragmatic yet underexplored. Thus, this work focuses on this pragmatic task and adopts a large-scale and manual-annotation-free synthetic dataset as the source domain.
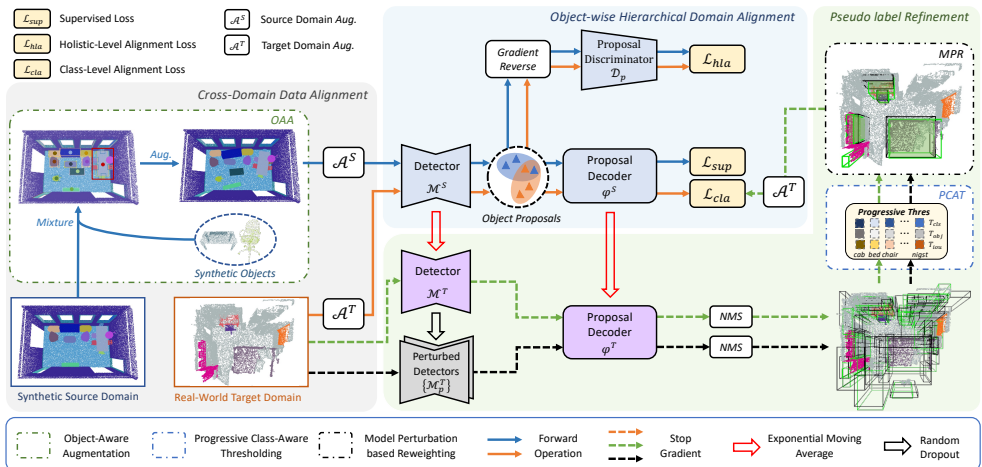
Figure 2: **Proposed Framework.** Our framework consists of three parts: cross-domain data alignment, Object-wise Hierarchical Domain Alignment (consisting of Class-Level Alignment (CLA) and Holistic-Level Alignment (HLA)), and Pseudo Label Refinement (PLR).

3 Our Approach

3.1 Overview

The framework of OHDA is shown in Figure 2. Formally, we denote the detector backbone as $\mathcal{M}$, which predicts the object proposals, and $\boldsymbol{\varphi}$ is the proposal decoder that predicts the bounding box attributes. We denote the student and teacher model as $\boldsymbol{\varphi}^S(\mathcal{M}^S(\cdot))$ and $\boldsymbol{\varphi}^T(\mathcal{M}^T(\cdot))$, respectively. Both student and teacher are initialized by a pre-trained model trained on the augmented source domain data. During adaptation, the student model is updated using gradient decent, and the teacher model is detached and is updated by Exponential Moving Average (EMA) of the student model. In SR-UDA, we are given N_s labeled point clouds from the synthetic source domain $\mathcal{D}_s = \{\boldsymbol{X}_s, \boldsymbol{Y}_s\}^{N_s}$ and N_t unlabeled point clouds from the real-world target domain $\mathcal{D}_t = \{\boldsymbol{X}_t\}^{N_t}$.

3.2 Cross-Domain Data Alignment

The syn-to-real domain gaps in indoor 3D scenes are mainly caused by the change of two aspects: the RGB-D camera scanning pattern and the room layout. The first type of domain gap can be largely solved by VSS [8], which is a recent technique that simulates sensor noise and occlusion similarly to real-world scenes. In this paper, we propose an *Object-Aware Augmentation* (OAA) strategy that can narrow down the second type of domain gap and is compatible with VSS.

Scene-Object Mixture. Although 3D-FRONT benefits from its large quantity of individual scenes, it still suffers from *regular localization* of objects. Besides, the *class-imbalance* problem can also lead to inferior performance of detector. We therefore propose to mix the scenes and the synthetic objects. Specifically, we randomly sample objects from 3D-FRONT [9] and randomly place them into the scene on-the-fly. The objects are placed on the floor, with no collision with original scene. Furthermore, we set the probability of sampling each

object category to be inversely proportional to their frequencies in 3D-FRONT [9], so as to mitigate the class-imbalance problem.

