2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Source-free Unsupervised Domain Adaptation for 3D Object Detection in Adverse Weather

Deepti Hegde
Johns Hopkins University
dhegde1@jhu.edu

Velat Kilic
Johns Hopkins University
velat_kilic@jhu.edu

Vishwanath Sindagi
Johns Hopkins University
vishwanath.sindagi@gmail.com

A Brinton Cooper
Johns Hopkins University
abcooper@jhu.edu

Mark Foster
Johns Hopkins University
mark.foster@jhu.edu

Vishal M Patel
Johns Hopkins University
vpatel36@jhu.edu

***Abstract*—A domain shift exists between the distributions of large scale, outdoor lidar datasets due to being captured using different types of lidar sensors, in different locations, and under varying weather conditions. Inclement weather in particular affects the quality of lidar data, adding artifacts such as scattered and missed points, leading to a drop in performance of 3D object detection networks trained on standard lidar datasets. Domain adaptation methods seek to adapt source-trained neural networks to a target domain. Pseudo-label based self training approaches are popular methods for source-free unsupervised domain adaptation. However, their efficacy depends on the quality of the labels generated by the source trained model. These labels may be incorrect with high confidence, rendering thresholding methods ineffective. In order to avoid reinforcing errors caused by label noise, we propose an uncertainty-aware mean teacher framework which implicitly filters incorrect pseudo-labels during training. Leveraging model uncertainty allows the mean teacher network to perform implicit filtering by down-weighing losses corresponding to uncertain pseudo-labels. Effectively, we perform automatic soft-sampling of pseudo-labeled data while aligning predictions from the student and teacher networks. We demonstrate our domain adaptation method on an adverse weather dataset created by augmenting lidar scenes from KITTI with rain, snow, and fog and show that it out-performs current domain adaptation frameworks. We make our code publicly available [1] .**

I. Introduction

Perception is an important part of autonomous driving systems, with navigation and decision making relying heavily on the ability of the vehicle to correctly localize and classify the objects around it. Recent pure lidar-based 3D object detectors have proven to perform extremely well on large public datasets and have topped their challenge leaderboards [1]–[3]. Although containing similar scenes of roads, pedestrians, and vehicles, they tend to differ from each other in terms of pointcloud density, the average size of lanes, as well as the types of vehicles present [4]. This is due to the fact that these datasets are collected using different types of lidar sensors in varying locations around the world, and at times, under varying weather conditions. In weather scenarios such as rain, snow and fog, data from lidar sensors may get corrupted due to reduced signal-to-noise ratio (SNR) or scattered power from the droplets and particles in the air.

This results in a gap in the domains of the training dataset (source) and the testing dataset (target), and poorly generalized 3D object detectors tend to drop in performance when evaluated on samples from the target domain [5]. Since highway systems, driving conventions, and traffic density vary from country to country, this poses a particular challenge in autonomous driving, where generalization is crucial. It would be impractical to collect and annotate every possible type of road scene or weather condition from around the world. Unsupervised domain adaptation (UDA) addresses this by attempting to improve the performance of a source domain trained model on the target domain without having access to the labels of the target dataset. This has been explored on 3D data for tasks such as pointcloud classification and segmentation [6]–[8] and less extensively for 3D object detection [9], [10]. Existing UDA methods require labeled source data along with the source-trained model during adaptation to the target domain [9], [11]. This limits applicability in scenarios where the source data is proprietary, unavailable due to privacy reasons, or too large to store. In order to address this issue we propose a source-free approach that performs domain adaptation of a network to a target domain using only a source-trained model, without the use of source domain data or labels.

Recent domain adaptation methods for 3D object detection include semi-supervised [12] as well as source-free unsupervised approaches [9], [10]. SFUDA^{3D} [10], is the first source free UDA method of this kind, but uses detection-based tracking that requires a sequence of lidar frames to estimate the quality of pseudo-labels, which limits its efficacy in real-time applications. ST3D, another unsupervised approach, obtains promising results on cross-dataset adaptation, but depends on the statistical normalization scheme from [4] to surpass oracle results on the KITTI lidar dataset, which uses label data from the target domain. We propose an unsupervised, single frame, source free domain adaptation method for 3D

[1]https://github.com/deeptibhegde/UncertaintyAwareMeanTeacher

This work was supported by ARO grant W911NF2110135.

object detection. Using a source domain pre-trained model, we follow an iterative training scheme to generate pseudo-labels for the target domain. Although this scheme, along with the use of confidence thresholds, improves the quality of the labels, noise and incorrect labels with high confidence still exist. In order to avoid enforcing these errors while training the mean teacher network, we leverage model uncertainty to perform soft-sampling through Monte Carlo dropout-based uncertainty estimation.

