Reliable Student: Addressing Noise in Semi-Supervised 3D Object Detection

Anonymous CVPR submission

Paper ID 49

Abstract

*Semi-supervised 3D object detection can benefit from the promising pseudo-labeling technique when labeled data is limited. However, recent approaches have overlooked the impact of noisy pseudo-labels during training, despite efforts to enhance pseudo-label quality through confidence-based filtering. In this paper, we examine the impact of noisy pseudo-labels on IoU-based target assignment and propose the **Reliable Student** framework, which includes two complementary approaches to mitigate errors. The first approach involves a class-aware target assignment strategy that reduces false negative assignments in challenging classes. The second approach involves a reliability weighting strategy that suppresses false positive assignment errors while also addressing remaining false negative ones from the first step. The reliability weights are determined by querying the teacher network for confidence scores of the student-generated proposals. Our work surpasses the previous state-of-the-art on KITTI 3D object detection benchmark on point clouds in the semi-supervised setting. On 1% labeled data, our approach achieves an improvement of 6.2% for the pedestrian class, despite having only 37 labeled samples available. The improvements become significant for the 2% setting, achieving 6.0% and 5.7% improvements for the pedestrian and cyclist classes, respectively.*

1. Introduction

Significant progress has been made in image classification [4] and object detection [2, 8, 13, 15–17, 27, 33] with recent developments in deep learning. The availability of large datasets [4, 11, 14, 20] has helped to accelerate these advancements. However, annotating massive datasets remains a bottleneck, particularly for 2D and 3D object detection. Semi-supervised approaches (SSA) have been proposed to address this problem. Unlike supervised methods, these approaches require only a limited amount of annotated data for model training, with the remaining data being unlabeled.

Several semi-supervised techniques have been proposed for object detection, including [5, 9, 12, 21, 22, 28]. Self-training using pseudo-labeling is the most commonly used method and has shown effectiveness in both object detection [9, 12, 19, 21] and classification [18, 29]. At its core, a student-teacher framework is used to incrementally train teacher and student models on unlabeled data in a mutually beneficial manner. The teacher model is initially trained on limited labeled data in a supervised manner to generate pseudo-labels (PL) to train the student model on unlabeled data via back-propagation. Mean-teacher based techniques [21, 22] use an exponential moving average (EMA) of the weights from the student model to update the teacher model's weights, which leads to more stable predictions on the unlabeled data.

Due to its limited pre-training on labeled data, the teacher model fails to generalize effectively, resulting in noisy pseudo-labels that hinder the learning of the student model. Existing methods overcome this problem by filtering out low-quality pseudo-labels with confidence-based thresholds, acting as a global quality-based filtering mechanism. However, even with strict filtering, the pseudo-labels still remain noisy, as shown in Fig. 1 (a). They have erro-

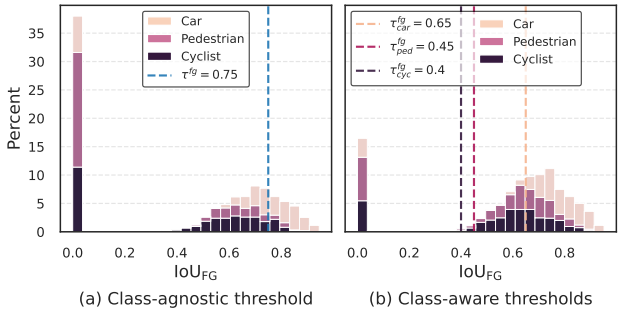

Figure 1. Illustrates the need for class-aware foreground thresholds for foreground/background target assignment. $\mathrm{IoU}_{\mathrm{FG}}$ on the x-axis shows the IoU of proposals with respect to pseudo labels, which are foreground relative to ground truths. (a) The default class-agnostic threshold in the PV-RCNN baseline. (b) Our class-aware thresholds. Lowering the threshold and including more foreground proposals can benefit challenging and uncommon classes. It also significantly reduces false negatives with IoUs close to zero. (Best viewed in color)

neous Intersection over Union (IoU) with proposals, which are foreground relative to ground truths. This poses a significant problem for downstream tasks such as target assignment in Region Proposal Network (RPN) and Region-based Convolution Neural Network (RCNN) modules, which rely on these noisy IoUs. The standard target assignment inevitably misclassifies the proposals with IoUs close to zero, *i.e*. the bar close to the y-axis in Fig. 1 (a), as background, leading to performance degradation.

Fig. 1 also shows distinct class-specific distributions of IoUs due to the different difficulty levels and the unbalanced distribution of classes in the dataset. Neglecting the difference in distributions poses a challenge for class-agnostic target assignment methods in detectors such as PV-RCNN. A high-value class-agnostic threshold exacerbates false-negative (FN) errors for difficult classes, such as pedestrians and cyclists, with lower distribution modes, while lowering the threshold causes many false positives (FP) for the car class, which is easier to learn.

We address these challenges from two perspectives: 1) reducing false-negative and false-positive errors using a new and simple class-aware target assignment approach, and 2) increasing the robustness of the training against the potential failure of our initial assignment by weighting the classification loss to suppress misclassified proposals. These two steps are complementary, with the first step aiming to minimize assignment errors by considering the difference between distribution modes of different classes, while the second step mitigates residual errors from the first step.