Local Pattern Preserved Augmentation. To diversify the synthetic source domain, the local augmentation on object level is essential, which includes random scaling, rotating and translating the objects on the floor. To prevent the generation of unrealistic scenes, it is crucial to perform collision checks between objects during object augmentation. However, collision checks in indoor scenes are more complex than those in outdoor scenes [29] due to the intricate local spatial patterns between objects. For example, some objects such as chairs and tables have naturally overlapping bounding boxes in the original scenes. Therefore, directly checking the overlap of bounding boxes may not be sufficient to validate object collision. To address this challenge, we propose preserving local patterns by merging overlapped objects. Specifically, before augmentation, we identify existing collided bounding boxes (where IoU> 0.01) and merge all points from the corresponding collided bounding boxes to obtain new object points. The collided bounding boxes are replaced by a minimum axis-aligned bounding box that covers them. We then iteratively merge the collided objects until there are no more collisions, followed by local augmentation on the merged objects.

3.3 Object-wise Hierarchical Domain Alignment

To enhance the domain adaptation under the indoor scenario with more complicated object distributions and cross-domain shifts, we propose a hierarchical domain alignment framework performed on object-level, which includes class-level and holistic-level alignment.

3.3.1 Class-Level Alignment

We first build a pseudo labeling branch to facilitate the domain alignment using labeled source domain data and unlabeled target domain data. Formally, the supervised loss on the labeled source domain samples is computed as $\mathcal{L}_{sup} = \mathcal{L}^l\left(\hat{\boldsymbol{Y}}_s, \boldsymbol{Y}_s\right)$, where $\hat{\boldsymbol{Y}}_s = \boldsymbol{\varphi}^S\left(\mathcal{M}^S(\boldsymbol{X}_s)\right)$ are the student predictions on source domain samples. The Class-Level Alignment (CLA) loss on the pseudo-labeled target domain data is computed as:

$$\mathcal{L}_{cla} = \frac{1}{M}\sum_{m=1}^{M}\mathcal{L}^u\left(\hat{\boldsymbol{Y}}_t^m, \bar{\boldsymbol{Y}}_t^m\right), \tag{1}$$

$$\hat{\boldsymbol{Y}}_t^m = \left[\boldsymbol{\varphi}^S(\mathcal{M}^S(\boldsymbol{X}_t))\right]^m, \bar{\boldsymbol{Y}}_t^m = \mathcal{S}\left(\mathcal{R}(\boldsymbol{\varphi}^T(\mathcal{M}^T(\boldsymbol{X}_t))), \hat{\boldsymbol{Y}}_t^m\right), \tag{2}$$

where M the number of proposals, $\mathcal{F}(\cdot)$ the pseudo label refinement strategy (as discussed in Sec 3.4), $\hat{\boldsymbol{Y}}_t^m$ the m-th bounding box predicted by student model, $\bar{\boldsymbol{Y}}_t^m$ the teacher's prediction that is matched with $\hat{\boldsymbol{Y}}_t^m$ using distance-based scheme $\mathcal{S}$ similarly to [26]. We supervise the student's predictions that are within $0.3m$ of pseudo bounding box centers. $\mathcal{L}^l$ is the standard supervised loss used in [26], and $\mathcal{L}^u$ only contains the box loss and the semantic loss.

3.3.2 Holistic-Level Alignment

The aforementioned Class-Level Alignment is similar to the standard mean teacher [34] in 2D semi-supervised learning, which however could lead to suboptimal performance due to the domain gap between synthetic and real domains under UDA setting. Therefore, we propose to further encourage the cross-domain proposals to share the same feature space,

such that the ground truth labels on the source domain and pseudo labels on the target domain can achieve better joint-supervision on the student model. Specifically, we add a Proposal Discriminator $\mathcal{D}_{\boldsymbol{p}}(\cdot)$ to predict the domain label (0 for source, 1 for target domain) on the object proposal level. The Holistic Level Alignment (HLA) loss is implemented as a cross-entropy loss on all object proposals:

$$\mathcal{L}_{hla} = \max_{\mathcal{M}^S} \min_{\mathcal{D}_{\boldsymbol{p}}} \left\{ -\log\left(1 - \boldsymbol{D}^S\right) - \log\left(\boldsymbol{D}^T\right) \right\}, \tag{3}$$

where $\boldsymbol{D}^S = \mathcal{D}_{\boldsymbol{p}}(\mathcal{M}^S(\boldsymbol{X}_s))$ and $\boldsymbol{D}^T = \mathcal{D}_{\boldsymbol{p}}(\mathcal{M}^S(\boldsymbol{X}_t))$ are the predicted possibilities of the aggregated proposal features $\mathcal{M}^S(\boldsymbol{X}_s), \mathcal{M}^S(\boldsymbol{X}_t)$ coming from target domain.