The following are the main contributions of our work:

- We propose an uncertainty-aware, self training framework for source-free unsupervised domain adaptation of 3D object detectors which implicitly selects confident samples out of a set of pseudo-labelled target data.
- We extensively experiment on domain shifts associated with adverse weather scenarios, and demonstrate results on several autonomous driving lidar datasets, and outperform recent domain adaptive works.

II. Related Works

3D Object Detection: 3D Object detection networks aim to localize and categorize objects occurring in a 3D scene by estimating the dimensions, positions in space, and the class predictions of their bounding boxes. The publication of PointNet [13] and its successor PointNet++ [14], has enabled the extraction of point features by directly consuming pointclouds in an end-to-end trainable neural network. These networks form the backbone of several point-based detectors [2], [15]–[18] which operate directly on the points in the Cartesian space and voxel-based detectors [1], [3], [19], which format the points into equally spaced 3D grids. In this work we perform experiments on two particular 3D object detection networks. PointRCNN [2] is a two stage, point-based object detector which generates and refines 3D box proposals in a bottom-up approach, through foreground-background segmentation followed by a second stage which performs bin-based box regression loss. SECOND [1] is a two stage object detector which extracts voxel features from a raw pointcloud through an encoding layer, followed by sparse convolution and an RPN head which generates the detected bounding boxes.

Domain Adaptation: Unsupervised domain adaptation (UDA) addresses the problem of distribution shift when annotations of data in the target domain are unavailable. The broad categories of approaches include adversarial training methods [20]–[23], in which the network is encouraged to learn domain invariant features by being trained jointly with a domain discriminator, self-training methods [24]–[26] that adapt a network with pseudo-labels created by source-model generated annotations, style transfer approaches that bring the target domain closer to the source domain through feature translation [27], [28].

Some of these ideas have been applied to object detection in the 3D domain, such as in [11], where Cane *et al.* leverage large amounts of pseudo-labeled target data along with labelled source domain data to train student networks to perform adaptation of a 3D object detector. In [9], Yang *et al.* propose a self training domain adaptation framework that alternately updates pseudo-labels generated by the source network and model training using curriculum data augmentation. While successful for several domain adaptation scenarios, their best results are obtained via an additional statistical normalisation step taken from [4], which uses bounding box statistics from the target labels, making it not fully unsupervised.

Source-free domain adaptation refers to adaptation methods which use only source-trained models and not the source data or labels during training. This becomes necessary when access to the source data is unavailable due to privacy, copyright, or storage restrictions. Saltori *et al.* proposed SF-UDA^{3D}, [10], a source free domain adaptation framework for 3D object detection that scales pseudo-labels generated by the source-trained model to varying levels and selects the best labels through a scoring method. However, this method relies on detection-based tracking across multiple lidar frames to estimate the score for each pseudo-label.

Mean teacher networks are a popular method for unsupervised, semi-supervised and self-supervised training methods. Liu *et al.* propose [29], a semi-supervised 2D object detector which jointly trains identically initialized student and teacher networks with inputs of differing levels of perturbations. The weights of the teacher network are updated by transfer from student to teacher through exponential moving average (EMA). In [30], Luo *et al.* propose a mean-teacher framework for unsupervised domain adaptation of 3D object detectors, and utilize point consistency, instance consistency, and neural consistency during joint learning. However, [30] does not explicitly account for pseudo-label noise. Although unsupervised, this approach is not source-free, and requires annotated source data during training. Additionally, our framework prefaces the central adaptation stage with a pseudo-label refinement stage that further aids in performance improvement. We propose a fully unsupervised framework for domain adaptation which does not utilize source data or a set of sequential lidar frames.

III. Proposed Method for Domain Adaptive 3D Object Detection

In this section, we provide an overview of the domain adaptation problem and a detailed explanation of our proposed methodology. The goal is to adapt a 3D object detector trained on a source dataset to an unlabelled target dataset without the use of the source data during adaptation.