To this end, we first modify the target assignment process in two key areas where IoU scores are used. We replace the standard foreground/background random subsampling with a top-k IoU-based subsampler to promote learning from uncertain or difficult background proposals. We also propose local class-aware foreground thresholds for target assignment. As shown in Fig. 1 (b), the new thresholds include more foreground proposals of difficult classes (leading to higher recall) while preserving a high value for the dominant car class to ensure learning from high-precision proposals. The foreground and background thresholds divide proposals into three categories: foreground (FG), background (BG), and uncertain (UC). We assign hard labels to FG and BG proposals and use soft labels for those in the UC category to consider their uncertainty.

Second, to address false negative/positive target assignment errors, we propose to use the teacher to provide reliability scores for the student-generated proposals. To this end, the teacher's RCNN head refines the student's proposals and assigns confidence scores to them, which we use to weight the RCNN classification loss on unlabeled data using different FG/UC/BG weighting options. Our results show that weighting uncertain and background proposals effectively suppress false positives and false negatives, respectively, and outperforms other proposed weighting schemes.

In summary, our key contributions are as follows :

- We thoroughly investigate the impact of noisy pseudo labels on the IoU-based target assignment.
- We propose a class-aware target assignment method to address the target misclassification problem present in recent pseudo-labeling approaches.
- We propose different reliability weighting options to suppress false negatives and positives using teacher confidence scores.
- We conduct extensive experiments and ablation studies to evaluate the effectiveness of our approach on the KITTI 3D object detection benchmark in a semi-supervised setting.

2. Related Work

2.1. 3D Object Detection

Research on 3D object detection from point clouds focused on a bird's eye view of the lidar point cloud [3, 7]. However, VoxelNet [33] employed a different approach by dividing the point cloud into 3D voxels and encoding each voxel using a feature encoding layer. Although 3D convolution layers were applied to further aggregate features, this method was considered time-consuming due to the 3D convolutions involved. To address this, SECOND [27] proposed a spatially sparse convolutional network to improve the speed of previous methods. PointPillars [8] then suggested using vertical columns instead of voxels and a 2D convolutional network to encode features. This approach was found to be faster and more robust than previous methods. Another approach by PointNet and PointNet++ [15,16] was to work directly on encoding points instead of voxels, resulting in more efficient and flexible approaches. In this study, we use PV-RCNN [17], a robust two-stage detector that combines the VoxelNet and PointNet approaches and achieves high performance.

2.2. Semi-Supervised Object Detection

In the field of semi-supervised 2D object detection, numerous studies have been conducted. PseCo [9] combines both pseudo-labeling and consistency approaches. It uses not only label-level consistency but also feature-level consistency, which further improves the performance of the final detector. This approach also uses focal loss similar to [12] to alleviate the class-imbalanced in pseudo-labeling. [10] considers the localization task as a classification task and proposes a certainty-aware pseudo-label approach. By quantifying the quality score of classification and regression, they adjust the threshold used for generating pseudo-labels. Instant-Teaching [32] proposes to generate pseudo

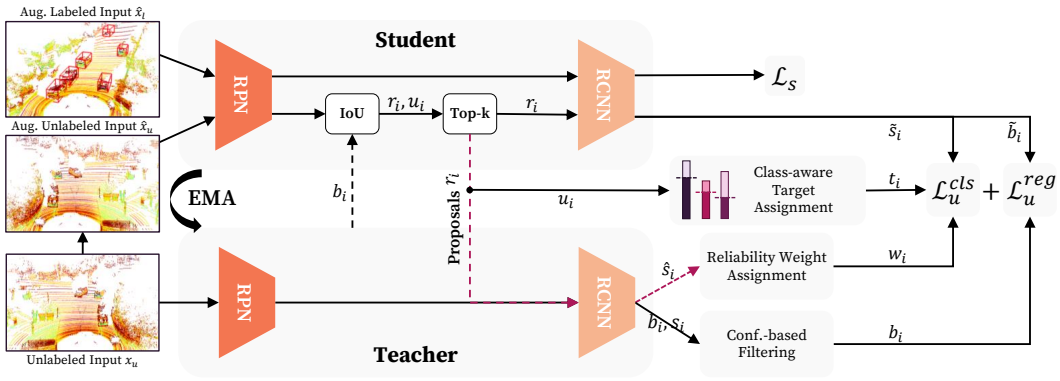

Figure 2. **Overview of our Reliable Student framework.** It uses a teacher-student network, where the EMA teacher produces high-quality pseudo-label boxes (b_i). In RCNN, we compute IoU (u_i) between b_i and post-NMS proposals (r_i) of the student, followed by a Top-k sampling of r_i based on u_i. Sampled proposals r_i are injected into student and teacher RCNN heads to predict objectness-based confidence scores $\tilde{s}_i$ and $\hat{s}_i$, respectively. While $\tilde{s}_i$ acts as an input to RCNN classification loss ($\mathcal{L}_u^{cls}$), $\hat{s}_i$ are converted into reliability weights w_i for $\mathcal{L}_u^{cls}$. The class-aware target assignment module uses thresholds for different classes on u_i to assign objectness targets (t_i) for $\mathcal{L}_u^{cls}$.

annotation for unlabeled data using a weak augmentation in mini-batch, then using these predicted annotations as ground truth of the same image with strong augmentation. For strong augmentation, the authors use Mixup [30].

Recent works also focused class-imbalance and confirmation bias issues. LabelMatch [1] deals with this in two ways. Firstly, it leverages the labeled data distribution for adaptive thresholding to filter out unbiased pseudo labels, and secondly, it re-calibrates the high-quality unreliable pseudo labels into reliable ones. Unbiased Teacher [12] attempts to address the class-imbalanced problem in pseudo-labeling by incorporating a focal loss that forces the model to focus on challenging samples from the underrepresented classes. Humble Teacher [21] achieved comparable results via soft labels instead of hard labels with a teacher ensemble network to improve the reliability of the pseudo labels.

