Uncertainty-aware Mean Teacher for Source-free
Unsupervised Domain Adaptive 3D Object Detection

Deepti Hegde, Vishwanath Sindagi, Velat Kilic, A. Brinton Cooper, Mark A. Foster and Vishal M. Patel
Department of Electrical and Computer Engineering, Johns Hopkins University
Baltimore, U.S.A

Abstract

Pseudo-label based self training approaches are a popular method 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 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 method on several domain adaptation scenarios, from cross-dataset to cross-weather conditions, and achieve state-of-the-art performance in these cases, on the KITTI lidar target dataset.

1. 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 [33, 27, 13].

The recent years have seen the publication of several large autonomous driving lidar datasets [2, 1, 28, 20]. 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 [31]. 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. Particularly in the case of adverse weather such as rain, snow, or fog, lidar returns are affected by scattering, resulting in missing data in the final pointcloud [11]. Thus, there exists a gap in the domains of the training dataset (source) and the testing dataset (target), and 3D object detectors that generalize poorly drop in performance when evaluated on samples from the target domain.

This poses a particular challenge in autonomous driving systems where generalization is crucial, since highway systems, driving conventions and traffic density vary from country to country. Naively, this problem may be mitigated by training the network with data from the various target domains. However, it would be impractical to collect and annotate every possible type of road scene 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 \[24, 10, 16\] and less extensively for 3D object detection \[34, 26\]. Recent works have also explored the semi-supervised approach by utilizing vehicle statistics from a particular area, along with the average size of objects in the target domain \[32\].

Existing UDA methods require labeled source data along with the source-trained model during adaptation to the target domain \[34, 8\]. 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 only using a source-trained model, without the use of source domain data or labels.

Recent domain adaptation methods for 3D object detection include semi-supervised \[32\] as well as source-free unsupervised approaches \[26, 24\]. SFUDA \[30, 26\] 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 \[31\] 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 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. We explore the cross-dataset domain shift between the Waymo Open Dataset \[28\] and the KITTI lidar dataset \[7\], as well as nuScenes \[1\] along with the domain shift corresponding to adverse weather conditions. An overview of our method along with a brief qualitative comparison of the result with that of a source-trained detector can be seen in Figure 1.

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 demonstrate our approach on the object detectors SECOND \[33\].

- We extensively experiment on domain shifts associated with cross dataset and adverse weather scenarios, and demonstrate results on several autonomous driving lidar datasets such as the Waymo Open Dataset \[28\], the KITTI lidar dataset \[7\], and nuScenes \[1\].

2. Related works

3D Object Detection has been well explored in the recent years, particularly for autonomous driving scenarios, with several pure lidar-based detectors topping the challenge leaderboards of popular public datasets. These detectors 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 \[22\] and its successor PointNet++ \[23\], 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 \[35, 30, 27, 21, 36\] which operate directly on the points in the Cartesian space and voxel-based detectors \[37, 13, 33\] which format the points into equally spaced 3D grids.

PointRCNN \[27\] 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 \[33\] is a single 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. Several networks use radar and image data in addition to lidar for a multimodal approach to object detection \[4, 15, 21\].

Unsupervised Domain Adaptive Object Detection addresses the problem of the drop in performance when networks are evaluated on samples from a distribution different than that of the training set, when annotations of data in the new domain is unavailable. This has been well explored for 2D detection on images \[19\], including approaches such as adversarial training with domain discriminators \[6\], image translation to bring the target domain closer to the source domain through style transfer \[25\], and self training approaches that adapt an object detector with pseudo-labels created by source-model generated annotations \[9\]. Some of these ideas have been applied to object detection in the 3D domain, such as in \[8\], 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 \[34\], Yang et al. propose a self training domain adaptation framework that alternatively updates pseudo-labels generated by the source network and model training using curriculum data augmen-
3. Proposed method

In this section, we provide an overview of the 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.

3.1. Preliminaries

Consider an object detector \( \phi^s \) 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 using a detailed explanation of our proposed methodology. The goal is to adapt a 3D object detector trained on a source dataset to an unannotated target dataset without the use of the source data during adaptation.