A. Preliminaries

Consider an object detector ϕ^s trained on an annotated source dataset $\{(X_i^s, Y_i^s)\}_{i=1}^N$, where X_i^s is the i^{th} sample in the set of N samples, and Y_i^s is the corresponding labels consisting of the location and dimensions of each bounding box. With access to this source-trained model, we adapt this detector to an unannotated target dataset $\{(X_i^t)\}_{i=1}^M$, where X_i^t is the i^{th} sample in the set of M samples. Initially, the

3D detector is trained on source data to give source model ϕ^s. In the case of SECOND-iou [1], [9], ϕ^s is supervised by four losses: RPN sigmoid focal classification loss, $\mathcal{L}_{cls}^{rpn}$, RPN weighted smooth $L1$ regression loss $\mathcal{L}_{reg}^{rpn}$, RPN direction classification cross entropy loss $\mathcal{L}_{dir}^{rpn}$, and region of interest (ROI) binary cross entropy classification loss $\mathcal{L}_{cls}^{roi}$.

We propose a framework for unsupervised domain adaptation for 3D object detection. Our approach consists of an iterative training scheme for pseudo-label generation and a student-teacher network to refine the generated pseudo-labels with Monte Carlo dropout based uncertainty estimation to mitigate label noise through soft sampling. An overview of this two-stage framework may be seen in Figure 1, which illustrates the functionality of each component in the architecture.

B. Iterative pseudo-label generation

Naively training the object detector on pseudo-labels generated by ϕ^s and filtered by a threshold could reinforce errors due to the fact that the source trained model may produce incorrect predictions of higher confidence as well as correct predictions of lower confidence. In [31], Xie *et al.* demonstrate the effectiveness of training a classifier with a combination of labelled and pseudo-labelled data over several repeated training sessions. We adapt this approach for the source-free setting in order to mitigate label noise. We propose an iterative generation step to provide better quality pseudo-labels to the mean-teacher network, which performs further soft-sampling.

The source-trained model is inferenced to generate pseudo-labels for target domain data $\{(X_i^t, Y_i^{pt})\}_{i=1}^{M}$, where Y_i^{pt} is the i^{th} source-generated pseudo-label for target sample X_i^t. The detector ϕ is then initialised with the weights from ϕ^s and trained on the pseudo-annotated target data to give the model ϕ^{pt}. This process is repeated J number of times to give target models $\{\phi_j^{pt}\}_{j=1}^{J}$, each initialised with weights from ϕ^s and supervised with pseudo-labels $\{\{Y_{i,j}^{pt}\}_{i=1}^{M}\}_{j=1}^{J}$, filtered with a threshold δ. The pseudo-labels obtained at the end of the J^{th} iteration is obtained by performing inference on ϕ_J^{pt} and used for training the student-teacher network.

C. Mean teacher with Monte-Carlo dropout uncertainty

Mean Teacher In order to avoid enforcing the errors present in the generated pseudo-labels during adaptation, we propose a joint learning framework based on the Mean Teacher method [32] to mitigate label noise while training the object detector. This framework consists of a student network and a teacher network, both identically initialized with the source trained model weights. The student model is supervised using the generated pseudo-labels, and the weights are updated during training through backpropogation. The weights of the teacher network are gradually transferred from the student network by EMA given by

$$W_t \leftarrow \alpha W_t + (1-\alpha) W_s, \tag{1}$$

where W_t and W_s are the weights of the teacher and student networks respectively, and α is the keep ratio. The weights of the teacher network are the average of the weights of the student network over multiple iterations, and thus the teacher becomes a temporal ensemble of the student.

Uncertainty Aware Student Training The epistemic uncertainty of the teacher model is utilized by casting the network as a Bayesian Neural Network as in [33], using the existing dropout layers to approximate variational inference on the network. The first moment of the predictive distribution may be calculated by performing a series of T forward passes of the teacher network and averaging the results in a process called Monte-Carlo dropout [33]. The second moment, or predictive variance of the model may be approximated by the sample variance of the T forward passes given by

$$Y_{i,var}^{pt} = \frac{\sum_{j=1}^{T}(Y_{i,j}^{pt} - Y_{i,mean}^{pt})^2}{T-1}, \tag{2}$$