Soft Teacher [26] deals with the misclassification of foreground proposals by suppressing the classification loss using the teacher's confidence scores. Our approach follows this but additionally considers the reliability of foreground targets with a foreground reliability weight. Our work also differs from Soft Teacher in that we use a third category of targets in the RCNN, called the Uncertain (UC) region, and assign soft labels to them. These targets may correspond to actual foreground or background boxes, and it is crucial to assign appropriate weights to this region to optimize the precision-recall trade-off further. Combating Noise [25] assumes the background proposals to be accurate and only deals with mitigating the noise due to foreground proposals by suppressing their losses. Whereas we show that dealing with both misclassified foreground as well as background proposals is important. Instead, class-specific thresholds representing the learning status of each class are used.

To date, there are few works on semi-supervised point-based 3D object detection, such as SESS [31] and 3DIoUMatch [24]. SESS employs asymmetric data augmentation techniques and enforces consistency between teacher-student predictions via different types of losses. 3DIoUMatch [24] proposes a pseudo-labeling approach for both indoor and outdoor 3D object detection. Inspired by FixMatch [18], they introduce a joint confidence-based pseudo-label filtering mechanism using predicted objectness and class probabilities. Additionally, they estimate IoU and use it as a localization quality to filter pseudo-labels. Unlike 3DIoUMatch, we only employ an objectness threshold, eliminating the complexity of employing multiple thresholds. Moreover, we adopt objectness supervision on unlabeled data, unlike 3DIoUMatch. Our findings indicate that this strategy enhances performance.

3. Method

3.1. Overview

An overview of our approach is depicted in Fig. 2. Our approach is based on the mean-teacher framework, in which the teacher creates PLs for unlabeled input to serve as a supervised signal for the student. The student is provided with the strongly augmented version of the unlabeled input as well as the labeled input, and its parameters are updated through backpropagation. The teacher's parameters, on the other hand, are gradually updated from the student's parameters using the exponential moving average strategy. To ensure the quality of generated PLs, we filter them based on their confidence scores. We introduce the Classwise Target Assignment module (Sec. 3.2) with class-aware foreground thresholds on IoU of proposals with PLs, with the goal of improving recall, particularly for challenging classes. This is based on the understanding that the learning status of

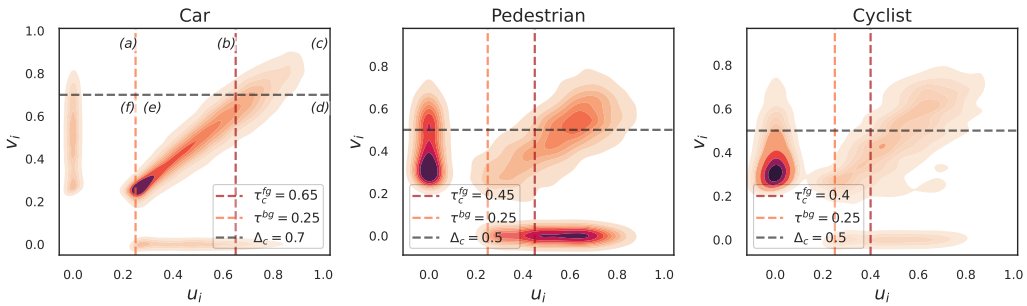

Figure 3. Illustrates the density of IoU values of proposals with their matched PL (u_i) and GT (v_i) on the x-axis and y-axis, respectively. Denser regions are shown with darker shades. The **red** and **orange** vertical lines denote the local foreground (FG) (τ_c^{fg}) and background (BG) (τ^{bg}) thresholds, while the **black** horizontal line represents the FG threshold (Δ_c) for evaluation mode, dividing the plot into six sub-regions. Sub-regions *(a)* and *(f)* represent false negative and true negative proposals, respectively. *(b)* and *(e)* depict proposals lying in the uncertain region and are assigned soft targets, while *(c)* and *(d)* depict true positive and false positive proposals, respectively. The proposals are obtained from the last few training iterations. We also omit proposals that are in the background with respect to both GT and PL for better visualization. All three plots follow the same sub-region breakdown. (Best viewed in color)

classes depends on their difficulty level and the availability of their instances in the dataset. Given these foreground thresholds and the default background threshold, we define hard classification targets for the foreground and background proposals, while uncertain proposals whose IoU lay between the FG and BG thresholds are assigned soft targets.

Due to the noisy IoU signal used for target assignment, some proposals may be mistakenly assigned to incorrect targets, leading to FPs and FNs. To mitigate this, we introduce the Classwise Weight Assignment module (Sec. 3.3) to assign reliability weights to the proposals of each category based on the dominant error type in that category, making the training more robust. To obtain the reliability weights, we use the teacher model to refine student proposals using its RCNN module and use its confidence score $\hat{s}_i$ as additional supervision to improve the student's performance. Given the student's RCNN refinement box and score $\{\tilde{b}_i, \tilde{s}_i\}$ and their corresponding targets, we use the teacher score $\hat{s}_i$ to weight the loss of classification on unlabeled data.

3.2. In-depth analysis of supervision from noisy pseudo labels

We investigate the problem of learning from noisy PLs, mainly used to supervise RPN and RCNN modules in the detector. We focus on the RCNN module and its classification target assignment, where the proposals are assigned with foreground/background labels.