3. Proposed method

In this section, we provide an overview of the 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.

3.1. Preliminaries

Consider an object detector \( \phi^s \) 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 using a data augmentation method from [34], where objects and labels are randomly scaled with a series of factors, to give
source model $\phi^s$. In the case of SECOND-iou\cite{33,34}, $\phi^s$ is supervised by four losses: RPN sigmoid focal classification loss, $L_{cls}^{rpn}$, RPN weighted smooth $L1$ regression loss $L_{reg}^{rpn}$, RPN direction classification cross entropy loss $L_{dir}^{rpn}$, and ROI binary cross entropy classification loss $L_{cls}^{roi}$. 

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 to mitigate label noise through soft sampling.

### 3.2. Iterative pseudo-label generation

Naively training the object detector on pseudo-labels generated by $\phi^s$ and filtered by a threshold could reinforce errors due to the fact that the source trained model may produce incorrect of higher confidence as well as correct predictions of lower confidence. We demonstrate the prevalence of label noise in Figure 4, in which we plot the density of correct and incorrect pseudo-labels with respect to their confidence scored at each step for the Waymo $\rightarrow$ KITTI domain scenario. The pseudo-labels generated by the last two iterations have higher IoU with a higher confidence score as compared to those generated by the source-trained detector inferred for the first time, and that of the first iteration. In order for thresholding to be effective, we desire correct labels to have a higher confidence score, and less confident pseudo-labels to correspond with lower IoU scores. This can be observed in the later two iterations.

In order to mitigate this 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 inferred to generate pseudo-labels for target domain data $\{(X^t_i, Y^\text{pt}_i)\}_{i=1}^M$.

where $Y^\text{pt}_i$ is the $i^{th}$ source-generated pseudo-label for target sample $X^t_i$. The detector $\phi$ is then initialised with the weights from $\phi^s$ and trained on the pseudo-annotated target data to give the model $\phi^pt$. This process is repeated $J$ number of times to give target models $\{\phi^pt_j\}_{j=1}^J$, each initialised with weights from $\phi^s$ and supervised with pseudo-labels $\{\{Y^\text{pt}_{ij}\}_{i=1}^M\}_{j=1}^J$, filtered with a threshold $\delta$. The pseudo-labels obtained at the end of the $J^{th}$ iteration is obtained by performing inference on $\phi^pt_j$, where $j$ being the $j^{th}$ iteration and used for training the student-teacher network. An overview of this process is illustrated in Figure 3.

### 3.3. 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\cite{29} 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 backpropagation. The weights of the teacher network are gradually transferred from the student network by EMA given by

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

where $W_t$ and $W_s$ are the weights of the teacher and student networks respectively, and $\alpha$ 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\cite{5}, 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\cite{5}. The second moment, or predictive variance of the model may be approximated by the sample variance of the $T$ forward passes given by

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

where

$$Y^\text{pt}_{i, mean} = \frac{\sum_{j=1}^T Y^\text{pt}_{i,j}}{T}.$$  

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

---

**Figure 3. An expanded diagram of the iterative pseudo-label training stage.** Initialized with the source trained model $\phi^s$ at each iteration, the object detector is trained on the pseudo-annotated target data, the labels of which are updated with each iteration and filtered with a threshold of $\delta$. The solid boxes indicate the inference steps where pseudo-labels are generated and the dotted-line boxes indicate training with the pseudo-labels generated in the previous iteration.
Figure 4. A density plot of correct (IoU ≥ 0.5) and incorrect (IoU < 0.5) pseudo-labels according to their confidence scores for the Wayo → KITTI domain scenario. The density curve of labels generated by the source model is in blue, iteration 1 in green, iteration 2 in red, and iteration 3 in purple. With successive iterations pseudo-label training, labels become better in quality, with higher confidence as well as higher IoU scores, as can be seen in the top graph. The peak of the last two iterations is higher and further to the left that those of the first two iterations. This indicates more correct samples of higher confidence. pseudo-labels from the source only model display higher confidence for incorrect samples, as can be seen in the bottom graph. The peaks of the red and purple curves are more defined and closer to 0, indicating incorrect pseudo-labels are of lower confidence.