where

$$Y_{i,mean}^{pt} = \frac{\sum_{j=1}^{T}(Y_{i,j}^{pt})}{T}. \tag{3}$$

In addition to being supervised by the iteratively generated pseudo-labels through the losses present in the network, the student network is also supervised by the pseudo-labels generated by the teacher network in each epoch. The degree of model uncertainty of the teacher is represented by the variance in predictions. The teacher supervises the student with a Binary Cross Entropy loss. The pseudo-labels generated by the teacher network are obtained by computing the sigmoid of the average predictions of T forward passes. The loss value for each ROI is weighted by the inverse of the computed variance. Predicted values with higher variance are thus down-weighted, effectively sampling the data to mitigate noise in the pseudo-labels. This loss can be written as

$$\mathcal{L}_{tea}^{roi} = \frac{1}{N}\sum_{N}^{i=1}\{\mathcal{C} * \mathcal{L}_{BCE}(Y_{pred}^{roi\_stu}, Y_{ps}^{roi\_tea})\}, \tag{4}$$

where $\mathcal{C}$ is the inverse of the predictive variance, Y_{pred}^{roi} is the predicted classification output of the ROI head of the student network, and $Y_{ps}^{roi\_tea}$ is the pseudo-label given by the teacher network after T forward passes.

During joint training, only the student network is supervised by the existing network losses, and $\mathcal{L}_{cls}^{roi}$ is scaled by the uncertainty weights obtained from the predictions of the teacher network. The total loss is thus given by

$$\mathcal{L}_{total} = \mathcal{L}_{cls}^{rpn} + \mathcal{L}_{reg}^{rpn} + \mathcal{L}_{dir}^{rpn} + \mathcal{L}_{cls\_unc}^{roi} + \mathcal{L}_{tea}^{roi}, \tag{5}$$

where

$$\mathcal{L}_{cls\_unc}^{roi} = \frac{1}{N}\sum_{N}^{i=1}(\mathcal{C} * \mathcal{L}_{BCE}(Y_{pred}^{roi}, Y_{ps}^{roi})), \tag{6}$$

and N is the total number of valid ROIs, Y_{pred}^{roi} is the predicted classification output of the ROI head of the student network, and Y_{ps}^{roi} is the pseudo ground truth label of the ROIs generated by the teacher network.

IV. Experiment settings

We demonstrate the proposed method on two 3D object detectors SECOND-iou [1], [9] and PointRCNN [2]. SECOND is a voxel-based 3D object detection network consisting of a voxel feature extractor based on [34], which applies

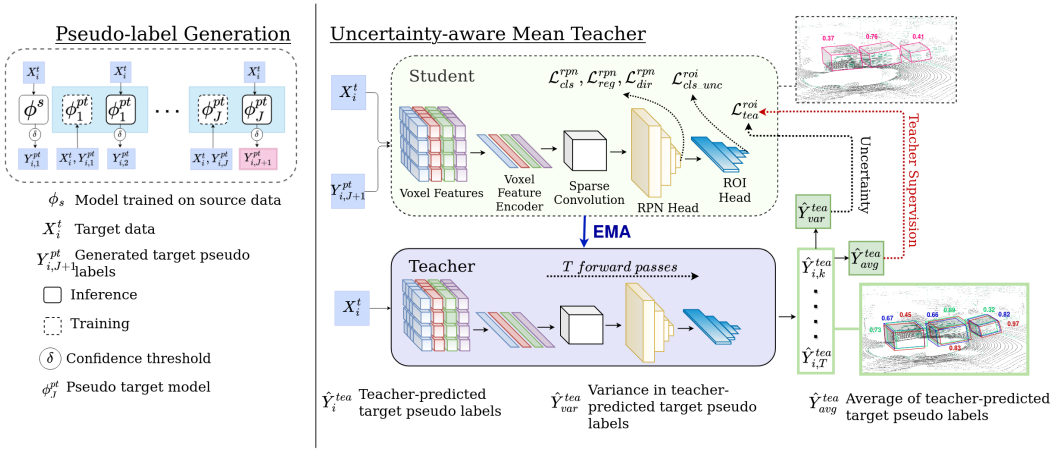

Fig. 1: The two stage architecture of our proposed domain adaptation method. Initialised with a source trained model ϕ^s, the object detector is iteratively trained with successively generated pseudo-labels to create the final set of pseudo-labels $Y^{pt}_{i,J+1}$ with which the mean-teacher network is trained. The mean teacher networks consists of two identically initialised object detection models that are jointly trained, with one network (the student) optimised through backpropogation, and the other (the teacher) optimized by exponential moving average update from the student.

a fully connected layer, batch normalization, and ReLU on each voxel, a sparse convolution layer, followed by a region proposal network (RPN) provides the category and dimensions of the bounding boxes predicted. SECOND-iou is a slightly modified version of this network proposed by Yang *et al.* in [9], which has an additional refinement ROI-head. PointRCNN is a two stage, point-based object detector which generates and refines 3D box proposals in a bottom-up approach, through foreground-background segmentation followed by a second stage which performs bin-based box regression loss.

