Hierarchical Supervision and Shuffle Data Augmentation for 3D Semi-Supervised Object Detection

Chuandong Liu[1,2], Chenqiang Gao[1,2*], Fangcen Liu[1,2], Pengcheng Li[1,2], Deyu Meng[3,4], Xinbo Gao[1]
[1]School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China
[2]Chongqing Key Laboratory of Signal and Information Processing, Chongqing, China
[3]Xi'an Jiaotong University, Xi'an, China
[4]Macau University of Science and Technology, Taipa, Macau

Abstract

State-of-the-art 3D object detectors are usually trained on large-scale datasets with high-quality 3D annotations. However, such 3D annotations are often expensive and time-consuming, which may not be practical for real applications. A natural remedy is to adopt semi-supervised learning (SSL) by leveraging a limited amount of labeled samples and abundant unlabeled samples. Current pseudo-labeling-based SSL object detection methods mainly adopt a teacher-student framework, with a single fixed threshold strategy to generate supervision signals, which inevitably brings confused supervision when guiding the student network training. Besides, the data augmentation of the point cloud in the typical teacher-student framework is too weak, and only contains basic down sampling and flip-and-shift (i.e., rotate and scaling), which hinders the effective learning of feature information. Hence, we address these issues by introducing a novel approach of Hierarchical Supervision and Shuffle Data Augmentation (HSSDA), which is a simple yet effective teacher-student framework. The teacher network generates more reasonable supervision for the student network by designing a dynamic dual-threshold strategy. Besides, the shuffle data augmentation strategy is designed to strengthen the feature representation ability of the student network. Extensive experiments show that HSSDA consistently outperforms the recent state-of-the-art methods on different datasets. The code will be released at https://github.com/azhuantou/HSSDA.

1. Introduction

Kinds of important applications, especially autonomous driving, have been motivating the rapid development of 3D object detection by the range sensor data (*e.g.*, LiDAR point cloud). Up to now, many point-based and point-voxel-based methods [10,30,31,49] have been proposed. Despite the impressive progress, a large amount of accurate instance-level 3D annotations have to be provided for training, which is more time-consuming and expensive than 2D object annotation. This hinders the application and deployment of exist-

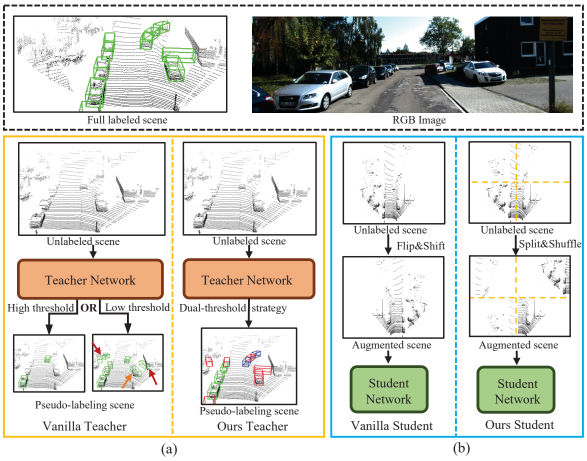

Figure 1. Illustration of (a) the previous teacher compared to our teacher framework and (b) the previous student compared to our student framework. The black dashed box includes the RGB image and the corresponding fully annotated 3D point cloud (green box). The left side of the yellow dotted line in (a) represents the pseudo-labeling scene generated by the single threshold of the vanilla teacher network, causing the student network may be severely misled due to missing mined objects (high threshold) or false positive objects (low threshold), while our proposed teacher network generates three groups of pseudo labels(shown as green, red, blue) to provide hierarchical supervision for the student network. (b) shows our student network adopts stronger shuffled data augmentation than the vanilla student network to learn the stronger ability of feature representation.

[*]Corresponding author.

ing advanced detection models. To this end, how to reduce the dependence on huge annotated datasets has achieved growing interest in the object detection community.