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

$$L_{roi}^{tea} = \frac{1}{N} \sum_{i=1}^{N} \left( C \ast L_{BCE}(Y_{roi}^{pred}, Y_{roi}^{tea}) \right),$$

(4)

where $C$ is the inverse of the predictive variance, $Y_{roi}^{pred}$ is the predicted classification output of the ROI head of the student network, and $Y_{roi}^{tea}$ 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 $L_{roi}^{tea}$ is scaled by the uncertainty weights obtained from the predictions of the teacher network. The total loss is thus given by

$$L_{total} = L_{rpn} + L_{reg} + L_{rpn}^{roi} + L_{cls_{unc}} + L_{tea},$$

where

$$L_{cls_{unc}}^{roi} = \frac{1}{N} \sum_{i=1}^{N} \left( C \ast L_{BCE}(Y_{roi}^{pred}, Y_{roi}^{ps}) \right),$$

(5)

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

4. Experiments

In this section, we explain the details of the experiments carried out and conduct an analysis of model uncertainty.

4.1. Experimental setup

Datasets We demonstrate our method on three well-known lidar detection datasets – Waymo[29], KITTI[7], and nuScenes[1]. The Waymo Open Dataset is collected from
Table 1. A comparison of results for 3 domain shift scenarios from the proposed method with that of a recent domain adaptive 3D object detector ST3D \[34\], a semi-supervised domain adaptive technique statistical normalization (SN) \[31\], the source-model performance, and the oracle performance of the object detector, i.e. trained in a fully supervised manner with target data. Our proposed framework outperforms the compared method in most categories across all the explored domain shifts.

| Domain shift       | Method | mAP (BEV/3D) | Easy     | Moderate | Hard     |
|--------------------|--------|--------------|----------|----------|----------|
|                    |        |              |          |          |          |
| Waymo \(\rightarrow\) KITTI | Source Only | 82.57 / 46.66 | 66.70 / 39.86 | 59.02 / 35.03 |
|                    | SN     | 86.86 / 73.23 | 72.46 / 56.36 | 68.63 / 54.25 |
|                    | ST3D   | 87.97 / 67.47 | 77.36 / 59.17 | 76.09 / 56.73 |
|                    | Proposed | 88.60 / 75.28 | 77.79 / 64.56 | 76.48 / 61.02 |
|                    | Oracle | 89.55 / 84.86 | 79.29 / 68.93 | 78.26 / 67.38 |
|                    |        |              |          |          |          |
| nuScenes \(\rightarrow\) KITTI | Source Only | 53.15 / 18.37 | 46.21 / 17.31 | 43.64 / 16.09 |
|                    | SN     | 39.43 / 22.03 | 31.88 / 18.51 | 31.61 / 18.04 |
|                    | ST3D   | 86.39 / 58.24 | 74.61 / 43.13 | 70.95 / 39.46 |
|                    | Proposed | 82.50 / 62.09 | 72.33 / 47.91 | 70.99 / 42.36 |
|                    | Oracle | 89.55 / 84.86 | 79.29 / 68.93 | 78.26 / 67.38 |
|                    |        |              |          |          |          |
| Waymo \(\rightarrow\) KITTI-rain | Source Only | 60.36 / 31.61 | 40.66 / 23.73 | 38.74 / 21.12 |
|                    | SN     | 68.23 / 40.67 | 47.99 / 30.95 | 46.32 / 28.05 |
|                    | ST3D   | 76.45 / 52.16 | 57.21 / 37.57 | 53.24 / 35.20 |
|                    | Proposed | 74.08 / 53.27 | 55.89 / 38.99 | 53.17 / 35.25 |
|                    | Oracle | 78.14 / 63.89 | 59.79 / 47.22 | 51.45 / 39.95 |

Table 2. A comparison of results for Waymo \(\rightarrow\) nuScenes domain shift scenario from the proposed method with that of a recent domain adaptive 3D object detector ST3D \[34\], statistical normalization (SN) \[31\], the source-model performance, and the oracle performance of the object detector. Our proposed framework outperforms the compared method in most categories.