Denote $\mathcal{P} = \{b_n, c_n, s_n\}_{n=1}^{N_{pl}}$ as the set of filtered PLs consisting of bounding box b_n, category label c_n, and the confidence score s_n. We define $\{r_i\}$ as the final set of proposals or Regions of Interest (RoIs) that the student generates after the IoU-guided filtering and deduplication of RPN proposals using Non-Maximum Suppression (NMS). Existing pseudo-labeling approaches use IoU between these RoIs and PLs to assign category labels and FG/BG targets to proposals of unlabeled data in RPN and RCNN modules of PV-RCNN, respectively. In RCNN, for a given proposal, if its maximum IoU with PLs, i.e., $u_i = \max_{p \in \mathcal{P}} \text{IoU}(r_i, p)$, exceeds a predefined class agnostic foreground threshold τ^{fg}, it is considered as a foreground proposal. We define these IoU thresholds used in these two modules as *local thresholds* (τ_c^{fg}), as opposed to the *global thresholds* (δ_c^{fg}), used to filter out low-quality PLs.

We analyze the sub-optimal classification target assignment from PLs with the optimal assignment from GTs. In Fig. 1, we evaluate the mean IoU of proposals that are foreground with respect to GTs, i.e., their IoUs with GTs are greater than the evaluation mode class-wise foreground threshold Δ_c^{fg}. We observe two crucial issues when using the standard target assignment.

First, the classes exhibit distinct mean IoU distributions. Therefore, the standard target assignment strategy based on a single class-agnostic foreground threshold, *e.g.* $\tau^{fg} = 0.75$, cannot reliably classify the proposals. For pedestrians and cyclists classes, which have lower distribution modes than the car, such a class-agnostic threshold results in many misclassified foreground proposals whose IoU cannot exceed the threshold by a small margin. To address this issue, we propose local class-aware foreground thresholds τ_c^{fg}, instead of a class agnostic τ^{fg}, on u_i IoUs, to construct the FG/BG target t_i for proposal r_i as follows:

$$t_i = \begin{cases} 1, & u_i > \tau_c^{fg} \\ \frac{u_i - \tau^{bg}}{\tau_c^{fg} - \tau^{bg}}, & \tau^{bg} \leq u_i \leq \tau_c^{fg} \\ 0, & u_i < \tau^{bg} \end{cases} . \quad (1)$$

Background proposals have consistently low IoUs, enabling a single class-agnostic threshold τ^{bg} to distinguish them from other proposals.

Second, the IoUs used for target assignment are noisy and unreliable. This is particularly the case for pedestrian

and cyclist classes which are challenging to detect/learn due to their object size and the imbalance class distribution of the dataset. Given the presence of excessively noisy IoUs, despite the implementation of class-specific local thresholds, the assignment carried out in Eq. (1) will inevitably result in the occurrence of false negative (FN) and false positive (FP) errors.

To examine how proposals in FG, UC, and BG categories are affected by the FP and FN errors, we illustrate the density plots in Fig. 3, showing the distribution of RoI IoUs relative to both PLs and GTs. The FP proposals are referred to as those which are foreground with respect to PL, but background with respect to GT, whereas those that are the opposite are referred to as FN proposals. As shown, each local class-aware threshold divides the plot into three columns showing FG, UC, and BG sections from right to left.

Ideally, we expect well-calibrated IoU scores such that the IoU of RoIs with respect to PLs are as close as possible to their corresponding IoUs with respect to GTs. However, in practice, there exist two sub-densities close to the axes contributing to the error. More specifically, in the foreground region, we observe the density of FP proposals in section (d), near the x-axis, for all classes. However, for the pedestrian class, we have significantly higher density compared to other classes. In the background region, FN proposals are present in (a) near the y-axis. The definitions of FP and FN have been extended to the uncertain region, i.e., sections (b) and (e), where FN and FP proposals are located in section (b) and at the bottom of section (e), close to the x-axis, respectively.

3.3. Learning from noisy pseudo labels via reliability-based weighting

To address these FP and FN erroneous proposals, we focus on making the training robust against a given set of uncertain PLs. We propose weighting the classification loss of such proposals based on the reliability of their target assignment, i.e., IoU between RoI and PL. We seek a reliability score that can consistently assign a low value to both FN and FP proposals. In this work, we evaluate the reliability score proposed by Soft Teacher. However, any other reliability score can also be plugged into our framework.

We estimate the reliability of the student's proposals based on their corresponding teacher's refined confidence scores. We use these scores to suppress the loss due to FP and FN targets. To this end, we first reverse the augmentation h on the student proposals before sending them to the teacher. The teacher refines each student's proposal r_i using its RoI pooling module and predicts $\hat{y}_i = \{\hat{b}_i, \hat{s}_i\}$, where $\hat{b}_i$ and $\hat{s}_i$ denote the corresponding refined bounding box and its confidence score, respectively. The confidence score $\hat{s}_i$, represents the foreground probability of the refined bounding box proposal, which acts as the reliability score for r_i.

We propose different reliability weighting schemes based on the teacher's confidence score $\hat{s}_i$, for the RCNN classification loss of unlabeled samples.