3.4 Pseudo Label Refinement

3.4.1 Progressive Class-Aware Thresholding

One issue that arises when devising the mean teacher paradigm under UDA settings for indoor 3D object detection is the occurrence of class-specific domain gaps, for which we first develop a straightforward *class-aware thresholding* strategy that accounts for various confidence distributions during the pseudo label filtering to handle the class-specific domain gaps. Initially, all target domain data is fed into the initialized teacher model to predict pseudo labels, and we store the per-class pseudo label confidence scores:

$$\left[\boldsymbol{Q}_{obj}^c, \boldsymbol{Q}_{cls}^c, \boldsymbol{Q}_{iou}^c\right] = \left[\boldsymbol{\tau}_{obj}^c(\bar{\boldsymbol{Y}}_t), \boldsymbol{\tau}_{cls}^c(\bar{\boldsymbol{Y}}_t), \boldsymbol{\tau}_{iou}^c(\bar{\boldsymbol{Y}}_t)\right], \tag{4}$$

where $\boldsymbol{\tau}_{obj}^c(\bar{\boldsymbol{Y}}_t), \boldsymbol{\tau}_{cls}^c(\bar{\boldsymbol{Y}}_t), \boldsymbol{\tau}_{iou}^c(\bar{\boldsymbol{Y}}_t)$ denote the objectness score, maximum classification score, and IoU score of the proposals in $\bar{\boldsymbol{Y}}_t$ with class c, respectively. We calculate the class-aware thresholds as:

$$\boldsymbol{T}_{c,m} = \text{Clamp}\left(\text{Percentile}_\alpha\left(\boldsymbol{Q}_m^c\right), T_l^m, T_h^m\right), \tag{5}$$

where $m \in \{obj, cls, iou\}$ is the confidence metric, $\text{Percentile}_\alpha(\cdot)$ is percentile function that outputs the value higher than $\alpha\%$ of input, and $\text{Clamp}(\cdot, T_l, T_h)$ is clamping between T_l^m and T_h^m. On top of the class-aware thresholds, we further perform progressive updates such that the thresholds are dynamically fit to the pseudo label confidence distributions, which leads to our *Progressive Class-Aware Thresholding (PCAT)*. Denoting $\bar{\boldsymbol{T}}_{c,m}^k$ as the progressive class-aware threshold for epoch k, we update it by:

$$\bar{\boldsymbol{T}}_{c,m}^k = \begin{cases} \boldsymbol{T}_{c,m}, & k = 0, \\ \beta \cdot \bar{\boldsymbol{T}}_{c,m}^{k-1} + (1-\beta) \cdot \boldsymbol{T}_{c,m}^{k-1}, & \text{otherwise.} \end{cases} \tag{6}$$

Here, $T_{c,m}^{k-1}$ is computed with Eq. (5) using the confidence scores in epoch $k-1$, and β is the update momentum.

3.4.2 Model Perturbation based Reweighting

Although the aforementioned PCAT enables adaptively filtering pseudo labels for diverse object classes, the pseudo labeling loss still assigns uniform weights for all valid pseudo labels after score filtering. To further mine the high-quality pseudo labels, we propose the Model Perturbation based Reweighting (MPR) module, which reweights the pseudo labels