1) Datasets: We demonstrate our adaptation method using three large-scale lidar datasets, the Waymo Open Dataset [35], KITTI [36], and nuScenes [37]. These datasets vary in size and richness in annotation, with Waymo containing 230K annotated samples, nuScenes containing 38K samples, and KITTI containing 7K samples. We use a subset of the Waymo dataset in a 40K/10K train/val split, and splits of 28K/6K and 3K/3K for nuScenes and KITTI respectively. We address both cross-dataset and adverse weather domain gaps. In particular, we address the domain shifts of Waymo $\rightarrow$ KITTI, KITTI-rain, KITTI-snow, KITTI-fog and nuScenes $\rightarrow$ KITTI, KITTI-rain, KITTI-snow, KITTI-fog. The adverse weather samples are simulated using LISA [38], a physics-based lidar light scattering model. We sample rain rates from an exponential distribution as supported by [39] and recommended by [38] at the rate $\lambda = 0.05\,mm/hr$. We follow a similar procedure for simulating snow precipitation and choose the moderate fog setting.

2) Implementation details: We implement the proposed framework on SECOND-iou using the codebase OpenPCDet [40]. We also refer to code from [9] for their implementation of the extra ROI head. We use the official code release of [2]. During iterative pseudo-label generation step as well as the mean-teacher training step, we use a series of confidence threshold of $\delta \in \{0.1, 0.6, 0.8\}$. We run our method with a batch size of 16. During source model pre-training, we follow the data augmentation procedure used by [9] and train for 50 epochs. The teacher network is trained with a series of $T = 15$ forward passes, and a keep ratio of $\alpha = 0.999$ for the EMA step. During mean teacher training, the entire network is trained for 50 epochs.

V. Results and Evaluation

In this section, we present the results of our proposed domain adaptation framework for two object detectors SECOND-iou [1] and PointRCNN [2], and compare[2] it with recent methods for domain adaptation for 3D object detectors, namely Statistical Normalization (SN) [4] and ST3D [9], SFUDA^{3D} [10], and MLCNet [30], along with the source-only and oracle performances of the object detector. The oracle performance is obtained by training the object detector with the ground-truth annotated target samples.

1) Evaluation metrics: We evaluate the model on the official metrics of the KITTI dataset [36], which divides each object in each sample into 3 categories based on the distance of the object from the sensor and amount of occlusion. The average score across these categories is considered. Mean average precision is calculated for the bird's eye view (BEV) as well as for the entire 3D bounding box, with an IoU threshold of 0.7. Evaluation of the networks is performed on the "Car" class in the KITTI dataset.

Quantitative results Table I tabulates the average mAP across categories for the domain scenarios mentioned previously. We compare the quantitative results of our method

[2]To ensure a fair comparison across all evaluation categories, we re-implement the comparative methods with the same batch size, number of epochs, and other hyperparameters as our models.

TABLE I: A comparison of the network performance of SECOND-iou [9], [41] when adapted to various weather simulations augmented on the KITTI dataset [36] using statistical normalization [12], ST3D [9], and the proposed method, along with the oracle performance, which is obtained by simply retraining the object detector with the target data in a fully supervised manner. The best performance in each category is in bold.