| Domain shift       | Method | mAP (BEV/3D) | Easy     | Moderate | Hard     |
|--------------------|--------|--------------|----------|----------|----------|
|                    |        |              |          |          |          |
| Waymo \(\rightarrow\) nuScenes | Source Only | 25.39 / 15.25 | 20.66 / 12.37 | 18.74 / 10.12 |
|                    | SN     | 33.23 / 18.57 | 29.79 / 15.52 | 28.05 / 13.04 |
|                    | ST3D   | 29.07 / 15.52 | 25.98 / 12.57 | 24.34 / 11.04 |
|                    | Proposed | 35.10 / 21.05 | 31.09 / 17.05 | 29.45 / 15.04 |
|                    | Oracle | 49.32 / 32.65 | 45.29 / 28.05 | 42.17 / 26.04 |

six cities across the United States, and consists of X training samples and Y validation samples of which we use a subset of X’ and Y’ samples for training and testing, respectively. The KITTI dataset is one of the earliest publicly available lidar datasets for scene understanding, and is much smaller in size with 3712 training samples and 3769 validation samples, which we use for testing. The nuScenes dataset consists of 34,149 samples in a 28130/6019 training/validation split. We address the domain shifts of cross-dataset detection when going from label-rich to label-poor scenarios, by exploring the following scenarios: Waymo \(\rightarrow\) KITTI, nuScenes \(\rightarrow\) KITTI, and Waymo \(\rightarrow\) nuScenes.

We also address the domain shift caused by adverse weather conditions, by simulating rain on the KITTI dataset using the physics-based lidar weather simulation algorithm proposed in \[11\]. Sampling from a range of rain rates from 0mm/hr to 100mm/hr to reflect realistic adverse weather conditions, each sample is augmented with the artifacts that occur when lidar data is captured in rainy weather. We denote this domain shift as Waymo \(\rightarrow\) KITTI-rain.

3D object detection network We demonstrate the proposed method on the 3D object detector SECOND-iou \[33, 34\]. SECOND is a voxel-based, single stage 3D object detection network consisting of a pointcloud-grouping step to create voxels and their corresponding coordinates, a feature extractor based on \[38\], which applies 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 \etal in \[34\], which has an additional refinement ROI-head. This is a single stage detector, and the entire network is used for the mean-teacher framework.

4.2. Implementation details

We implement the proposed framework on SECOND-iou using the codebase OpenPCDet \[30\], a popular open-source repository for 3D object detection networks. We also refer to code from \[34\] for their implementation of the extra ROI head. During iterative pseudo-label generation step as well as the mean-teacher training step, we use a series of confidence threshold of \(\delta \in \{0.1, 0.6, 0.8\}\). Due to com-
Table 3. A comparison of mean average precision (mAP) values for the 3D Moderate category of a mean teacher framework with and without uncertainty aware weighting of regions during teacher supervision. As can be observed, performance in most categories is increased by several points when the network is trained to be uncertainty-aware.

| Domain shift     | Mean teacher | Uncertainty-aware mean teacher |
|------------------|--------------|-------------------------------|
|                  | Easy | Moderate | Hard | Easy | Moderate | Hard |
| nuScenes → KITTI | 52.63 | 42.04 | 40.70 | 62.09 | 47.91 | 42.36 |
| Waymo → KITTI    | 75.28 | 58.11 | 56.55 | 75.35 | 64.56 | 61.02 |
| Waymo → KITTI-rain | 52.03 | 35.90 | 30.73 | 53.27 | 38.99 | 35.25 |
| Waymo → nuScenes | -    | 17.08  | -   | -    | 21.05  | -   |

Table 4. The left column shows a comparison of mean average precision (mAP) values of the object detection network for the “Moderate” difficulty at each iteration of the pseudo-label generation process for the nuScenes → KITTI domain scenario. As is apparent in the table, with each successive iteration, the quality of the pseudo-labels increases, as does the performance of the self-training method. The left column demonstrates a similar effect, with the performance of the uncertainty-aware mean teacher framework with each iteration of pseudo labels.

| 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 |

In this section, we present the results of our proposed framework, and compare it with recent methods for domain adaptation for 3D object detectors, namely Statistical Normalization (SN) [31] and ST3D [34], along with the source-only and oracle performances of the object detector.