As one of the important machine learning schemes for reducing data annotation, semi-supervised learning (SSL) aims to improve the generalization ability of model training with a small number of labeled data together with large-scale unlabeled samples. In the semi-supervised object detection community, most works focus on 2D object detection [13, 17, 34, 40, 43], among which the teacher-student model is the mainstream framework. Specifically, the teacher network with weakly augmented labeled data generates pseudo labels to train the student network with strong data augmentation. This pipeline has been widely verified to be effective, in which the quality of pseudo labels and the data augmentation strategy are the key and many works have been proposed to tackle them [2,6,33,34]. Benefiting from 2D semi-supervised object detection, several 3D semi-supervised object detection methods have been proposed [23, 39, 47, 52], which still mainly adopted the teacher-student model. These methods, as well as 2D semi-supervised object detection methods [21,34,55], mainly use a hard way, *e.g.*, a score threshold, to get pseudo labels to train the student network. This kind of strategy is difficult to guarantee the quality of pseudo labels. Taking the score threshold strategy as an example, if the threshold is too low, the pseudo labels will contain many false objects, while if it is too high, the pseudo labels will miss many real objects which will be improperly used as background (see in Fig. 1 (a)). As exhibited in [39], only about 30% of the objects can be mined from the unlabeled scenes even at the end of the network training. Thus, both of those two cases will bring the student network confused supervision, which harms the performance of the teacher-student model. This would inevitably happen for the single threshold strategy, even adopting some optimal threshold search method [39]. Thus, how to produce reasonable pseudo labels from the teacher network output is an important issue to address for better training the student networks.

Besides the quality of pseudo labels, data augmentation is also the key to the teacher-student model as mentioned previously. Extensive works in 2D semi-supervised object detection have shown that strong data augmentation is very important to learn the strong feature representation ability of the student network. Thus, kinds of strong data augmentation strategies, *e.g.*, Mixup [48], Cutout [6], and Mosaic [4] have been widely adopted. However, current 3D semi-supervised object detection methods adopt some weak data augmentation strategies, *e.g.*, flip-and-shift. These kinds of data augmentations are not able to well drive the student network to learn strong feature representation ability. Thus, the good effect of data augmentation in 2D semi-supervised object detection does not appear obviously in 3D semi-supervised object detection.

To tackle the above issues of the quality of pseudo labels and data augmentation, we propose a Hierarchical Supervision and Shuffle Data Augmentation (HSSDA) method for 3D semi-supervised object detection. We still adopt the teacher-student model as our mainframe. For obtaining more reasonable pseudo labels for the student network, we design a dynamic dual-threshold strategy to divide the output of the teacher network into three groups: (1) high-confidence level pseudo labels, (2) ambiguous level pseudo labels, and (3) low-confidence level pseudo labels, as shown in Fig. 1 (a). This division provides hierarchical supervision signals for the student network. Specifically, the first group is used as the strong labels to learn the student network, while the second join learning through a soft-weight way. The higher the score is, the more it affects learning. The third group is much more likely to tend to be false objects. We directly delete them from the point cloud to avoid confusing parts of the object point cloud into the background.

For strengthening the feature representation ability of the student network, we design a shuffle data augmentation strategy in this paper. As shown in Fig. 1 (b), we first generate shuffled scenes by splitting and shuffling the point cloud patches in BEV (bird-eye view) and use them as inputs to the student model. Next, the feature maps extracted from the detector backbone are unshuffled back to the original point cloud geometry location.

To summarize, our contributions are as follows:

- We propose a novel hierarchical supervision generation and learning strategy for the teacher-student model. This strategy can provide the student network hierarchical supervision signal, which can fully utilize the output of the teacher network.
- We propose a shuffle data augmentation strategy that can strengthen the feature representation ability of the student network.
- Our proposed hierarchical supervision strategy and shuffle data augmentation strategy can be directly applied to the off-the-shelf 3D semi-supervised point cloud object detector and extensive experiments demonstrate that our method has achieved state-of-the-art results.