according to the pseudo label consistency. Specifically, given each batch of data, we first perform random dropout on the teacher detector backbone to get a set of perturbed teacher models $\{\mathcal{M}_p^T\}_{p=1}^P$, where P is the number of perturbed models. We then feed it with target domain data and go through Non-Maximum Suppression (NMS) and PCAT similarly to get the perturbed pseudo labels $\{\bar{\mathbf{Y}}_{t,p}\}_{p=1}^P$. Consequently, we assess the robustness of $\bar{\mathbf{Y}}_t$ according to their IoU with $\{\bar{\mathbf{Y}}_{t,p}\}_{p=1}^P$, such that each pseudo bounding box is reweighted based on their consistency *w.r.t* model perturbation. Formally, we compute:

$$\mathcal{U}^m = \frac{1}{P}\sum_{p=1}^P \max_n \text{IoU}(\bar{\mathbf{Y}}_t^m, \bar{\mathbf{Y}}_{t,p}^n), \ \boldsymbol{\omega}(\mathcal{U}^m) = 1 + \lambda_{mpr}\cdot\mathcal{U}^m, \tag{7}$$

$$\mathcal{L}_{cla}^{mpr} = \frac{\sum_{m=1}^M \boldsymbol{\omega}(\mathcal{U}^m)\cdot\mathcal{L}^u\left(\hat{\mathbf{Y}}_t^m, \bar{\mathbf{Y}}_t^m\right)}{\sum_{m=1}^M \boldsymbol{\omega}(\mathcal{U}^m)}. \tag{8}$$

Consequently, the MPR is aimed at mining the consistent pseudo labels, which is seamlessly integrated with the score filtering to further enhance pseudo labeling.

3.5 Training Objective

The overall loss is given by:

$$\mathcal{L} = \mathcal{L}_{sup} + \lambda_{hla}\mathcal{L}_{hla} + \lambda_{cla}\mathcal{L}_{cla}^{mpr}, \tag{9}$$

where λ_{hla} and λ_{cla} are weights to balance different losses. The losses work synergically on object proposal level, aiming to achieve hierarchical cross-domain alignments.

4 Experiments

4.1 Implementations Details

Dataset settings. For the data preprocessing of 3D-FRONT [9], we obtain point clouds by uniformly sampling on the mesh surfaces, and select scenes with $3 \sim 20$ interested objects, getting 10,515 scenes for training and 1,104 scenes for validation. As for the label mapping, we select 10 and 7 object categories for 3D-FRONT $\rightarrow$ ScanNetV2 and 3D-FRONT $\rightarrow$ SUN RGB-D settings, respectively.

Experiment settings. We leverage VoteNet [26] as the detector and first train it on 3D-FRONT for 50 epochs, then adapt it to target domains ScanNetV2 and SUN RGB-D individually. We regard traversing target domain as one epoch, and respectively train 100 and 50 epochs for ScanNetV2 and SUN RGB-D.

4.2 Baselines

We set up baselines under source-only and UDA settings. For UDA setting, we use OAA+VSS for source domain augmentation. Since there is no prior work focusing on UDA for indoor 3D object detection, we set up the baselines adapted from 2D and outdoor detection:

Naive MT [34]: Based on the mean teacher paradigm, we reproduce the losses in VoteNet [26] for the unlabeled target domain data, using the pseudo labels provided by the mean teacher.

	Method	cab	bed	chair	sofa	tabl	door	wind	bkshf	desk	nigst	mAP$_{25}$	mAP$_{50}$
SO	*w/o Aug.* [76]	2.1	50.1	63.9	45.7	36.3	<u>2.5</u>	0.7	25.1	23.1	15.8	26.5	13.5
	VSS [8]	9.3	<u>61.3</u>	55.2	67.3	37.4	2.4	4.8	28.3	25.7	40.6	33.2	19.1
	OAA+VSS	9.7	**61.7**	55.3	74.8	41.0	1.8	4.4	29.9	29.8	42.6	35.1	20.3
UDA	Naive MT† [54]	7.6	60.7	65.5	63.3	40.4	2.1	3.3	30.5	11.9	44.3	33.0	20.0
	ST3D† [41]	10.7	54.9	66.5	77.0	39.2	3.0	5.9	<u>32.7</u>	27.1	49.1	36.3	22.6
	Global-DAT† [6]	6.0	63.4	57.3	66.6	<u>42.9</u>	1.6	3.6	22.3	<u>40.7</u>	25.3	32.9	19.3
	Ours *w/o* PLR	<u>12.0</u>	60.8	<u>69.1</u>	77.9	41.3	2.0	<u>6.5</u>	31.2	29.2	<u>55.2</u>	<u>38.5</u>	<u>23.9</u>
	Ours	**12.2**	58.8	**71.0**	**80.2**	**45.7**	**4.4**	**16.0**	**35.4**	**42.7**	**64.6**	**42.9**	**27.6**
FS	Oracle	32.0	87.4	86.3	88.3	55.5	39.1	34.7	39.0	69.6	76.7	60.9	39.9