5.1. Evaluation metrics

We evaluate the model on the official metrics of the KITTI dataset [7], which divides each object in each sample into 3 categories, Easy, Moderate, and Hard based on the distance of the object from the sensor and amount of occlusion. 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. For the Waymo and nuScenes datasets, we consider a full 360° view whereas only the front view is considered for samples from the KITTI dataset. Evaluation of the networks is performed on the “Car” class in the KITTI dataset, with the equivalent class in Waymo and nuScenes being “Vehicle” and “car”, respectively.

5.2. Comparison with SOTA methods

Quantitative results We compare the quantitative results of our method against that of ST3D [34], 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 [31], as well as the oracle performance of the network, obtained by fully supervising the detector with the target dataset during training. These results for each domain shift scenario may be seen in Table 4. We obtain the highest mean average precision in almost all categories across all domain shifts. Particularly in the Waymo → KITTI and Waymo → nuScenes scenarios, our method outperforms the rest.

Qualitative results We also compare the visual quality of the results against that of the source-only network and ST3D [34] for the Waymo → KITTI domain scenario. As can be seen in Figure 6, both the source trained model and ST3 suffer from incorrect samples with high confidence, negatively affecting the precision. Our method avoids this, and also demonstrated better localization.

5.3. Analysis of model uncertainty

We examine the model uncertainty of the object detector during adaptation at the beginning and end of training. We evaluate the pseudo labels by comparing the pseudo class annotation with the ground truth annotation for each detected object. We then plot the variance of the teacher network for the associated region of interest (ROI) of incorrect pseudo labels. We plot the distributions of these variances in Figure 7. As is apparent in the figure, by the end of training, the model learns to better identify incorrect pseudo labels, since fewer incorrect samples have low variance, ensuring
that they are down-weighed during training.

5.4. Ablation study

We conduct several ablation experiments to examine the contribution of each part of the architecture to the performance of the object detector.

Iterative pseudo-label generation As mentioned in Section 3.2, 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 the second column of Table 4, 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 $\rightarrow$ KITTI domain setting. As observed in the table, the performance of simple self-training in mAP improves with each iteration of pseudo-label generation. This can likewise be observed in the other domain shifts in Figure 5 here the mean average precision (mAP) values of each self-trained object detector $\{\phi^j\}_{j=1}^J$ in the “Moderate” difficulty category for each iteration are plotted for each domain shift scenario. As observed in the figure, the precision values increase with each iteration before either saturating or decreasing. We observe that most models reach this point at $J \approx 3$. We also compare the performance of the uncertainty-aware mean teacher framework in this setting when trained with each set of pseudo-labels in the third column of Table 4. It is apparent that the mean teacher framework benefits from the iterative pseudo-label generation, due to the improved quality.

As mentioned in Section 3.2, Figure 4 shows that pseudo-labels generated in later iterations tend to be more confident as well as more correct (i.e. higher IoU with ground truth labels) when compared to pseudo-labels generated by the source trained model and the first iteration model.

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 3, one can observe that there is significant performance improvement in most categories of the explored domain shifts. The mean teacher framework clearly benefits from being made uncertainty-aware through Monte-Carlo Dropout, and significantly mitigates noise not only on source-model generated pseudo labels, but the iteratively refined labels as well.

6. 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] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Q. Xu, Anush Krishnan, Yu Pan, Gian-carlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 11618–11628, 2020.

[2] Qi Cai, Yingwei Pan, C. Ngo, Xinmei Tian, Ling yu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 11449–11458, 2019.

[3] Benjamin Caine, R. Roelofs, Vijay Vasudevan, Jiquan Ngiam, Yuning Chai, Z. Chen, and Jonathon Shlens. Pseudo-labeling for scalable 3d object detection. ArXiv, abs/2103.02093, 2021.