2. Related Work

2.1. 3D Object Detection

3D object detection is a fundamental task in the autonomous driving area. The mainstream 3D object detection methods can be roughly divided into three types: voxel-based methods [8, 14, 20, 44, 53, 54], point-based methods [24, 31, 32, 45, 46, 51], and point-voxel-based methods [9,19,28–30]. For voxel-based methods, voxelization is a common operation that transforms irregular point clouds into voxel grids for applying traditional 2D or 3D convolution. In Voxelnet [56], the voxel-wise encoding layer was adopted for collective feature representation extraction from the voxel-wise LiDAR point cloud. VoxSeT [8] presented a novel transformer-based framework that encoded features from a larger receptive field. Point-based approaches directly use the raw point cloud to capture spatial structure information for feature extraction through networks of the PointNet series [25, 26]. PointRCNN [31] is the representative, which directly generates point-level RoIs and uses the point-level features for further refinement. To accelerate the inference speed for applications, IA-SSD [50] proposed an efficient downsampling way and a contextual centroid perception module to capture geometrical structure. Point-Voxel-based methods combined voxel representations with point representations from the point cloud. Built on the PV-RCNN [29], PV-RCNN++ [30] leveraged a Vector-Pool aggregation to learn structure features and a sectorized proposal-centric keypoints sampling strategy to obtain more keypoints. All the above fully supervised methods can be easily embedded into our HSSDA framework, *e.g.*, PV-RCNN [29] and Voxel-RCNN [5].

2.2. Semi-supervised Learning (SSL)

SSL can greatly reduce the annotations for model training and most existing works focus on image classification, which can be broadly divided into two types: consistency regularization [2, 3, 22, 41] and pseudo-labeling methods [1, 11, 42]. The former approaches assume the model's predictions to be consistent under input perturbations/augmentations (*e.g.*, different contrast, flip, etc.) and penalize the inconsistency of predictions. Techniques range from simple augmentation to more complex transformations such as MixUp [48], as well as stronger automatic augmentation such as Cutout [6] and CTAugment [2]. The latter methods exploit pseudo-labeling, where the model first is trained trains on labeled data and then iteratively generates the pseudo labels of unlabeled data to add highly confident predictions for training. It is revisited in deep neural networks to learn from large amounts of unlabeled data. Notably, perturbation mechanisms of the above two types of methods play a key role in promoting the model robustness against noise in network parameters or structure but have not been explored in SSL for 3D object detection.

2.3. Semi-supervised Object Detection

Inspired by the SSL works in image classification, SSL is also applied to the 2D object detection to alleviate the heavy annotation problem [13, 17, 34, 40]. STAC [34] generated pseudo labels for unlabeled data in an offline manner. To further improve the quality of pseudo labels, Instant-Teaching [55] rectified the false predictions via the co-rectify scheme and experimented with MixUp [48] and Mosaic [4]. This work aims to tackle a more challenging task, SSL for 3D object detection, where large spaces of 7 Degrees-of-Freedom of 3D objects need to be searched.

Recently, several works have been proposed in the 3D SSL domain. SESS [52] and 3DIoUMatch [39] are the pioneer approaches for 3D object detection from indoor and outdoor point cloud data. Similar to 2D SSL methods, SESS [52] leveraged a triple consistency regularization strategy to align the 3D proposal from the teacher and student network. Following the pseudo-labeling line in SSL, 3DIoUMatch [39] designed a series of filtering strategies such as objectness, classification, and localization threshold to obtain high-quality pseudo labels, and a unique IoU estimation branch to further deduplicate the predictions. Different from 3DIoUMatch, Proficient Teachers [47] developed several necessary modules to improve the recall and precision of pseudo labels and removed the necessity of threshold setting. DetMatch [23] generated more precise pseudo labels by matching 2D and 3D detection from each modality. These works employed the EMA weight update strategy to train a student network and then gradually update the teacher network. Although achieving impressive performances with high-quality pseudo labels, the missing-mined objects and heuristic strong augmentation are ignored. By contrast, our HSSDA leverages the hierarchical supervision and shuffle data augmentation to alleviate these issues and further improve performance.

