3DIoUMatch: Leveraging IoU Prediction for Semi-Supervised 3D Object Detection

He Wang[1*] Yezhen Cong[2*] Or Litany[3] Yue Gao[2] Leonidas J. Guibas[1]
[1]Stanford University [2]Tsinghua University [3]NVIDIA

Abstract

3D object detection is an important yet demanding task that heavily relies on difficult to obtain 3D annotations. To reduce the required amount of supervision, we propose 3DIoUMatch, a novel semi-supervised method for 3D object detection applicable to both indoor and outdoor scenes. We leverage a teacher-student mutual learning framework to propagate information from the labeled to the unlabeled train set in the form of pseudo-labels. However, due to the high task complexity, we observe that the pseudo-labels suffer from significant noise and are thus not directly usable. To that end, we introduce a confidence-based filtering mechanism, inspired by FixMatch. We set confidence thresholds based upon the predicted objectness and class probability to filter low-quality pseudo-labels. While effective, we observe that these two measures do not sufficiently capture localization quality. We therefore propose to use the estimated 3D IoU as a localization metric and set category-aware self-adjusted thresholds to filter poorly localized proposals. We adopt VoteNet as our backbone detector on indoor datasets while we use PV-RCNN on the autonomous driving dataset, KITTI. Our method consistently improves state-of-the-art methods on both ScanNet and SUN-RGBD benchmarks by significant margins under all label ratios (including fully labeled setting). For example, when training using only 10% labeled data on ScanNet, 3DIoUMatch achieves 7.7 absolute improvement on mAP@0.25 and 8.5 absolute improvement on mAP@0.5 upon the prior art. On KITTI, we are the first to demonstrate semi-supervised 3D object detection and our method surpasses a fully supervised baseline from 1.8% to 7.6% under different label ratio and categories.

1. Introduction

Object detection is a key task in 3D scene understanding. It provides a concise representation of raw sensor measurements in the form of semantically meaningful 3D bounding boxes. This low-dimensional representation can already serve numerous applications in autonomous driving and AR/VR, as well as in robot navigation and manipulation. As a result, in recent years there has been a surge of interest in developing improved object detection pipelines and indeed current state-of-the-art methods show impressive performance. Yet, much of their success is attributed to the availability of large datasets of 3D scenes that are carefully annotated. While rapid advances in sensor technology facilitate the collection of 3D scenes at scale, annotating them remains the main bottleneck. This calls for detection methods that can leverage both labeled and unlabeled data at train time.

In this work, we aim to address this requirement by proposing a novel semi-supervised 3D object detection method which we dub 3DIoUMatch. As a generally applicable method, 3DIoUMatch can be applied to both indoor scene datasets, *i.e.* ScanNet[4] and SUN-RGBD[28], and outdoor datasets, *i.e.* KITTI[7]. We adopt popular point-based object detectors, VoteNet [18] and PV-RCNN [24], as our backbone object detection networks for the indoor and outdoor scenes, correspondingly. To provide supervision to the unlabeled scenes, we leverage a teacher-student mutual learning framework [29] and use the bounding box predictions from the teacher network as pseudo-labels to supervise the student network on unlabeled data. However, unlike most pseudo-label techniques that were designed for classification, in the highly complex (joint regression and classification) task of object detection, we observe that the pseudo-labels suffer from significant noise, and using them directly is suboptimal.

Inspired by FixMatch [26], the state-of-the-art semi-supervised learning (SSL) method for 2D image classification that proposed confidence-based filtering to improve pseudo-label quality, we adopt a pseudo-label filtering mechanism for 3D object detection by setting thresholds on predicted class probabilities (and objectness scores for VoteNet), so as to filter out teacher proposals with potentially erroneous semantic labels or ones not belong to foreground. While effective, these criteria alone are not sufficient to capture localization quality, and the pseudo-labels may still have large errors in the bounding box parameters.

*: equal contribution
Project page: http://THU17cyz.github.io/3DIoUMatch

To that end, we further propose to leverage estimated IoU (intersection over union) as a localization quality measure for pseudo-label filtering. IoU estimation was first proposed in the context of 2D object detection as a localization confidence in the pioneering work IoU-Net [12], where estimated IoU was proven successful in replacement of class confidence for test-time Non-Maximal Suppression (NMS). To the best of our knowledge, leveraging IoU estimation for pseudo-label filtering is a novel idea for SSL on both 2D and 3D object detection. Equipping the detectors with a 3D IoU estimation module, we are able to filter out poorly localized pseudo-labels and leverage estimated IoU for both train-time and test-time NMS.

A key challenge when filtering based on IoU estimation is how to properly set the threshold. Unlike objectness and class confidence for which high threshold values (e.g. 0.9) work well, 3D IoU is more sensitive to small errors. Setting the threshold too high would reduce the number of pseudo-labels to very few, from which little could be learned. To balance between quality and coverage, we propose a two-stage filtering process: first, using a relatively low IoU threshold; then, an IoU-guided class-aware *Lower-Half Suppression* (LHS) that removes only half of the highly-overlapping boxes with low predicted IoU. Our proposed LHS thus naturally sets a threshold that is both dynamic and class-aware. Our experiments show that LHS outperforms IoU-guided NMS, which suppresses all but the top one during semi-supervised training.