Table 1: **3D-FRONT → ScannetV2 with per object category mAP@0.25 results.** We compare the results under Source-Only (SO), UDA and Fully-Supervised (FS) settings. *Aug.* denotes data augmentations. We indicate the best and runner-up adaptation results by **bold** and <u>underline</u>, respectively.

	Method	bed	tabl	sofa	chair	desk	nigst	bkshf	mAP$_{25}$	mAP$_{50}$
SO	*w/o Aug.* [76]	35.3	28.1	19.6	40.2	7.3	1.4	4.4	19.5	6.4
	VSS [8]	58.2	30.9	39.9	45.5	11.7	5.4	5.4	28.2	11.0
	OAA+VSS	61.5	29.9	42.4	45.2	**13.8**	15.2	4.8	30.4	14.1
UDA	Naive MT† [54]	64.4	27.3	37.9	41.0	5.3	26.2	4.1	29.5	10.6
	ST3D† [41]	70.3	33.4	37.7	50.4	11.4	26.9	5.8	33.8	15.1
	Global-DAT† [6]	55.3	27.9	37.6	39.8	8.9	5.1	4.9	25.6	9.6
	Ours *w/o* PLR	<u>71.8</u>	<u>35.7</u>	42.5	<u>54.0</u>	8.5	**29.4**	<u>6.2</u>	<u>35.5</u>	<u>15.6</u>
	Ours	**73.7**	**36.2**	**46.1**	**54.7**	<u>12.0</u>	<u>28.8</u>	**9.8**	**37.3**	**18.3**
FS	Oracle	84.1	50.6	60.7	74.2	23.5	56.9	28.0	54.0	29.1

Table 2: **3D-FRONT → SUN RGB-D with per object category mAP@0.25 results.**

ST3D [41]: We reproduce their adaptation results for indoor object detection using their consistency ensemble, where we only use box loss and semantic loss for self-training, similarly to our $\mathcal{L}_{cla}$. Their consistency ensemble can be regarded as a variant of our MPR module.

Global-DAT [6]: We average the seed point features and feed to a global discriminator to predict the domain, and perform min-max optimization similarly.

4.3 Main Results

3D-FRONT → ScanNetV2 results. As shown in Table 1, for data augmentation, OAA enhances mAP by $1\% \sim 2\%$ upon integration with VSS [8]. In the UDA scenario, the direct application of naive MT [54] and Global-DAT [6] to the SR-UDA problem results in performance drop, and the ST3D only boosts mAP marginally by $1 \sim 2\%$. In contrast, our OHDA *w/o* PLR already yields substantial gains, and the introduction of PLR elevates the mAP by $\sim 4\%$, surpassing the VSS Source-Only baseline by 9.7% mAP.

3D-FRONT → SUN RGB-D results. In Table 2, under the 3D-FRONT → SUN RGB-D setting, our approach achieves an mAP increase of 9.1% compared to the VSS-augmented source-only baseline, outperforming other UDA methods adapted from different contexts. Note that the improvements when adding OAA with VSS demonstrate its effectiveness in diversifying object layout distribution, which potentially enhances the learning of the object headings.