Domain Shift	Method	mAP (BEV/3D)	Domain Shift	Method	mAP (BEV/3D)
Waymo → KITTI	Source Only	69.43 / 40.52	nuScenes → KITTI	Source Only	47.66 / 17.26
	SN	75.98 / 61.28		SN	35.30 / 19.53
	ST3D	80.47 / 61.12		ST3D	**77.31** / 46.94
	Proposed	**80.95 / 66.97**		Proposed	75.27 / **50.78**
	Oracle	82.36 / 73.72		Oracle	82.36 / 73.72
Waymo → KITTI-rain	Source Only	46.59 / 25.49	nuScenes→ KITTI-rain	Source Only	18.89 / 10.01
	SN	54.17 / 33.22		SN	16.79 / 10.85
	ST3D	**62.30** / 41.64		ST3D	26.10 / 16.03
	Proposed	61.04 / **42.50**		Proposed	**38.15 / 25.72**
	Oracle	63.13 / 50.35		Oracle	63.13 / 50.35
Waymo → KITTI-snow	Source Only	43.51 / 24.72	nuScenes → KITTI-snow	Source Only	17.92 / 9.89
	SN	50.38 / 30.83		SN	17.55 / 10.91
	ST3D	54.86 / 22.08		ST3D	36.69 / 19.75
	Proposed	**56.41 / 39.81**		Proposed	**41.21 / 26.78**
	Oracle	65.33 / 50.62		Oracle	65.33 / 50.62
Waymo → KITTI-fog	Source Only	38.51 / 24.25	nuScenes → KITTI-fog	Source Only	17.82 / 9.98
	SN	41.55 / 22.89		SN	20.74 / 15.34
	ST3D	**49.78** / 24.58		ST3D	26.57 / 15.20
	Proposed	48.74 / **35.85**		Proposed	**34.52 / 21.90**
	Oracle	54.69 / 44.14		Oracle	54.69 / 44.14

TABLE II: A comparison of the network performance of PointRCNN [2] when adapted to various weather simulations augmented on the KITTI dataset [36] using statistical normalization [12], SF-UDA^{3D} [10], MLC-Net [30], and the proposed method, along with the oracle performance, which is obtained by simply retraining the object detector with the target data in a fully supervised manner. The best performance in each category is in bold, where the mean average precision (mAP) is calculcated for 3D bounding boxes with an IOU threshold of 0.7.

Domain shift	Method	mAP			Domain shift	Method	mAP		
		easy	mod.	hard			easy	mod.	hard
Waymo → KITTI	Source only	28.86	22.91	19.67	nuScenes→KITTI	Source only	20.62	17.50	15.20
	MLCNet [30]	69.35	59.44	**58.44**		MLCNet	71.26	55.42	48.99
	SFUDA^{3D} [10]	-	-	-		SFUDA^{3D}	68.8	49.8	45.0
	Proposed	**78.68**	**60.98**	55.98		Proposed	**77.07**	**56.44**	**49.12**
	Oracle	87.31	68.70	62.86		Oracle	87.31	68.70	62.86
Waymo→ KITTTI-rain	Source only	12.59	9.41	8.07	nuScenes → KITTI-rain	Source only	14.40	10.20	8.82
	MLCNet	51.23	31.82	31.04		MLCNet	53.14	34.97	32.44
	SFUDA^{3D}	-	-	-		SFUDA^{3D}	-	-	-
	Proposed	**55.63**	**38.57**	**34.42**		Proposed	**56.18**	**40.23**	**35.49**
	Oracle	71.42	48.82	44.43		Oracle	71.42	48.82	44.43

against that of ST3D [9], which uses the same base object detection network (SECOND-IoU) as well as with the semi-supervised method of statistical normalization proposed by Wang *et al.* in [4], and the oracle performance of the network, obtained by fully supervising the detector with the target dataset during training. In the evaluation of our adaptation method, we address two types of domain shift: that associated with a change in dataset (Waymo → KITTI, nuScenes → KITTI) and that associated with both a change in dataset and change in weather conditions (Waymo → KITTI-X, nuScenes → KITTI-X). The degree to which each domain shift affects the performance of the object detector network can be seen in the "Source-Only" row. Networks trained on large datasets tend to generalize better, and show a smaller drop in performance. This can be seen when comparing performance drop of Waymo → KITTI and nuScenes → KITTI. Waymo, with 230K annotated samples, is much larger than nuScenes' 34K samples. The effect of rain, snow and fog can also be observed by making a similar comparison of weather-related domain shifts. Across all shifts, we demonstrate better adaptation performance than recent comparative methods in most categories.In order to compare against the domain adaptive works of SFUDA^{3D} [10] and MLCNet [30], we implement our adaptation method on the

TABLE III: A performance comparison of the object detection network for the "Moderate" difficulty at each iteration of the pseudo-label generation process (center column) and the domain adaptation network trained with the corresponding labels (leftmost column) for the nuScenes → KITTI domain scenario.