Our method consistently improves upon the previous state-of-the-art method, SESS [34], on both ScanNet and SUN-RGBD benchmarks by significant margins. When using only 10% labeled data on ScanNet, 3DIoUMatch outperforms SESS by 7.7 absolute improvement on mAP@0.25 and by 8.5 absolute improvement on mAP@0.5. When using 5% labeled data on SUN-RGBD, 3DIoUMatch outperforms SESS by 4.8 absolute improvement on mAP@0.25 and by 8.0 absolute improvement on mAP@0.5. On KITTI, we are the first to demonstrate semi-supervised 3D object detection work and surpass fully-supervised baseline by large margins under all label ratios.

Our main contributions can be summarized as follows:

1. We propose a novel semi-supervised method for 3D object detection in point clouds based on pseudo-label propagation along with a carefully designed filtering mechanism.
2. For the first time, we leverage predicted 3D IoU as a localization confidence score for pseudo-label filtering, and further propose IoU-guided Lower-Half Suppression for robust pseudo-label deduplication. This idea is generally applicable and can be coupled to different 3D detectors on both indoor and outdoor scenes.
3. We achieve markedly improved performance over the previous state-of-the-art semi-supervised 3D object detection methods on the two major indoor object detection benchmarks, ScanNet and SUN-RGBD, under low label ratios and fully labeled setting. As the first semi-supervised 3D object detection work on KITTI, we also achieve significant improvements compared to fully supervised method.

2. Related Works

Semi-Supervised Learning (SSL) Many of the recent SSL methods [2, 31, 1] leverage consistency regularization, first proposed in [23, 13], which enforces the model to predict consistently across label-preserving data augmentation of different intensity. Borrowing the concept from Mean Teacher [29], the model with frozen weight can be viewed as the teacher model, otherwise student model. Some methods [2], following Mean Teacher, make the teacher model as the EMA of the student model for further regularization. Pseudo labeling [15] is another popular class of SSL method which can also be treated as a kind of consistency regularization, as one output of the unlabeled data is enforced to be consistent with the other (the pseudo-labels) by being supervised with the other. To improve the quality of pseudo-labels, FixMatch [26], a state-of-the-art SSL work on image classification, has shown that the student network can improve significantly by setting a classification confidence threshold τ_{cls} and filtering out low-confidence predictions from the teacher. With the filtered pseudo-labels, the student model only gets supervised on the unlabeled data whose pseudo-labels are kept. Another key factor to the success of these methods is strong data augmentation. It has been shown crucial to many SSL works [23, 13, 31]. Recent works [1, 26] proposed to adopt even more powerful augmentation such as RandAugment [3] and Cutout [5].

Semi-Supervised Object Detection Since the beginning of the deep learning era, tremendous progress has been made in 2D object detection, e.g region-based detectors [9, 8, 22] and single-stage detectors [16, 21, 30]. Similarly in 3D object detection, a number of deep learning methods have been proposed for different 3D data modalities, e.g. RGBD-based detectors [19, 17], point-based detectors [33, 25, 14, 18], voxel-based detectors [35], point-voxel-based detectors [24], etc.

Despite the great progress in both 2D and 3D object detection, most works focused on a fully-supervised setting. A few works [10, 6] have proposed to leverage unlabeled data or weakly-annotated data for 2D object detection. Under a standard SSL setting as we follow, CSD [11] proposed a consistency regularization method to enforce the consistency between predictions from an image and its flipped version. STAC [27] adopts a two-stage scheme for training Faster R-CNN [22]: in the first stage it pre-trains a detector with labeled data only and then predicts the pseudo labels

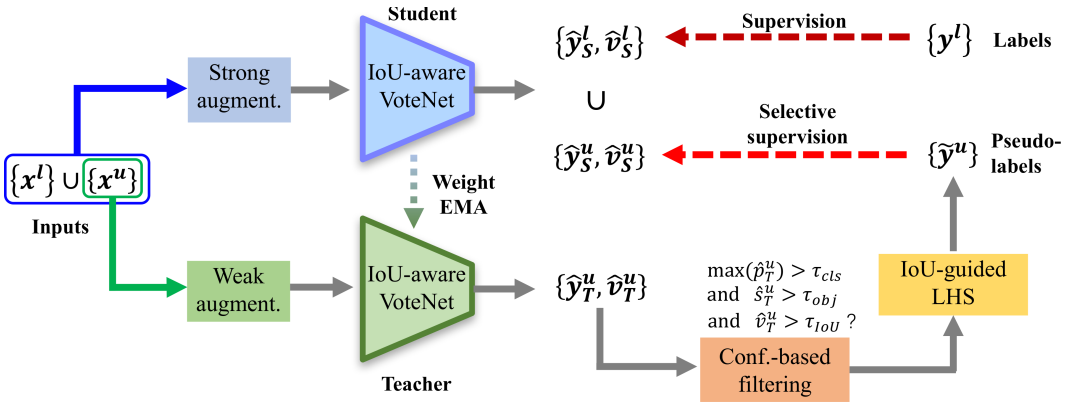

Figure 1. **3DIoUMatch pipeline at semi-supervised training stage.** We adopt as our backbone an extended version of VoteNet with an additional 3D IoU estimation module. For SSL, we utilize a teacher-student mutual learning framework, composed of a learnable student taking strongly augmented input data and an EMA teacher taking weakly augmented input samples. On labeled data, the student network is supervisedly trained. On unlabeled data, the student network takes pseudo-labels from its EMA teacher. To improve the quality of pseudo-label, we adopt a confidence-based filtering mechanism that filters out predictions that fail to pass all thresholds on class probability, objectness, and 3D IoU. We further use IoU-guided Lower-Half Suppression to remove the duplicated predictions. Using the filtered pseudo-labels, we selectively supervise the student predictions that are around the bounding boxes in the pseudo-labels.

for the unlabeled data; in the second stage, STAC leverages asymmetric data augmentation and the pseudo-label filtering mechanism to remove object proposals with low confidence. Note that the pseudo-labels are only generated once at the end of the first stage.