3. Method

3.1. Preliminary

Problem Definition We first provide the definition of semi-supervised 3D object detection. In detail, the model is trained with a set of labeled scenes $D_\mathrm{s} = \{p_i^\mathrm{s}, y_i^\mathrm{s}\}_{i=1}^{N_\mathrm{s}}$ and a set of unlabeled scenes $D_\mathrm{u} = \{p_i^\mathrm{u}\}_{i=1}^{N_\mathrm{u}}$, where $p_i \in \mathcal{R}^{n\times\{3+r\}}$ represents a point cloud scene p_i which has n points with three-dimensional coordinates and additional r-dimensional information (*e.g.*, color, intensity) that can be treated as extra features, N_s and N_u are the numbers of labeled and unlabeled point cloud scenes, respectively. Generally speaking, $N_u >> N_s$. For a scene p_i^s, the annotation y_i^s is composed of both 7-dimensional location information which includes center, size, and orientation, and category of the 3D bounding boxes.

Teacher-Student Framework Similar to the mainstream researches [15, 39, 52], our learning paradigm also builds up on the teacher-student framework which includes two 3D detectors with the same configurations. Here, we can use any off-the-shelf state-of-the-art 3D object detector, *e.g.*, PV-RCNN [29] and Voxel-RCNN [5]. Following those works, we build the teacher detector via exponential

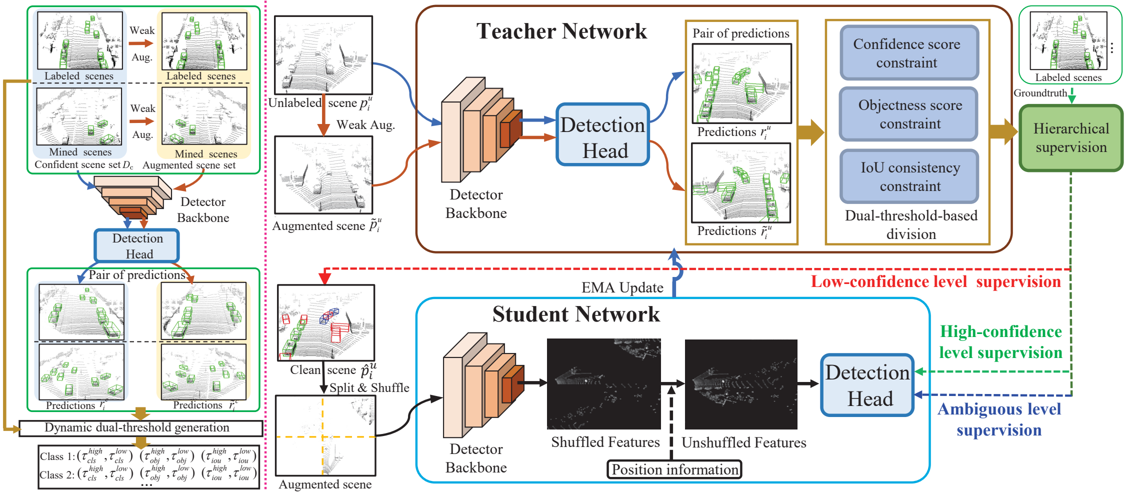
Labeled scenes
Weak Aug.
Labeled scenes
Mined scenes
Weak Aug.
Mined scenes
Confident scene set D_c
Augmented scene set
Detector Backbone
Detection Head
Pair of predictions
Predictions r_i^c
Predictions $\tilde{r}_i^c$
Dynamic dual-threshold generation
Class 1: $(\tau_{cls}^{high}, \tau_{cls}^{low})$ $(\tau_{obj}^{high}, \tau_{obj}^{low})$ $(\tau_{iou}^{high}, \tau_{iou}^{low})$
Class 2: $(\tau_{cls}^{high}, \tau_{cls}^{low})$ $(\tau_{obj}^{high}, \tau_{obj}^{low})$ $(\tau_{iou}^{high}, \tau_{iou}^{low})$
...
Unlabeled scene p_i^u
Weak Aug.
Augmented scene $\tilde{p}_i^u$
Teacher Network
Detector Backbone
Detection Head
Pair of predictions
Predictions r_i^u
Predictions $\tilde{r}_i^u$
Confidence score constraint
Objectness score constraint
IoU consistency constraint
Dual-threshold-based division
Labeled scenes
Groundtruth
Hierarchical supervision
EMA Update
Low-confidence level supervision
Clean scene $\hat{p}_i^u$
Split & Shuffle
Augmented scene
Student Network
Detector Backbone
Shuffled Features
Unshuffled Features
Position information
Detection Head
High-confidence level supervision
Ambiguous level supervision