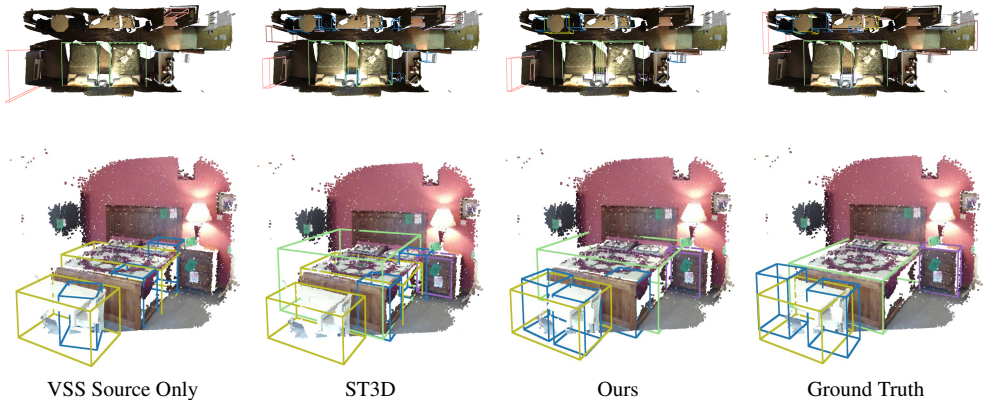**Qualitative results.** The qualitative comparisons are shown in Figure 3. We demonstrate

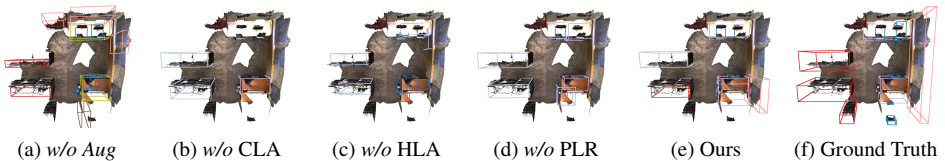Figure 3: **Qualitative results on ScanNetV2 (first row) and SUN RGB-D (second row).**

Figure 4: **Qualitative results for ablation study.** Better viewed in color with zoom in.

qualitative improvements in various object categories with different domain gaps, where we can predict more precise bounding boxes with higher recall.

4.4 Ablation Study

We conducted ablation study to assess the efficacy of the key components in our proposed model. This evaluation is detailed both quantitatively in Table 3 and qualitatively in Figure 4. The results substantiate the necessity of each component in our proposed methodology. Furthermore, an integration of the results from Table 1 and Table 2 reveals a notable insight: while domain adversarial training and naive pseudo labeling might lead to only marginal improvements or even a decline in performance when implemented separately, their synergistic application markedly boosts overall performance. This compellingly underscores the merits of our proposed OHDA framework. Additional ablation studies are available in the supplementary material.

Aug.	CLA	HLA	PLR	ScanNetV2		SUN RGB-D	
				mAP$_{25}$	mAP$_{50}$	mAP$_{25}$	mAP$_{50}$
✘	✓	✓	✓	37.0	22.2	29.3	13.5
✓	✘	✓	✓*	35.6	22.0	31.6	13.4
✓	✓	✘	✘	35.9	21.0	33.1	14.7
✓	✓	✘	✓	39.3	24.1	35.5	16.8
✓	✓	✓	✘	38.5	23.9	35.5	15.6
✓	✓	✓	✓	**42.9**	**27.6**	**37.3**	**18.3**

Table 3: **Ablation study.** * means the PLR components are invalid without CLA.

5 Conclusion

In this paper, we present the OHDA framework which is a pioneering solution to the SR-UDA challenge in indoor 3D object detection. Our key contribution is the integration of proposal-level domain adversarial training and pseudo labeling, which simultaneously align cross-domain proposal feature space and perform effective joint-training under complicated domain shifts. Extensive experiments underline the impressive efficacy of our approach in tackling the SR-UDA problem, where we achieve $9\% \sim 10\%$ consistent mAP gains over Source-Only baselines and consistently surpass existing UDA methods adapted from outdoor scenario. We hope that our work could inspire further research in this direction.

References

[1] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11457–11466, 2019.

[2] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8869–8878, 2020.

[3] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018.

[4] Bowen Cheng, Lu Sheng, Shaoshuai Shi, Ming Yang, and Dong Xu. Back-tracing representative points for voting-based 3d object detection in point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8963–8972, 2021.

[5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[6] Robert DeBortoli, Li Fuxin, 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.

[7] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4091–4101, 2021.