The only prior work on semi-supervised point-based 3D object detection, is SESS [34]. SESS is built upon VoteNet [18] and adopts a two-stage training scheme. It leverages a mutual learning framework composed of an EMA teacher and a student, uses asymmetric data augmentation, and enforces three kinds of consistency losses between the teacher and student outputs. Although SESS brings noticeable improvements upon a vanilla VoteNet when using only a small portion of labeled data, we find their consistency regularization suboptimal, as it is uniformly enforced on all the student and teacher predictions. In this work, we instead propose to apply confidence-based filtering to improve the quality of pseudo-labels from the teacher predictions and we are the first (in both 2D and 3D object detection) to introduce IoU estimation for localization filtering.

IoU Estimation IoU estimation was first proposed in a 2D object detection work IoU-Net [12], which proposed an IoU head that runs in parallel to bounding box refinement and is differentiable w.r.t. bounding box parameters. IoU-Net adds an IoU estimation head to several off-the-shelf 2D detectors and uses IoU estimation instead of classification confidence to guide NMS, which improves the performance consistently over different backbones. Thanks to its differentiability, IoU-Net can perform IoU optimization on bounding box parameters for iterative refinement, which further brings noticeable performance improvement.

For 3D object detection, STD [32] follows IoU-Net to add a simple IoU estimation branch parallel with the box estimation branch and to guide NMS with IoU estimation. PV-RCNN [24] devises a similar 3D IoU estimation module and use it at IoU-guided NMS stage. These two modules, unfortunately, are not suitable for IoU optimization as the features fed to the IoU estimation branch are not differentiable w.r.t. the bounding box size. Since the original VoteNet is not equipped with an IoU module, we devise a differentiable point-cloud-based 3D IoU estimation module is simple yet effective that can support the IoU optimization.

3. Method

In this section, we describe our solution in detail. We first formulate our problem in 3.1 and then summarize the two object detection backbones, PV-RCNN and VoteNet, in 3.2. We use VoteNet as an example to illustrate our proposed 3DIoUMatch pipeline in 3.3. We further explain how we use the estimated 3D IoU for pseudo-label filtering and deduplication in 3.4. Finally, we illustrate how we leverage the pseudo-labels for supervision in 3.5.

3.1. Problem Definition

Given a 3D point cloud representation of a scene $\boldsymbol{x} \in \mathbb{R}^{N\times 3}$ containing a set of objects $O = \{o^{(j)}\}$, we aim at detecting the amodal oriented 3D bounding boxes of all objects in O, along with their semantic class labels. In particular, we are interested in accomplishing this task under challenging conditions of limited supervision where we have access to a (small) set of labeled scenes $\{\boldsymbol{x}_i^l, \boldsymbol{y}_i^l\}_{i=1}^{N_l}$ and

a set of unlabeled scenes $\{\boldsymbol{x}_i^u\}_{i=1}^{N_u}$, where N_l and N_u are the number of labeled and unlabeled scenes, respectively. For a labeled scene $\boldsymbol{x}$, the label $\boldsymbol{y}$ comprises bounding box parameters $\{\boldsymbol{b}^{(j)}\}$ and semantic class labels $\{q^{(j)}\}$ of all ground truth objects $\{o^{(j)}\}$.

3.2. IoU-aware 3D Object Detection

We experiment our SSL method on two 3D detectors, VoteNet [18] and PV-RCNN [24]. VoteNet is a single-stage indoor point cloud detector while PV-RCNN is a two-stage outdoor point cloud detector. They both take point clouds only for inputs and output a list of bounding boxes after Non-Maximum Suppression (NMS) for each scene, which contain the prediction of center, size, orientation and semantic class. However, their architecture are very different partly due to the great discrepancy between indoor and outdoor scenes.

Indoor scene detector: VoteNet VoteNet [18] is built upon PointNet++ [20] backbone, and first processes the input point cloud $\{x_i\}_{i=1}^{N}$ to generate a sub-sampled set of $M < N$ seed points enriched with high-dimensional features $\{[x_i; f_i] \in \mathbb{R}^{3+C}\}_{i=1}^{M}$. Next, each seed point votes for the center of the object it belongs to, and the votes are grouped into K clusters. Finally, each of the K vote clusters is aggregated to make a prediction of a 3D bounding box parameters $\boldsymbol{b}^{(k)}$, a corresponding objectness score $s_k = \mathrm{P}(\boldsymbol{b}^{(k)}$ is an object), and a probability distribution $\{p_{cls}\}$ over L possible semantic classes. The bounding box parameters $\boldsymbol{b}$ are its center location $\mathbf{c} \in \mathbb{R}^3$, scale $\mathbf{d} \in \mathbb{R}^3$, and orientation θ around the upright axis.