Based on our error breakdown in the previous section, we introduce reliability-based weighting options as follows:

- **Background proposals** (BG): suppress the FN proposals in sub-region (f) of Fig. 3 by incorporating the teacher's background score as a weight ($w_i = 1 - \hat{s}_i$) for classification loss in sub-regions (a) and (f).
- **Uncertain FN proposals** (UC_{FN}): suppress the FN proposals in sub-regions (b) of Fig. 3 by incorporating the teacher's background score as a weight ($w_i = 1 - \hat{s}_i$) for classification loss for sub-regions (b) and (e).
- **Uncertain FP proposals** (UC_{FP}): suppress the FP proposals in sub-region (e) of Fig. 3 by incorporating the teacher's foreground score as a weight ($w_i = \hat{s}_i$) for classification loss for sub-regions (b) and (e).
- **Foreground proposals** (FG): suppress the FP proposals in sub-region (d) of Fig. 3 by incorporating the teacher's background score as a weight ($w_i = 1 - \hat{s}_i$) for classification loss for sub-regions (c) and (d).

In all the weighting options, proposals belonging to the remaining categories are assigned with the reliability weight $w_i = 1$. Later in Sec. 4.3.1, we evaluate the application of different weighting options individually and in combinations and achieve the best performance from $\text{UC}_{\text{FP}} + \text{BG}$ by suppressing FPs from uncertain proposals and FNs from background proposals.

Moreover, in contrast to [17, 24], we further leverage these reliability-based weights to let the student model learn more about challenging and uncertain proposals instead of the easy backgrounds. For the target assignment in RCNN of the student model, the detector calculates the IoU between the post-NMS proposals and pseudo-labels. Prior works perform sampling on these IoUs such that, at most, 50% of the foreground proposals are randomly sampled before they are passed on for refinement. The remaining background proposals are further randomly sub-sampled, ensuring that 20% of them have low IoU (*e.g.* < 0.1), which are easy to classify as backgrounds. Our approach differs in that it avoids sub-sampling of such easy backgrounds on unlabeled data and instead uses a top-k sampling strategy on the IoU. Thus, allowing the model to learn more about the challenging backgrounds.

Let $\{\tilde{b}_i, \tilde{s}_i\}$ denote the student refinement of the proposal r_i. The RCNN classification loss on unlabeled data is summarized as follows:

$$\mathcal{L}_u^{cls} = \frac{\sum_i^{N_b} w_i l_{cls}(\tilde{s}_i, t_i)}{\sum_i w_i}, \tag{2}$$

where N_b are the total number of proposals for a single unlabeled sample.

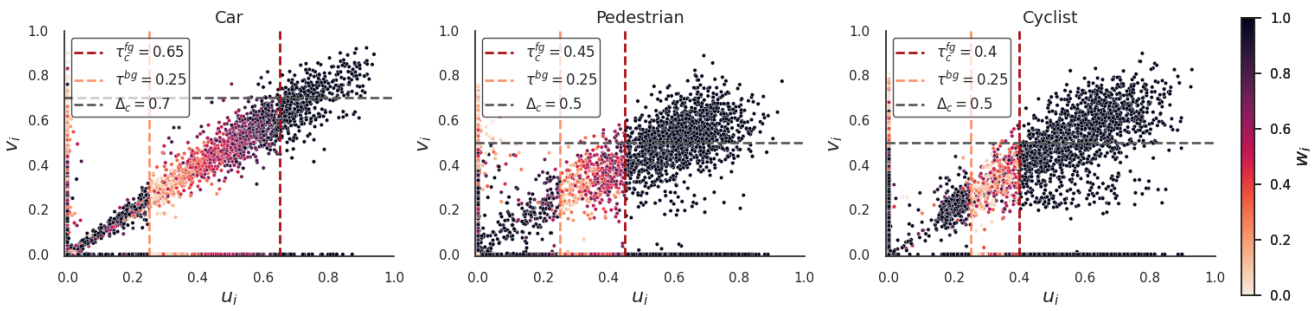

Figure 4. Illustrates the assigned reliability weights for RCNN classification loss based on proposal's IoU with PL (u_i) on the x-axis and GT (v_i) on the y-axis. The **red** and **orange** vertical lines depict local class-aware foreground (FG) (τ_c^{fg}) and background (BG) (τ^{bg}) thresholds, respectively, while the **black** horizontal line depicts the FG threshold (Δ_c) for evaluation mode. The color bar on the right shows the intensity of reliability weights. Plots are based on the last few training iterations for better visualization. (Best viewed in color)

Given N_l labeled samples, we define $\mathcal{D}_l = \{(x_i^l, y_i^l)\}_{i=1}^{N_l}$, where y_i^l contains the class labels and bounding box coordinates information, and use N_u unlabeled samples for $\mathcal{D}_u = \{x_i^u\}_{i=1}^{N_u}$. The unsupervised RCNN loss $\mathcal{L}_u$ consists of classification loss $\mathcal{L}_u^{cls}$ from Eq. (2), and box regression loss $\mathcal{L}_u^{reg}$, which is defined as:

$$\mathcal{L}_u^{RCNN} = \frac{1}{N_u}\sum_{i=1}^{N_u}(\mathcal{L}_u^{cls}(\tilde{s}_i^u, t_i^u) + \mathcal{L}_u^{reg}(\tilde{b}_i^u, b_i^u)), \quad (3)$$

where t_i^u is the target for classification loss from Eq. (1), and b_i^u is the bounding box of the assigned pseudo box based on u_i, acting as the regression loss target. We follow 3DIoUMatch for RCNN box regression loss $\mathcal{L}_u^{reg}$, as well as for RPN classification and regression losses to formulate the unsupervised loss $\mathcal{L}_u$. The supervised loss $\mathcal{L}_s$ is calculated similarly on labeled data using ground truth y_i^l. The overall loss of the student model is defined as

$$\mathcal{L} = \mathcal{L}_s + \lambda_u \mathcal{L}_u, \quad (4)$$

where λ_u is a coefficient balancing the unsupervised loss. The teacher weights are updated as the exponential moving average of the student model.