Figure 2. Overview of the proposed HSSDA pipeline. We propose a dual-threshold strategy to help the teacher network to generate hierarchical supervision to train the student network. Besides, we also propose a data augmentation strategy to strengthen the ability of feature representation of the student network.

moving average (EMA) [37]:

$$\theta_t^{i+1} = \theta_t^i \cdot \alpha + \theta_s^i \cdot (1 - \alpha), \tag{1}$$

where α is the EMA decay rate, θ_t and θ_s represent the parameters of the teacher and student networks, respectively, and i denotes the training step.

3.2. Overview

The pipeline of our HSSDA framework is illustrated in Fig. 2, which is derived from the basic teacher-student mutual learning framework. In the burn-in stage of training, we train the detector in a fully supervised manner following OpenPCDet [38] with the labeled scenes and keep the same setting as the used detector. Then, both the teacher network and student network are initialized with the same pre-trained weight parameters.

In the mutual learning stage, there are three steps in each training epoch. The first step is to generate three kinds of dual-thresholds for each class in a global view, as shown on the left of the pink dotted line in Fig. 2. Specifically, we construct a confident scene set D_c composed of labeled scenes and mined scenes (in the first epoch, it just contains all labeled scenes). Then we sequentially input each scene from D_c and its weak augmentation (rotation and scaling) into the teacher network to produce a pair of predictions. Based on those pairs of predictions and the object information from D_c, we design a dynamic dual-threshold generation strategy to obtain three kinds of dual-thresholds for each class in terms of confidence score, objectness score, and IoU consistency: $(\tau_{cls}^{high}, \tau_{cls}^{low})$, $(\tau_{obj}^{high}, \tau_{obj}^{low})$, $(\tau_{iou}^{high}, \tau_{iou}^{low})$. The second step (see the right of the pink dotted line in Fig. 2) is mainly to mine the hierarchical pseudo labels. Specifically, each unlabeled scene p_i^u and its weak augmented scene $\tilde{p}_i^u$ are sequentially input to the teacher detector to produce a pair of predictions. Through three measure rules based on the dual-thresholds obtained in the first step, we can generate the hierarchical supervision: (1) high-confidence level pseudo labels, (2) ambiguous level pseudo labels, and (3) low-confidence level pseudo labels. We add all high-confidence pseudo labels into the confident scene set D_c for the next dual-threshold generation. We also add the ground-truth from labeled scenes into the first group for following student network training. In the third step, we use the hierarchical supervision composed of three groups of pseudo labels to train the student network with our designed shuffle data augmentation, and then update the teacher network by the EMA strategy according to Eq. (1).

After the mutual training step, we use the 3D detector from the student network as our final detector. Through the above procedure, we can see that our designed framework can train any off-the-shelf 3D detector which consists of a backbone and a detection head.