At train time, VoteNet jointly minimizes a weighted combination of the following target losses: vote coordinate regression, objectness score binary classification, box center regression, bin classification and residual regression for heading angle, scale regression, and category classification. At test time, VoteNet applies Non-Maximum Suppression (NMS) based on objectness score to remove duplicated bounding boxes. Here, we instead rely on a 3D IoU estimation module designed for VoteNet. For more details, refer to the supplementary materials.

Outdoor scene detector: PV-RCNN PV-RCNN[24] is a high-performance and efficient LiDAR point cloud detector that deeply integrates both 3D voxel CNNs and PointNet++-style set abstraction to learn more discriminative point cloud features. Specifically, PV-RCNN first passes the 3D scene through a novel voxel set abstraction module based on sparse 3D CNN to get a set of keypoints with representative scene features. Then RoI grid pooling is then applied to the keypoints to abstract proposal-specific features into RoI grid points. The RoI grid points containing rich context information are finally used to accurately estimate bounding box parameters.

PV-RCNN itself incorporates an IoU-estimation module which can predict the IoU of each bounding box and use it to guide the sorting of the boxes.

3.3. 3DIoUMatch for SSL on 3D object detection

We take VoteNet as our example and our method with PV-RCNN is similar. With the incorporation of 3D IoU module into VoteNet, we construct an IoU-aware VoteNet for SSL on 3D object detection. Our proposed solution is comprised of two training stages: a pre-training stage, where we train our IoU-aware VoteNet on the labeled data, followed by an SSL stage where the entire data is utilized by pseudo-labeling the unlabeled scenes.

Pre-training. We start by training our IoU-aware VoteNet in a supervised manner, using the labeled set $\{\boldsymbol{x}_i^l, \boldsymbol{y}_i^l\}_{i=1}^{N_l}$. The training loss is a sum over the original VoteNet losses L_{votenet} and 3D IoU loss L_{IoU}. Once converged, we clone the network to create a pair of student and teacher networks.

Semi-supervised training through a teacher-student framework. We follow a teacher-student mutual learning framework [29] and train our networks on both labeled $\{\boldsymbol{x}_i^l, \boldsymbol{y}_i^l\}_{i=1}^{N_l}$ and unlabeled data $\{\boldsymbol{x}_i^u\}_{i=1}^{N_u}$. Each training batch contains a mixture of $\{\boldsymbol{x}_i^l\}_{i=1}^{B_l}$ labeled samples and $\{\boldsymbol{x}_i^u\}_{i=1}^{B_u}$ unlabeled samples.

For labeled samples, we supervise the student network using ground truth supervisions (as done in the pre-training stage) whereas for unlabeled samples, the student networks is supervised using pseudo-labels $\{\tilde{\boldsymbol{y}}_i^u\}_{i=1}^{N_u}$ generated from the teacher network. The final loss is formed as:

$$L = L_l(\{\boldsymbol{x}_i^l\}_{i=1}^{N_l}, \{\boldsymbol{y}_i^l\}_{i=1}^{N_l}) + \lambda_u L_u(\{\boldsymbol{x}_i^u\}_{i=1}^{N_u}, \{\tilde{\boldsymbol{y}}_i^u\}_{i=1}^{N_u})$$

where λ_u is the unsupervised loss weight.

To succeed in semi-supervised learning, it is crucial for the teacher network to generate high-quality pseudo-labels and maintain a reliable performance margin over the student network throughout the training. As commonly used in SSL literature, e.g. Mean Teacher [29] and SESS [34], we adopt an EMA teacher. We further leverage asymmetric data augmentation and pseudo-label filtering (see Sec.3.4).

To be in a position of advantage, the teacher network takes input data with weak augmentation only while the student network uses stronger data augmentation. We share the same data augmentation strategy with SESS. The input point clouds to our teacher network are augmented only by random sub-sampling while the inputs to the student network further undergo a set of stochastic transformation $\mathcal{T}$, including random flip, random rotation around the upright axis, and a random uniform scaling.

3.4. Pseudo-Label Filtering and Deduplication

In the teacher-student framework, the performance gap between the teacher and the student is usually quite

marginal given that these two models are just different by EMA on weight and data augmentation strength. Hence, it is not always true that the teacher prediction is more accurate than the student's on a specific training sample. On unlabeled data, the student model will only benefit from the pseudo-labels that are more accurate than its predictions. Therefore we should filter out low-quality predictions from the teacher model and only supervise the student model with the rest of the teacher model predictions.

Jointly filtering based on class, objectness, localization confidences For VoteNet, we propose to set an objectness threshold τ_{obj} and filter out bounding box predictions with objectness score $s < \tau_{obj}$. We further propose to set a classification confidence threshold τ_{cls} for filtering out predictions that are likely to contain a wrong class label.

Note that none of these two confidence measures capture the accuracy of bounding box parameter predictions. We propose to predict a 3D IoU for each predicted bounding box, use the 3D IoU estimation as a localization confidence, and set a localization threshold τ_{IoU} to filter out poorly localized predictions. Formally, we remove all the predictions that fail to satisfy all three confidence thresholds, i.e. $s > \tau_{obj}$, $\max(p_{cls}) > \tau_{cls}$, and $v > \tau_{\text{IoU}}$.