4. Experiments

4.1. Experimental Setup

We evaluate our method on KITTI [6] dataset, consisting of 7,481 training samples and 7,518 test samples. The training samples are divided into the train set (3,712 samples) for training the model and the validation set (3,769 samples) for evaluation. We use 1% and 2% labeled data splits with three folds each, provided by 3DIoUMatch [24]. For each fold, we carry out three trials with different random seed values and report the mean Average Precision (mAP) over all fold-trial combinations. The mAP is computed using a rotated IoU threshold of 0.7, 0.5, and 0.5 for car, pedestrian, and cyclist classes, respectively, at 40 recall positions. The experiments are conducted over all three object difficulty levels - Easy, Moderate, and Hard.

Implementation Details

For a fair comparison with [24], we utilize PV-RCNN [17] as the object detection backbone. We used OpenPCDet v0.5 framework [23] to implement our method and adapted the original 3DIoUMatch from OpenPCDet v0.3 to v0.5 for a fair comparison. The data augmentation on the student model is based on the 3DIoUMatch settings. Unlike 3DIoUMatch, which uses both RPN classification and RCNN objectness scores to filter pseudo labels, our approach utilizes only RCNN objectness threshold, i.e., $\tau_{car}^{pl} = 0.95$ for car, and $\tau_{ped}^{pl} = \tau_{cycl}^{pl} = 0.85$ for pedestrian and cyclist. Different to 3DIoUMatch, both the RPN and RCNN modules are supervised using labeled and unlabeled data through classification and regression losses, with the unlabeled loss weight $\lambda_u = 1$. On small amounts of data (1% and 2%), we pre-train PV-RCNN over 80 epochs with 10 repeated traversals in each epoch and use 60 epochs with 5 repeated traversals in each epoch for the training stage, similar to [24]. We employ a batch size of 8, consisting of 8 labeled and 8 unlabeled samples in both stages. For the evaluation stage, we used the student model.

4.2. Main Results

Tab. 1 shows the results of our approach, the original state-of-the-art 3DIoUMatch method, referred to as 3DIoUMatch†, and our adapted version of 3DIoUMatch, which is referred to as the baseline. Note that the original 3DIoUMatch does not use the RCNN classification loss on unlabeled data. Conversely, our approach benefits from unlabeled data supervision for RCNN classification. Hence, for a more accurate comparison, we have also included the results of our adapted baseline with RCNN classification loss on unlabeled data, which shows an improvement over the naive baseline. We refer to our method as the best option

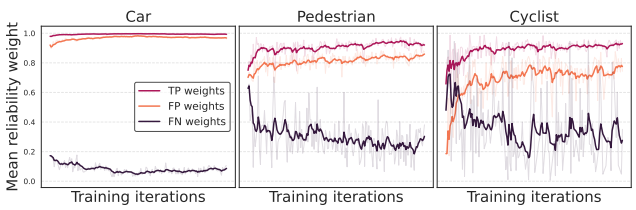

Figure 5. Teacher's mean reliability weights, averaged over every few iterations, using the FG + UC_{FP} + BG weighting type.

selected from the weighting schemes evaluated in Tab. 2, i.e., UC_{FP} + BG.

Our framework demonstrates superiority in performance over both the 3DIoUMatch and its improved version across all labeled data splits, particularly for the pedestrian and cyclist classes. Furthermore, the improvements become increasingly significant as the percentage of labeled samples increases. The results also indicate that our adapted 3DIoUMatch performs similarly to the original work, except for the cyclist class in the 2% split, where it exhibits a minor decrease in performance, less than 3%.

We evaluate our reliability-based weighting as shown in Tab. 2. To conduct a more detailed examination, we assess each option separately, as well as some interconnected options together. We consider the UC_{FN} and UC_{FN} + BG options to assess the effect of FN suppression. All other options are designed to evaluate the suppression of both FN and FP errors. The last two options are designated to assess how UC proposals should be weighted efficiently, i.e., to suppress FN or FP errors. All options except UC_{FN} show similar performance. The decrease in performance for UC_{FN} may be due to the down-weighting of truly uncertain proposals with incorrect ones. The results also suggest that utilizing the teacher's foreground score as a weight in the FG option is not as efficient as in the BG option.

4.3. Ablation Studies

4.3.1 Effects of reliability weights

Tab. 2 ablates the performance over different weighting options, improving the mAP over the baseline by 2.7%-3.2%. While the reliability weights help in all these options, UC_{FP} + BG has the highest gain in mAP of 3.2% over the baseline. In Fig. 5, we show the mean reliability weights of all foreground proposals relative to PLs with the weighting option of FG + UC_{FP} + BG. As shown, the weights from this option effectively suppress the loss due to FP and FN proposals at the cost of suppressing the loss of some true positives (TP). Moreover, the weights of FPs are relatively higher (close to 1), especially for the car class, and less effective than those for the FNs. We conjecture that this is due to the unbalanced number of FG/BG proposals in the RCNN module. Fig. 6 illustrates this by showing the percentage of FG proposals used to train the RCNN classification branch.

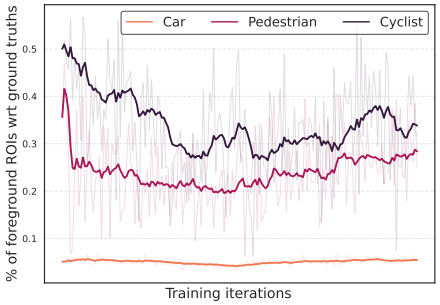

Figure 6. Shows the percentage of foreground proposals with respect to GT used for training FG/BG classification head, highlighting the imbalanced FG/BG ratios across different classes.