[4] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1907–1915, 2017.

[5] Yarin Gal and Zoubin 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, ICML’16, page 1050–1059. JMLR.org, 2016.

[6] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. J. Mach. Learn. Res., 17(1):2096–2030, Jan. 2016.

[7] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[8] Chenhang He, Hui Zeng, Jianqiang Huang, Xian-Sheng Hua, and Lei Zhang. Structure aware single-stage 3d object detection from point cloud. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11873–11882, 2020.

[9] Naoto Inoue, Ryosuke Furuta, T. Yamasaki, and K. Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5001–5009, 2018.

[10] 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), pages 12602–12611, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society.

[11] Vela Kilic, Deepthi Hegde, Vishwanath Sindagi, A. Brinton Cooper, Mark A. Foster, and Vishal M. Patel. Lidar light scattering augmentation (LISA): physics-based simulation of adverse weather conditions for 3d object detection. CoRR, abs/2107.07004, 2021.

[12] Jogendra Nath Kundu, Naveen Venkat, V. RahulM., and R. Venkatesh Babu. Universal source-free domain adaptation. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 4543–4552, 2020.

[13] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 12697–12705, 2019.

[14] Xianfeng Li, Weijie Chen, Di Xie, Shicai Yang, Peng Yuan, Shiliang Pu, and Yueting Zhuang. A free lunch for unsupervised domain adaptive object detection without source data. In AAAI, 2021.

[15] Ming Liang, Bin Yang, Yun Chen, Rui Hu, and Raquel Urtasaun. Multi-task multi-sensor fusion for 3d object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7345–7353, 2019.

[16] Wei Liu, Zhiming Luo, Yuanzheng Cai, Ying Yu, Yang Ke, José Marcato Junior, Wesley Nunes Gonçalves, and Jonathan Li. Adversarial unsupervised domain adaptation for 3d semantic segmentation with multi-modal learning. ISPRS Journal of Photogrammetry and Remote Sensing, 176:211–221, 2021.

[17] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In Proceedings of the International Conference on Learning Representations (ICLR), 2021.

[18] Zhipeng Luo, Zhongang Cai, Changqing Zhou, Gongjie Zhang, Haiyu Zhao, Shuai Yi, Shijian Lu, Hongsheng Li, Shanghang Zhang, and Ziwei Liu. Unsupervised domain adaptive 3d detection with multi-level consistency, 2021.

[19] Poosan Oza, Vishwanath A. Sindagi, Vibashan VS, and Vishal M. Patel. Unsupervised domain adaptation of object detectors: A survey, 2021.

[20] Matthew Pitropov, Danson Evan Garcia, Jason Rebello, Michael Smart, Carlos Wang, Krzysztof Czarnecki, and Steven Waslander. Canadian adverse driving conditions dataset. The International Journal of Robotics Research, 40(4-5):681–690, 2021.

[21] 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.

[22] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. Proc. Computer Vision and Pattern Recognition (CVPR), IEEE, 2017.

[23] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. arXiv preprint arXiv:1706.02413, 2017.

[24] Can Qin, Haoyuan You, Lichen Wang, C.-C. Jay Kuo, and Y. Fu. PointIdan: A multi-scale 3d domain adaption network for point cloud representation. In NeurIPS, 2019.

[25] Adrian Lopez Rodriguez and K. Mikolajczyk. Domain adaptation for object detection via style consistency. In BMVC, 2019.
[26] Cristiano Saltori, Stéphane Lathuilière, N. Sebe, E. Ricci, and Fabio Galasso. Sf-uda3d: Source-free unsupervised domain adaptation for lidar-based 3d object detection. 2020 International Conference on 3D Vision (3DV), pages 771–780, 2020.

[27] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 770–779, 2019.

[28] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2446–2454, 2020.

[29] Antti Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In NIPS, 2017.

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

[31] Yan Wang, Xiangyu Chen, Yurong You, Li Erran, 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.