IoU-guided lower-half suppression for deduplication After the confidence-based filtering, there is still a lot of duplicated bounding box predictions that may introduce harmful noise to our pseudo-labels. NMS is a standard process in object detection for duplicate removal before evaluation, which takes a set of highly overlapped bounding box predictions that share the same class prediction, ranks them according to a confidence score and removes all but the top-1 prediction. STAC [27] applies class confidence based NMS to teacher predictions during pseudo-label generation.

The default NMS used in VoteNet is based on objectness confidence. Given that objectness score doesn't capture the localization quality, a train-time IoU-guided NMS will naturally perform better (see Table.2), where we use the product of predicted IoU and predicted objectness as the ranking metric. However, using the top one selected by IoU-guided NMS can still be suboptimal, since the predicted IoU will inevitably carry some errors. We argue that different from the test time scenario, pseudo-labels do not need to be fully deduplicated. Imagine this situation: if a bounding box predicted by the student is $0.2m$ to the left of its corresponding ground truth, it is a foreground object and will get bounding box supervision in VoteNet. However, if unfortunately the pseudo-label survives after non-maximal suppression is to the right of the ground truth more than $0.1m$, this predicted bounding box may lose supervision and be treated as a background box. This example shows that strict non-maximal suppression can lead to a smaller number of student model predictions that can receive supervision. Since we cannot know the best pseudo label among a bunch of highly-overlapped ones, it's fine to be less strict. To this end, we propose a novel Lower-Half Suppression, or in short, LHS, that only discards half of the proposals with lower predicted IoU. We argue that since LHS suppresses bounding boxes sharing the same class label, this suppression can be seen as a second-step class-aware self-adjusted filtering, which sets dynamic thresholds among the overlapping bounding boxes to keep the ones with higher confidence and hence find a better balance between pseudo-label quality and the amount of supervision. We also use the product of predicted IoU and predicted objectness as the confidence metric.

Final-step pseudo-label processing After the filtering and IoU-guided LHS, we now have high-quality predictions $\{\hat{y}_T^u\}_{k=1}^{K'}$ from the teacher network, where K' is the number of bounding boxes remains. Given that the student model inputs go through a stronger augmentation including an additional geometric transformation $\mathcal{T}$, in synchronize with the student model inputs, the bounding box parameters of the pseudo-labels need to go through the same transformation $\mathcal{T}$, namely $\tilde{\boldsymbol{b}}^u = \mathcal{T}(\hat{\boldsymbol{b}}_T^u)$. We further take convert the predicted class probability distribution $\hat{p}_T^u$ into semantic class label via $\tilde{q}^u = \max(\hat{p}_T^u)$. Now we obtain the filtered pseudo-labels $\tilde{y} = \{\tilde{\boldsymbol{b}}^u, \tilde{q}^u\}_{k=1}^{K'}$.

3.5. Selective Supervision using Pseudo-Labels

For our generated pseudo-labels, there is no guarantee that the labels can cover all the ground truth objects from O due to the filtering and potentially inaccurate teacher predictions. Given the incompleteness of our filtered pseudo-labels, we are relatively confident about the bounding boxes in this set but student predictions far away from all of our pseudo-labels are not necessarily negative. Our experiments show that supervising objectness on unlabeled data using the pseudo-labels seriously hurts the performance. For similar reasons, we do not supervise vote loss, which is a unique element in VoteNet and not shown in other detectors. For more analysis and experimental proof for this, we refer the readers to the supplementary materials. In this case, we will only supervise the bounding boxes in the vicinity of the pseudo bounding boxes and aim to improve their bounding box quality. More specifically, we stick to the way how VoteNet select foreground objects for bounding box parameter supervision: we supervise bounding box parameters and class for a prediction only if the vote that generates this prediction is within 0.3m of any bounding box in the pseudo-labels. For this set of pseudo-foreground predictions, we adopt the same way that VoteNet establishes association and enforce original VoteNet losses except for objectness loss and vote loss.

Dataset	Model	5%		10%		20%		100%	
		mAP @0.25	mAP @0.5	mAP @0.25	mAP @0.5	mAP @0.25	mAP @0.5	mAP @0.25	mAP @0.5
ScanNet	VoteNet	27.9±0.5	10.8±0.6	36.9±1.6	18.2±1.0	46.9±1.9	27.5±1.2	57.8	36.0
	SESS reported	\	\	39.7±0.9	18.6	47.9±0.4	26.9	62.1	38.8
	SESS	32.0±0.7	14.4±0.7	39.5±1.8	19.8±1.3	49.6±1.1	29.0±1.0	61.3	39.0
	Ours	**40.0±0.9**	**22.5±0.5**	**47.2±0.4**	**28.3±1.5**	**52.8±1.2**	**35.2±1.1**	**62.9**	**42.1**
	Abs. improve.	+8.0	+8.1	+7.7	+8.5	+3.2	+6.2	+1.6	+3.1
SUN-RGBD	VoteNet	29.9±1.5	10.5±0.5	38.9±0.8	17.2±1.3	45.7±0.6	22.5±0.8	58.0	33.4
	SESS reported	\	\	42.9±1.0	14.4	47.9±0.5	20.6	61.1	37.3
	SESS	34.2±2.0	13.1±1.0	42.1±1.1	20.9±0.3	47.1±0.7	24.5±1.2	60.5	38.1
	Ours	**39.0±1.9**	**21.1±1.7**	**45.5±1.5**	**28.8±0.7**	**49.7±0.4**	**30.9±0.2**	**61.5**	**41.3**
	Abs. improve.	+4.8	+8.0	+3.4	+7.9	+2.6	+6.4	+1.0	+3.2