Note that the car class is highly skewed, with almost 95% of the proposals as BGs. As a result, the network is biased towards the BG class, and the teacher model cannot provide a reliable FG score for the FP proposals. Whereas, UC_{FP} + BG option compensates this by avoiding suppression of loss due to TP proposals, instead mainly suppressing the FPs and FNs, as shown in Fig. 4.

4.3.2 Effects of class-aware target assignment

In Tab. 3, we analyze the effects of local class-aware foreground thresholds over class-agnostic thresholds and their sensitivity to different values. Class-agnostic model is built by replacing the class-aware thresholds from our best-performing model with the default threshold of 0.75 for all classes. We show that the class-aware thresholds not only perform better than the default threshold with a great margin, but also they are consistent in performance across different values. We leverage our previous finding that pedestrian and cyclist classes require lower thresholds than the car class by adjusting our baseline thresholds by 10%.

4.3.3 Effects of top-k based sampler

Tab. 4 shows that using the balanced random sampler with the class-aware target assignment and unreliability weighing scheme improves the results over the baseline. However, our top-k sampler improves the baseline further by 0.2%-4.4% across different classes.

5. Conclusion

In this paper, we deal with the semi-supervised 3D object detection and show that while generating high-quality pseudo labels using quality-based filtering is a useful approach in this paradigm, it is not the only factor to consider. We investigate the impact of residual errors from noisy pseudo-labels even after filtering, particularly on a crucial IoU-based target assignment module, which is the first stage where pseudo-labels are employed. Additionally, we emphasize the significance of distinct learning curves

Methods	1%									2%								
	Car			Pedestrian			Cyclist			Car			Pedestrian			Cyclist		
	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
PV-RCNN† [17]	87.7	73.5	67.7	32.4	28.7	26.2	48.1	28.4	27.1	\	76.6	\	\	40.8	\	\	45.5	\
3DIoUMatch† [24]	89.0	76.0	70.8	37.0	31.7	29.1	60.4	36.4	34.3	\	78.7	\	\	48.2	\	\	56.2	\
PV-RCNN	87.6	74.1	67.9	36.5	31.7	28.9	49.9	28.8	27.3	88.9	76.8	71.9	45.1	40.4	35.6	63.0	42.3	38.9
3DIoUMatch (Baseline)	89.2	76.4	71.3	41.8	35.7	32.9	**59.9**	36.0	33.8	90.7	78.9	74.3	52.9	47.0	41.8	74.2	53.3	49.6
3DIoUMatch + ULB RCNN CLS	89.8	76.6	72.0	41.9	36.0	33.1	59.0	35.6	33.3	91.1	79.3	75.3	54.6	48.6	42.8	75.9	54.4	50.7
Reliable Student	**89.7**	**77.0**	**72.5**	**48.0**	**41.9**	**38.4**	59.1	**36.4**	**34.2**	**90.9**	**79.5**	**75.0**	**59.3**	**53.0**	**46.9**	**83.1**	**59.0**	**55.1**
% Improvement over Baseline	+0.5	+0.6	+1.2	+6.2	+6.2	+5.5	-0.8	+0.4	+0.4	+0.2	+0.6	+0.7	+6.4	+6.0	+5.1	+8.9	+5.7	+5.5

Table 1. Results on the KITTI evaluation set based on mAP over 40 recall positions. PV-RCNN† is the supervised-only baseline, and 3DIoUMatch† is the original work (both based on OpenPCDet v0.3). 3DIoUMatch (Baseline) is our adaptation of the original work to OpenPCDet v0.5, and 3DIoUMatch + ULB RCNN CLS is our modified version of the baseline with objectness supervision from unlabeled data. (†) denotes borrowed results from [24], (\) denotes non-available results, and **Bold** indicates the best results from OpenPCDet v0.5.

Methods	1%			2%			mAP %
	Car	Ped.	Cycl.	Car	Ped.	Cycl.	
Baseline	76.4	35.7	36.0	78.9	47.0	53.3	54.6
BG	76.8	40.5	36.7	79.1	**53.2**	57.2	57.3 (+2.7)
$\mathrm{UC_{FN}} + \mathrm{BG}$	76.9	41.6	36.6	79.4	51.3	58.1	57.3 (+2.7)
$\mathrm{UC_{FP}} + \mathrm{BG}$*	**77.0**	**41.9**	36.4	79.5	53.0	**59.0**	**57.8 (+3.2)**
$\mathrm{FG} + \mathrm{UC_{FN}} + \mathrm{BG}$	76.8	39.9	**37.2**	**79.6**	53.0	55.5	57.0 (+2.4)
$\mathrm{FG} + \mathrm{UC_{FP}} + \mathrm{BG}$	77.0	41.4	35.9	79.5	53.2	56.8	57.3 (+2.7)

Table 2. Ablation study on different reliability-based weighting options on 1% and 2% data splits for moderate difficulty level. For a fair comparison, we show the mAP across all classes in the last column, where $\mathrm{UC_{FP}} + \mathrm{BG}$ performs the best. (*) indicates our chosen weighting option and **Bold** indicates the best results.