Iteration	Self training	Uncertainty-aware mean teacher
source only	46.21 / 17.31	60.03 / 40.48
iteration 1	58.59 / 37.71	62.89 / 44.01
iteration 2	65.33 / 46.42	74.19 / 42.77
iteration 3	64.76 / 47.17	72.33 / 47.91

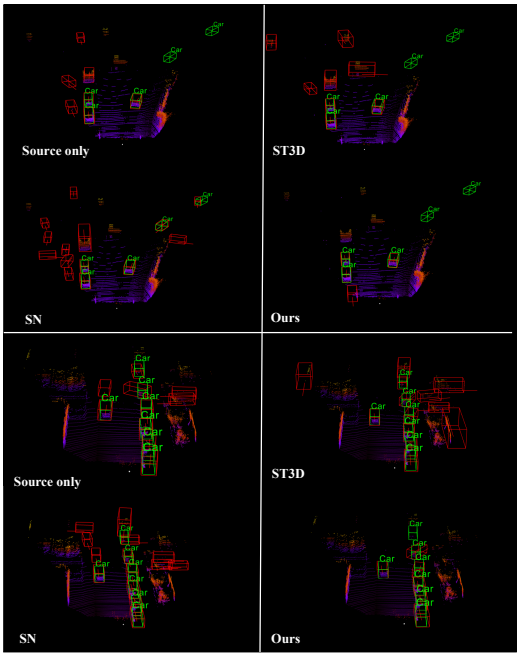

Fig. 2: A qualitative comparison our results on two scenes (top and bottom blocks) against that of the source trained model, ST3D [9] and statistical normalization (SN) [12] for the Waymo → KITTI-snow. The ground truth bounding boxes are in green and the predictions are in red. (Best viewed zoomed in and in color.)

object detector PointRCNN [2] for four domain shifts. Due to a lack of availability of code, we compare against the reported numbers of [10]. In Table II, we maintain the same evaluation metrics as [30] and [10] and compare the 3D mAP performance across the three difficulty categories. Note that despite [30] using source data during adaptation, we still beat their performance under a source-free setting.

Qualitative results We also compare the visual quality of the results against that of the source-only network and ST3D [9] for the Waymo → KITTI-rain domain scenario. As can be seen in Figure 2, both the source trained model and ST3D suffer from incorrect samples with high confidence, negatively affecting the precision. This results in a large number of false positives. Our method avoids this, and also demonstrates better localization. At times, our method suffers from missed detections and a few false positives, despite the presence of which our method mitigates this problem better than the comparative methods.

TABLE IV: A comparison of mean average precision (mAP) values for the 3D Moderate category of a mean teacher framework with and without uncertainty aware weighing of regions during teacher supervision.

Domain shift	Mean teacher			Uncertainty-aware mean teacher		
	easy	mod.	hard	easy	mod.	hard
Waymo / KITTI	72.58	58.11	56.55	73.75	64.56	61.02
Waymo / KITTI-rain	53.03	35.009	30.73	53.27	38.99	35.25

A. Ablation Study

Iterative Pseudo-label Generation As mentioned above, we use an iterative training strategy to initially generate pseudo labels to train the mean-teacher network. With confidence thresholds $0.1 < \delta < 0.8$ at each iteration, this process helps to filter low confidence pseudo labels at several levels. In Table III we compare the performance of the detector when trained with source-only generated pseudo labels and at each subsequent iteration of the pseudo-label generation process for the nuScenes → KITTI domain setting. As observed in the table, the performance of self-training improves with each iteration of pseudo label generation. We also compare the performance of the uncertainty aware mean teacher framework in this setting when trained with each set of pseudo-labels, with the performance demonstrating a similar upward trend. It is clear that the mean teacher framework benefits from the iterative pseudo-label generation, due to the improved quality of the pseudo-supervision.

Uncertainty-aware supervision In order to examine the role of uncertainty-aware teacher supervision of the student network, we compare the Moderate 3D performance of the mean-teacher framework with and without down-weighing losses obtained from the variance of the T forward passes of the network. From Table IV, one can observe that there is significant performance improvement in most categories of the explored domain shifts. The mean teacher framework thus benefits from being made uncertainty-aware, and mitigates noise in both source-model generated pseudo labels and the iteratively refined labels.

VI. Conclusion

We proposed an uncertainty-aware mean teacher framework for domain adaptive 3D object detection, which improves upon naive pseudo-label based self training methods through the mitigation of label noise by down weighing samples the teacher model is uncertain about. We show improved performance across four domain shift scenarios, outperforming the improvement demonstrated by recent unsupervised and semi-supervised domain adaptation methods.

REFERENCES

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

[2] S. Shi, X. Wang, and H. Li, "Pointrcnn: 3d object proposal generation and detection from point cloud," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 770–779.