Table 1. Comparison with VoteNet and SESS on ScanNet val set and SUN RGB-D val set under different ratios of labeled data. We report the mAP@0.25 and mAP@0.5 as mean±standard deviation across 3 runs under different random data splits. Due to the randomness of the data splits and our better pre-training protocol, SESS results provided by us are higher than those reported in the paper on mAP@0.5, and the mAP@0.25 results differ a little (the only difference is the pre-trained weights and data splits). The final improvement is the absolute improvement of our method over SESS results provided by us. Following SESS, we also report the results with 100% labeled data, where we simply make a copy of the full dataset as unlabeled data and train our method.

4. Experiments

4.1. Datasets and Evaluation Metrics

Indoor Datasets: ScanNet and SUNRGB-D We evaluate our VoteNet-based 3DIoUMatch on two major indoor datasets, ScanNet [4] and SUN RGB-D [28]. ScanNet is an indoor scene dataset consisting of 1513 reconstructed meshes, among which 1201 are training samples and the rest are validation samples. SUN RGB-D contains 10335 RGB-D images of indoor scenes which are split into 5285 training samples and 5050 validation samples. For both datasets, we follow [18, 34] for pre-processing data and labels to train our method and we report mAP@0.25 (mean average precision with 3D IoU threshold 0.25) and mAP@0.5 in the following experiments.

Outdoor Dataset: KITTI As for our PV-RCNN-based 3DIoUMatch, we use KITTI for evaluation. KITTI [7] is a very popular dataset for autonomous driving which consists of fine annotations for 3D detection. There are 7481 outdoor scenes for training and 7518 for testing, and the training samples are generally divided into a train split of 3712 samples and a validation split of 3769 samples. We follow [24] for data pre-processing and report the mAP with 40 recall positions, with a rotated IoU threshold 0.7, 0.5, 0.5 for the three classes, car, pedestrian, and cyclist, respectively.

4.2. Experiments on Indoor Scene Datasets

For experiments on indoor datasets, *i.e.*, ScanNet and SUNRGB-D, we use IoU-aware VoteNet as our backbone detector.

4.2.1 Result Comparison

Table 1 shows the results of our method compared to SESS and VoteNet under different ratios of labeled data on ScanNet and SUN RGB-D, respectively. The results illustrate that, with our effective train-time filtering and test-time improvement leveraging IoU estimation, we are able to significantly outperform current state-of-the-art, SESS, under all labeled ratio settings. With 5% labeled data, our method outperforms SESS by **8.1** and **8.0** on mAP@0.5 on ScanNet and SUN RGB-D, respectively. Note that our method gains more improvement on mAP@0.5, thanks to the high quality of pseudo labels and the IoU guidance for test-time NMS.

4.2.2 Ablation Study

Filtering and Deduplication Mechanism. We study the effect of each component of the filtering and deduplication mechanism. In Table 2, the second row shows the results of naive pseudo labeling, which takes all predictions from the teacher model for supervision. Expectedly the results are not satisfying, only a little higher than VoteNet. Simply applying the dual filtering of classification and objectness confidence gives significant improvement, as the filtering picks out the teacher model proposals that are very likely to be close to true objects and have the correct class. The conventional objectness-based NMS in VoteNet, however, fails to improve further, since the remaining proposals already have high objectness scores and the objectness-based NMS is not capable of picking the ones with higher localization accuracy.

As shown in the fifth and sixth row, after we introduce IoU during train time, IoU filtering and train-time IoU-

Obj&Cls Filter	IoU Filter	Train-time Suppression	Test-time Suppression	Test-time IoU opt.	ScanNet 10%		SUN-RGBD 5%	
					mAP@0.25	mAP@0.5	mAP@0.25	mAP@0.5
			Obj-NMS		38.4	19.8	32.9	12.5
✓			Obj-NMS		44.5	24.7	36.9	17.5
✓		Obj-NMS	Obj-NMS		44.2	25.2	37.1	17.4
✓	✓	IoU-NMS	Obj-NMS		45.9	26.8	37.4	18.7
✓	✓	IoU-LHS	Obj-NMS		46.5	26.9	37.9	18.5
✓	✓	IoU-LHS	IoU-NMS		47.0	28.2	38.8	20.8
✓	✓	IoU-LHS	IoU-NMS	✓	**47.2**	**28.3**	**39.0**	**21.1**

Table 2. Effects of the different components, including train-time filtering and deduplication, and test-time improvements.

guided NMS contribute to better performance under both settings. Our proposed IoU-guided LHS improves over IoU guided NMS on mAP@0.25, since LHS finds a better balance between quality and coverage. With better filtering and deduplication leveraging IoU estimation during train time, we gain 2.3 and 1.7 absolute improvement over the without-IoU version on mAP@0.25 and mAP@0.5 respectively on ScanNet 10%. This verifies that considering localization confidence is important for getting high-quality pseudo labels. With test-time improvements, our method gains in total 3.0 and 3.1 absolute improvement respectively.