	Methods	Car	Pedestrian	Cyclist
	Baseline	76.4	35.7	36.0
C-Ag	0.75	76.6	37.0	33.2
C-Aw	0.75, 0.55, 0.5	76.5	**41.9**	**36.6**
	0.65, 0.45, 0.4*	**77.0**	**41.9**	36.4
	0.55, 0.35, 0.3	76.9	41.1	36.5

Table 3. Ablation study of local class-aware (C-Aw) and class-agnostic (C-Ag) foreground thresholds. C-Aw thresholds are shown for the car, pedestrian, and cyclist (in the same order). We used 1% labeled data for the moderate difficulty level. (*) indicates our chosen thresholds, and **Bold** indicates the best results.

Methods	Car	Pedestrian	Cyclist
Baseline	76.4	35.7	36.0
Default sampler	76.8	37.5	35.5
Top-k sampler	**77.0**	**41.9**	**36.4**

Table 4. Ablation study of default random sampler and our top-k sampler. We use 1% labeled data for the moderate difficulty level.

for different classes and the necessity for class-specific target assignments, especially when using pseudo-labeling techniques. Our suggested thresholds are only applied to the RCNN target assignment of the PV-RCNN. However, we believe it is crucial to highlight the importance of incorporating quality-based filtering in an earlier stage of the pseudo-label generation process, e.g., in the student model's IoU-based target assignment, to prevent the propagation of errors. In addition, we utilize the teacher model to obtain a reliability score, which can serve as a basis for filtering out inaccurate signals and retaining clear supervision signals from unlabeled data during the learning process. Nonetheless, our research offers a structured framework for analyzing errors, which can be applied alongside other reliability-based metrics to enhance the overall reliability of the system. Our experiments demonstrate that our proposed method achieves state-of-the-art results across different settings and classes. In the future, we plan to expand its application to additional autonomous driving datasets and object detectors.

References

[1] Shicai Yang Yunyi Xuan JieSong Di Xie Shiliang Pu Mingli Song Yueting Zhuang. Binbin Chen, Weijie Chen. Label matching semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer, 2020. 1

[3] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. pages 6526–6534, Honolulu, HI, USA, 2017. IEEE. 2

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. pages 248–255. IEEE, 2009. 1

[5] Jinhao Dong and Tong Lin. Margingan: Adversarial training in semi-supervised learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 10440–10449, 2019. 1

[6] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *Int. J. Robotics Res.*, 32(11):1231–1237, 2013. 6

[7] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven L. Waslander. Joint 3d proposal generation and object detection from view aggregation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2018, Madrid, Spain, October 1-5, 2018*, pages 1–8. IEEE, 2018. 2

[8] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 12697–12705. Computer Vision Foundation / IEEE, 2019. 1, 2

[9] Gang Li, Xiang Li, Yujie Wang, Yichao Wu, Ding Liang, and Shanshan Zhang. Pseco: Pseudo labeling and consistency training for semi-supervised object detection. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IX*, volume 13669 of *Lecture Notes in Computer Science*, pages 457–472. Springer. 1, 2

[10] Hengduo Li, Zuxuan Wu, Abhinav Shrivastava, and Larry S. Davis. Rethinking pseudo labels for semi-supervised object detection. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 1314–1322. AAAI Press. 2

[11] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer. 1

[12] 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 *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 1, 2, 3

[13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9992–10002. IEEE. 1

[14] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Hanxue Liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Jie Yu, Chunjing Xu, and Hang Xu. One million scenes for autonomous driving: ONCE dataset. In Joaquin Vanschoren and Sai-Kit Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*. 1

[15] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 77–85. IEEE Computer Society, 2017. 1, 2

[16] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5099–5108, 2017. 1, 2

[17] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. PV-RCNN: point-voxel feature set abstraction for 3d object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10526–10535. Computer Vision Foundation / IEEE, 2020. 1, 2, 5, 6, 8

[18] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia

Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 1, 3

[19] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv e-prints*, page arXiv:2005.04757, May 2020. 1

[20] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 2443–2451. Computer Vision Foundation / IEEE. 1

[21] Yihe Tang, Weifeng Chen, Yijun Luo, and Yuting Zhang. Humble teachers teach better students for semi-supervised object detection. pages 3131–3140. IEEE, 2021. 1, 3

[22] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. 1

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

[24] He Wang, Yezhen Cong, Or Litany, Yue Gao, and Leonidas J. Guibas. 3dioumatch: Leveraging iou prediction for semi-supervised 3d object detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14615–14624. Computer Vision Foundation / IEEE, 2021. 3, 5, 6, 8

[25] Zhenyu Wang, Ya-Li Li, Ye Guo, and Shengjin Wang. Combating noise: semi-supervised learning by region uncertainty quantification. *Advances in Neural Information Processing Systems*, 34:9534–9545, 2021. 3

[26] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 3040–3049. IEEE. 3

[27] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, Oct. 2018. 1, 2

[28] Qize Yang, Xihan Wei, Biao Wang, Xian-Sheng Hua, and Lei Zhang. Interactive self-training with mean teachers for semi-supervised object detection. pages 5937–5946, Nashville, TN, USA, 2021. IEEE. 1

[29] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 18408–18419, 2021. 1

[30] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 3

[31] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Sess: Self-ensembling semi-supervised 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11079–11087, 2020. 3

[32] Qiang Zhou, Chaohui Yu, Zhibin Wang, Qi Qian, and Hao Li. Instant-teaching: An end-to-end semi-supervised object detection framework. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 4081–4090. Computer Vision Foundation / IEEE. 2

[33] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4490–4499. Computer Vision Foundation / IEEE Computer Society, 2018. 1, 2