[3] 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 Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 697–12 705.

[4] Y. Wang, X. Chen, Y. You, L. Erran, 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.

[5] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: A survey of recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 53–69, 2015.

[6] C. Qin, H. You, L. Wang, C.-C. J. Kuo, and Y. Fu, "Pointdan: A multi-scale 3d domain adaption network for point cloud representation," in *NeurIPS*, 2019.

[7] M. Jaritz, T. Vu, R. de Charette, E. Wirbel, and P. Perez, "xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2020, pp. 12 602–12 611. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.01262

[8] W. Liu, Z. Luo, Y. Cai, Y. Yu, Y. Ke, J. M. Junior, W. N. Gonçalves, and J. Li, "Adversarial unsupervised domain adaptation for 3d semantic segmentation with multi-modal learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 176, pp. 211–221, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271621001131

[9] 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.

[10] C. Saltori, S. Lathuili'ere, N. Sebe, E. Ricci, and F. Galasso, "Sf-uda3d: Source-free unsupervised domain adaptation for lidar-based 3d object detection," *2020 International Conference on 3D Vision (3DV)*, pp. 771–780, 2020.

[11] B. Caine, R. Roelofs, V. Vasudevan, J. Ngiam, Y. Chai, Z. Chen, and J. Shlens, "Pseudo-labeling for scalable 3d object detection," *ArXiv*, vol. abs/2103.02093, 2021.

[12] Y. Wang, X. Chen, Y. You, L. Erran, B. Hariharan, M. Campbell, K. Q. Weinberger, and W.-L. Chao, "Train in germany, test in the usa: Making 3d object detectors generalize," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11 710–11 720, 2020.

[13] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[14] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *arXiv preprint arXiv:1706.02413*, 2017.

[15] Z. Yang, Y. Sun, S. Liu, and J. Jia, "3dssd: Point-based 3d single stage object detector," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 040–11 048.

[16] C. He, H. Zeng, J. Huang, X.-S. Hua, and L. Zhang, "Structure aware single-stage 3d object detection from point cloud," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 873–11 882.

[17] 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.

[18] Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia, "Ipod: Intensive point-based object detector for point cloud," 12 2018.

[19] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," 2018.

[20] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, p. 2096–2030, Jan. 2016.

[21] Z. Pei, Z. Cao, M. Long, and J. Wang, "Multi-adversarial domain adaptation," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[22] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[23] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International conference on machine learning*. Pmlr, 2018, pp. 1989–1998.

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

[25] Y. Ge, D. Chen, and H. Li, "Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification," *arXiv preprint arXiv:2001.01526*, 2020.

[26] K. Saito, Y. Ushiku, and T. Harada, "Asymmetric tri-training for unsupervised domain adaptation," in *International Conference on Machine Learning*. PMLR, 2017, pp. 2988–2997.

[27] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3722–3731.

[28] A. L. Rodriguez and K. Mikolajczyk, "Domain adaptation for object detection via style consistency," in *BMVC*, 2019.

[29] Y.-C. Liu, C.-Y. Ma, Z. He, C.-W. Kuo, K. Chen, P. Zhang, B. Wu, Z. Kira, and P. Vajda, "Unbiased teacher for semi-supervised object detection," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[30] 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," 2021.

[31] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 687–10 698.

[32] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *NIPS*, 2017.

[33] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ser. ICML'16. JMLR.org, 2016, p. 1050–1059.

[34] 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.

[35] 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.

[36] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[37] 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," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11 618–11 628, 2020.

[38] V. Kilic, D. Hegde, V. Sindagi, A. B. Cooper, M. A. Foster, and V. M. Patel, "Lidar light scattering augmentation (LISA): physics-based simulation of adverse weather conditions for 3d object detection," *CoRR*, vol. abs/2107.07004, 2021. [Online]. Available: https://arxiv.org/abs/2107.07004

[39] D. Moisseev and V. Chandrasekar, "Examination of the μ - λ relation suggested for drop size distribution parameters," *Journal of Atmospheric and Oceanic Technology*, vol. 24, 06 2007.

[40] O. D. Team, "Openpcdet: An open-source toolbox for 3d object detection from point clouds," https://github.com/open-mmlab/OpenPCDet, 2020.
[41] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, 2018. [Online]. Available: https://www.mdpi.com/1424-8220/18/10/3337