We set 0.9 for both classification and objectness confidence threshold following STAC [27] and investigate the effect of different IoU thresholds on ScanNet 10%, as shown in Figure 2. The performance (with test-time improvements) is higher than the without-IoU baseline by large margins when $\tau_{IoU} \leq 0.5$. Note that the performance peaks at $\tau_{IoU} = 0.25$ for mAP@0.25 while peaking at 0.5 for mAP@0.5, simply because mAP@0.5 prefers a stronger filtering on localization quality. When $\tau_{IoU} > 0.5$, further increasing τ_{IoU} may lead to a drastic drop in pseudo-label coverage and hence is detrimental to the performance.

Test-time IoU-guided NMS and IoU optimization. We then evaluate the improvement brought by using IoU estimations at test time. The last two rows in Table 2 shows that IoU-guided NMS and IoU optimization improves the performance further.

4.2.3 Result Analysis

In this section, we examine how our 3DIoUMatch works during training on ScanNet 10%. The upper two curves in Figure 3 show that as the training goes, the performance on unlabeled data and test data increases conformably, which indicates the increasing quality of pseudo-labels. We also show how the coverage of the pseudo-labels changes on the unlabeled data over the training. Here coverage at a certain threshold simply means the class-agnostic recall, measuring the percentage of ground truth objects that can find a pseudo-label with an IoU larger than the threshold.

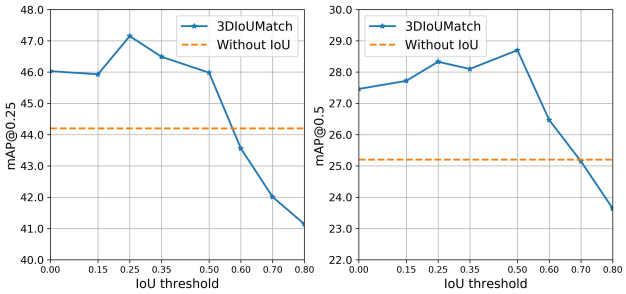

Figure 2. 3DIoUMatch results with different IoU thresholds on ScanNet 10%.

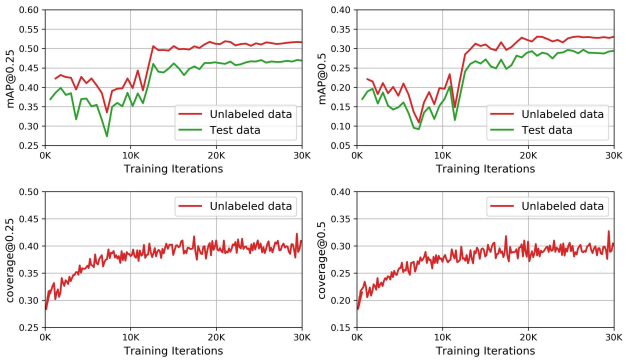

Figure 3. The performance improvements and pseudo-label coverage over the semi-supervised learning stage on ScanNet 10%.

As we can see from the lower two curves in Figure 3: at the beginning, the coverage of the pseudo-labels is relatively low due to the strict filtering mechanism; as the semi-supervised learning goes on, the improving detection performance leads to a higher passing rate of the filter and hence a higher coverage of the pseudo-labels, which in return fuels SSL; by the end of training, the coverage at 0.25 and at 0.5 both increase by about 10%.

4.2.4 Implementation Details

Training For the pre-training stage, we train with a batch size of 8 and follow the same data augmentation of

SESS [34]. We then use those pre-trained weights to initialize the student and teacher networks. For the SSL stage, we construct each batch by taking 4 labeled samples and 8 unlabeled samples, with the same data augmentation. The weights of different loss terms are the same as VoteNet and we set $\lambda_u = 2$. The student network is trained for 1000 epochs (the labeled data is traversed in one epoch), optimized by an ADAM optimizer with an initial learning rate of 0.002, and the learning rate is decayed by 0.3, 0.3, 0.1, 0.1 at the 400$^{\text{th}}$, 600$^{\text{th}}$, 800$^{\text{th}}$ and 900$^{\text{th}}$ epoch, respectively. The number of generated 3D proposals is 128. We use $k = 3, D = 4$ for the IoU module. The three thresholds are set to be $\tau_{obj} = 0.9, \tau_{cls} = 0.9, \tau_{IoU} = 0.25$. For more details, we refer the readers to the supplementary materials.

Inference We forward the input to the student network to generate proposals. We first apply IoU optimization to refine box parameters following IoU-Net [12], followed by an IoU-guided NMS with a 3D IoU threshold of 0.25.

4.3. Experiments on KITTI

For experiments on the KITTI dataset, we adopt PV-RCNN[24] as our backbone. PV-RCNN itself comes with a 3D IoU module, allowing to use it in our semi-supervised learning pipeline without modifying its architecture.

4.3.1 Results

We evaluate our method on KITTI validation set. Table 3 demonstrates significant and consistent improvement across all categories with 1%, 2%, and 100% labeled data, compared to supervised training only. Similar to our experiments on indoor scene datasets, here the 100% labeled data setting means making a copy of the full dataset as unlabeled data and train the network using our devised semi-supervised pipeline. With 1% labeled data, our method outperforms the labeled-data-only baseline by 6.6 mAP@0.5 on pedestrian, which is the most challenging class.