[32] Yan Wang, X. Chen, Yurong You, Li Erran, Bharath Hariharan, M. Campbell, Kilian Q. Weinberger, and Wei-Lun Chao. Train in germany, test in the usa: Making 3d object detectors generalize. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 11710–11720, 2020.

[33] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. Sensors, 18(10):3337, 2018.

[34] 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, 2021.

[35] Zetong Yang, Yanan Sun, Shu Liu, and Jiaya Jia. 3dssd: Point-based 3d single stage object detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11040–11048, 2020.

[36] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Ipod: Intensive point-based object detector for point cloud. 12 2018.

[37] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. 2018.

[38] 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.
A. Algorithms

In this section, we detail the training procedures and algorithms for our two stage framework for source free unsupervised domain adaptive 3D object detection. Algorithm 1 details the iterative pseudo label generation procedure illustrated in Figure 3 of the main paper. Algorithm 2 explains the steps involved in training the uncertainty aware mean teacher with the pseudo-labels generated in the previous stage.

Algorithm 1 Iterative pseudo-label generation

**Input:** source trained model \(\phi^s\), unannotated target data

**Output:** \({Y^{pt}_{i,J}}^M_{i=1}\)

- Perform inference on \(\phi^s\) to obtain \({Y^{pt}_{i,0}}^M_{i=1}\)
- Threshold with \(\delta[0]\)

for \(1 \leq j \leq J\) do

- Init \(\phi\) with \(\phi^s\)
- train \(\phi\) with target data of \(M\) samples annotated with \({Y^{pt}_{i,j-1}}^M_{i=1}\)
- Obtain trained model \({\phi^{pt}_j}\)
- Perform inference on \({\phi^{pt}_j}\) to obtain pseudo labels \({Y^{pt}_{i,j}}^M_{i=1}\)
- Threshold with \(\delta[j]\)

end for

- Output pseudo labels = \({Y^{pt}_{i,J}}^M_{i=1}\)

B. Additional Qualitative Results

In this section, we provide further qualitative comparisons with recent domain adaptation method [34] as well as the source trained object detector. In Figure 8 we demonstrate the improved localization of the proposed method as compared to the source-trained network and that of ST3D [34]. Both these methods suffer from incorrect predictions of high confidence, resulting in spurious detections and lower precision.

Algorithm 2 Training the uncertainty-aware mean teacher network

**Input:** source trained model \(\phi^s\), unannotated target data, pseudo labels \({Y^{pt}_{i,J}}^M_{i=1}\)

**Output:** Predictions

- Initialize teacher model with \(\phi^s\)
- Freeze autograd for teacher network.
- Initialize student model with \(\phi^s\)

for \(1 \leq epoch \leq num\_epochs\) do

- Train student net with target data annotated with pseudo-labels \({Y^{pt}_{i,J}}^M_{i=1}\)
- init \(\text{pred\_teacher}\)

for \(1 \leq i \leq T\) do

- Perform a forward pass on the teacher network
- Append output to \(\text{pred\_teacher}\)

end for

- Transfer weights from student network to teacher network through exponential moving average (EMA).
- Compute losses \(L_{\text{rpn}}\), \(L_{\text{reg}}\), \(L_{\text{dir}}\), and \(L_{\text{cls}}\) from the output of the student network.
- Compute \(\hat{Y}\_\text{var}\) and \(\hat{Y}\_\text{avg}\) from \(\text{pred\_teacher}\).
- Compute loss weight \(C = \text{clip}(\frac{1}{\hat{Y}\_\text{var}}, 10^{-5}, 1)\)
- Compute \(L_{\text{cls}}\_\text{roi} \leftarrow C \times L_{\text{cls}}\_\text{roi}\)
- Compute \(L_{\text{roi}}\)
- Backpropagate losses through student network.

end for

- Output bounding box predictions.
Figure 8. A qualitative comparison of detection results from the source-trained model, ST3D [34], and our proposed framework. The ground truth bounding boxes for the “Car” class are in green and the detected bounding boxes are in red. The first and fourth rows show the results from the source trained model, the second and fifth from [34], while the third and final rows show the results of the proposed method.