[8] Runyu Ding, Jihan Yang, Li Jiang, and Xiaojuan Qi. Doda: Data-oriented sim-to-real domain adaptation for 3d indoor semantic segmentation. *arXiv preprint arXiv:2204.01599*, 2022.

[9] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10933–10942, 2021.

[10] Juergen Gall, Angela Yao, Nima Razavi, Luc Van Gool, and Victor Lempitsky. Hough forests for object detection, tracking, and action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 33(11):2188–2202, 2011.

[11] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[12] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[13] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Francois Laviolette, Mario Marchand, and Victor Lempitsky. *Domain-Adversarial Training of Neural Networks*. 09 2017.

[14] Dayan Guan, Jiaxing Huang, Shijian Lu, and Aoran Xiao. Scale variance minimization for unsupervised domain adaptation in image segmentation. *Pattern Recognition*, 112: 107764, 2021.

[15] Ji Hou, Angela Dai, and Matthias Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4421–4430, 2019.

[16] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Cross-view regularization for domain adaptive panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10133–10144, 2021.

[17] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Multi-level adversarial network for domain adaptive semantic segmentation. *Pattern Recognition*, 123:108384, 2022.

[18] Jean Lahoud and Bernard Ghanem. 2d-driven 3d object detection in rgb-d images. In *Proceedings of the IEEE international conference on computer vision*, pages 4622–4630, 2017.

[19] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6936–6945, 2019.

[20] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2949–2958, 2021.

[21] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.

[22] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. *Advances in neural information processing systems*, 29, 2016.

[23] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017.

[24] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 922–928. IEEE, 2015.

[25] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 918–927, 2018.

[26] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019.

[27] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[28] SC Radopoulou, M Sun, F Dai, I Brilakis, and S Savarese. Testing of depth-encoded hough voting for infrastructure object detection. In *Computing in Civil Engineering (2012)*, pages 309–316. 2012.

[29] Matthias Reuse, Martin Simon, and Bernhard Sick. About the ambiguity of data augmentation for 3d object detection in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 979–987, 2021.

[30] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732, 2018.

[31] Cristiano Saltori, Stéphane Lathuiliére, Nicu Sebe, Elisa Ricci, and Fabio Galasso. Sf-uda 3d: Source-free unsupervised domain adaptation for lidar-based 3d object detection. In *2020 International Conference on 3D Vision (3DV)*, pages 771–780. IEEE, 2020.

[32] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.

[33] Peng Su, Kun Wang, Xingyu Zeng, Shixiang Tang, Dapeng Chen, Di Qiu, and Xiaogang Wang. Adapting object detectors with conditional domain normalization. In *European Conference on Computer Vision*, pages 403–419. Springer, 2020.

[34] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[35] Darren Tsai, Julie Stephany Berrio, Mao Shan, Stewart Worrall, and Eduardo Nebot. See eye to eye: A lidar-agnostic 3d detection framework for unsupervised multi-target domain adaptation. *IEEE Robotics and Automation Letters*, 7(3):7904–7911, 2022.

[36] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[37] He Wang, Yezhen Cong, Or Litany, Yue Gao, and Leonidas J Guibas. 3dioumatch: Leveraging iou prediction for semi-supervised 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14615–14624, 2021.

[38] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. Domain adaptive semantic segmentation with self-supervised depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8515–8525, 2021.

[39] Yan Wang, Xiangyu Chen, Yurong You, Li Erran Li, Bharath Hariharan, Mark Campbell, Kilian Q Weinberger, and Wei-Lun Chao. Train in germany, test in the usa: Making 3d object detectors generalize. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11713–11723, 2020.

[40] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11724–11733, 2020.

[41] Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. St3d: Self-training for unsupervised domain adaptation on 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10368–10378, 2021.

[42] Zaiwei Zhang, Bo Sun, Haitao Yang, and Qixing Huang. H3dnet: 3d object detection using hybrid geometric primitives. In *European Conference on Computer Vision*, pages 311–329. Springer, 2020.

[43] Yu Zheng, Yueqi Duan, Jiwen Lu, Jie Zhou, and Qi Tian. Hyperdet3d: Learning a scene-conditioned 3d object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5585–5594, 2022.

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

[45] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5982–5991, 2019.