4.3.2 Ablation Study

We conduct ablation studies on KITTI with 1% labeled data. Table 4 shows our improvements compared with a pseudo-label baseline and filtering based on class confidence only. The results validate the effectiveness of our IoU-based localization confidence filtering.

4.3.3 Implementation Differences with VoteNet

In KITTI, we only care about three classes, car, pedestrian, and cyclist, which differ a lot in the difficulty to detect. Instead of using LHS, we follow PV-RCNN to set different IoU thresholds for each individual class, *i.e.*, $\tau_{car} = 0.8$, $\tau_{ped} = \tau_{cyc} = 0.4$. We selectively supervise the predictions that meet the foreground bar in PV-RCNN according to our pseudo-labels. In contrast to VoteNet, PV-RCNN is a two-stage detector containing an RPN. Bounding box objectness (or foreground probability) has been predicted in the RPN and used to pick top 100 proposals to forward to the RoI head. We therefore only additionally filter according to classification confidence with the threshold $\tau_{cls} = 0.2$. Due to the non-differentiability of the IoU module of PV-RCNN, we also do not apply IoU optimization.

We also adopt a two-stage training scheme for our PV-RCNN-based 3DIoUMatch. We use an unlabeled weight $\lambda_u = 2$ and only supervise anchor classification and bounding box regression in PV-RCNN on unlabeled data. Please refer to the supplementary materials for more details.

	1%			2%			100%		
	Car	Ped.	Cyc.	Car	Ped.	Cyc.	Car	Ped.	Cyc.
PVR.	77.3	47.8	62.9	80.4	47.1	63.5	83.0	57.9	73.1
Ours	**80.7**	**54.4**	**67.3**	**82.0**	**54.6**	**69.5**	**84.8**	**60.2**	**74.9**

Table 3. **3D detection results on KITTI val set with different labeled ratios.** The results are for moderate difficulty level evaluated by the mAP with 40 recall positions, with a rotated IoU threshold 0.7, 0.5, 0.5 for the three classes, respectively.

	Car			Pedestrian			Cyclist		
	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
PVR.	89.6	77.3	74.1	54.9	47.8	42.3	80.4	62.9	58.7
naive psd.-lb.	91.1	78.8	76.1	58.9	51.3	45.4	82.6	65.5	60.8
cls. thres. only	90.8	79.7	76.8	63.2	**55.0**	**49.7**	84.9	65.0	61.2
Ours	**91.6**	**80.7**	**78.1**	**63.3**	54.4	49.5	**86.5**	**67.3**	**62.8**

Table 4. **Ablation study on KITTI 1% labeled data.** Same evaluation metric as Table 1.

5. Conclusion

In this paper, we propose 3DIoUMatch, a novel semi-supervised 3D object detection method leveraging IoU estimation. Built upon a teacher-student mutual learning framework, we leverage asymmetric data augmentation and pseudo-label filtering and deduplication to facilitate the student learning from the EMA teacher. With our IoU estimation module, we make filtering and deduplication aware of localization confidence and apply test-time IoU-guided NMS and IoU optimization, leading to further improvement. Experiment results on the ScanNet, SUN-RGBD, and KITTI datasets validate the effectiveness of our method: we achieve significant gain over the previous state-of-the-art methods and baselines under all settings. Our idea of leveraging IoU estimation for semi-supervised learning is generally applicable to different kinds of 3D object detectors and can be extended to 2D detectors as future works.

Acknowledgement: This research is supported by a grant from the SAIL-Toyota Center for AI Research, NSF grant CHS-1528025, a Vannevar Bush Faculty fellowship, a TUM/IAS Hans Fischer Senior Fellowship, and gifts from the Adobe, Amazon AWS, and Snap corporations.

References

[1] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019.

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5049–5059, 2019.

[3] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

[4] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017.

[5] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[6] Jiyang Gao, Jiang Wang, Shengyang Dai, Li-Jia Li, and Ram Nevatia. Note-rcnn: Noise tolerant ensemble rcnn for semi-supervised object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 9508–9517, 2019.

[7] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[8] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[10] Judy Hoffman, Sergio Guadarrama, Eric S Tzeng, Ronghang Hu, Jeff Donahue, Ross Girshick, Trevor Darrell, and Kate Saenko. Lsda: Large scale detection through adaptation. In *Advances in Neural Information Processing Systems*, pages 3536–3544, 2014.

[11] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. In *Advances in neural information processing systems*, pages 10759–10768, 2019.

[12] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 784–799, 2018.

[13] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

[14] 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.

[15] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 2013.

[16] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[17] Charles R. Qi, Xinlei Chen, Or Litany, and Leonidas J. Guibas. Imvotenet: Boosting 3d object detection in point clouds with image votes, 2020.

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

[19] 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.

[20] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, pages 5099–5108, 2017.

[21] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[23] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in neural information processing systems*, pages 1163–1171, 2016.

[24] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10529–10538, 2020.

[25] 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.

[26] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

[27] 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 preprint arXiv:2005.04757*, 2020.

[28] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 2015.

[29] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017.

[30] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 9627–9636, 2019.

[31] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019.

[32] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Std: Sparse-to-dense 3d object detector for point cloud. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1951–1960, 2019.

[33] Li Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas J Guibas. Gspn: Generative shape proposal network for 3d instance segmentation in point cloud. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3947–3956, 2019.

[34] 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.

[35] 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.