3.3. Dynamic dual-threshold generation

The dual threshold generation is dynamically conducted in each training epoch. Algorithm 1 describes the whole process of generating dual-thresholds for one class in one training epoch, given a confidence set $D_c = \{p_i^c, y_i^c\}_{i=1}^{N_c}$, its pairs of predictions r^c, $\tilde{r}^c$ through the 3D detector, and an IoU matching threshold τ_{pair}. Here τ_{pair} is determined experientially and is used to judge if two 3D bounding boxes match. Initially, we create three empty sets $\mathcal{P}_{cls}$, $\mathcal{P}_{obj}$ and $\mathcal{P}_{iou}$ to collect the confidence score, objectness score and IoU. These three sets will be used to search three optimal dual-thresholds. Concretely, for the i-th scene, we can fetch the predicted bounding boxes b_i^r and ground truth bounding

Algorithm 1: Dynamic dual-threshold generation

```
Input: confident scene set $D_c$, pairs of predictions
       $r^c$ and $\tilde{r}^c$, IoU matching threshold $\tau_{pair}$.
Output: $(\tau_{cls}^{high}, \tau_{cls}^{low})$, $(\tau_{obj}^{high}, \tau_{obj}^{low})$, $(\tau_{iou}^{high}, \tau_{iou}^{low})$
initialize empty sets $\mathcal{P}_{cls}$, $\mathcal{P}_{obj}$, and $\mathcal{P}_{iou}$;
for each $\{p_i^c, y_i^c\} \in D_c$ do
    fetch bounding boxes $b_i^{gt}$ from $y_i^c$;
    fetch bounding boxes $b_i^r$ from $r_i^c$;
    fetch bounding boxes $\tilde{b}_i^r$ from $\tilde{r}_i^c$;
    for $b_{ij}^{gt}$ in $b_i^{gt}$ do
        compute the matrix $\mathcal{M} \leftarrow IoU(b_{ij}^{gt}, b_i^r)$;
        if $max(\mathcal{M}) > \tau_{pair}$ then
            choose index $k$ with $max(\mathcal{M})$;
            $\mathcal{P}_{cls} = \mathcal{P}_{cls} \cup s_{cls}^c$,
            $s_{cls}^c$ is the confidence score of $r_{ik}^c$;
            $\mathcal{P}_{obj} = \mathcal{P}_{obj} \cup s_{obj}^c$,
            $s_{obj}^c$ is the objectness score of $r_{ik}^c$;
            $\mathcal{P}_{iou} = \mathcal{P}_{iou} \cup v^c$,
            $v^c = max(IoU(b_{ik}^r, \tilde{b}_i^r))$;
$(\tau_{cls}^{high}, \tau_{cls}^{low}), (\tau_{obj}^{high}, \tau_{obj}^{low}), (\tau_{iou}^{high} \tau_{iou}^{low}) \leftarrow$
    $JNB(\mathcal{P}_{cls}, \mathcal{P}_{obj}, \mathcal{P}_{iou})$;
```

boxes b_i^{gt} from r_i^c and y_i^c, respectively. Further, we utilize the common IoU-based strategy to pair a predicted box b_{ik}^r for each ground truth b_{ij}^{gt} in the i-th scene, aiming to obtain the predicted confidence scores and predicted objectness scores of ground truth objects, and collect these scores into the confidence score set $\mathcal{P}_{cls}$ and objectness score set $\mathcal{P}_{obj}$, respectively. In this way, we can distinguish predictions with different classifications and objectness reliability in a global view. At the same time, we can also get the consistency IoU set $\mathcal{P}_{iou}$ based on the r_i^c and $\tilde{r}_i^c$, which facilitates grading predictions with different localization quality in a consistency constraints manner. After handling all scenes, we solve the dual-threshold search problem through a global clustering algorithm. Specifically, we adopt the Jenks Natural Breaks (JNB) [12] algorithm to search the natural turning points or breakpoints based on the three sets. As shown in Fig. 3 (a) and (b), we can automatically obtain the dual-threshold τ_{cls}^{high} and τ_{cls}^{low} for each class. Similarly, we can automatically obtain the other two dual-thresholds τ_{obj} and τ_{iou}, which are detector-agnostic and category-aware. As mentioned in Sec. 3.2, in each training epoch, the confident scene D_c will be updated, so all three dual-thresholds will dynamically change during training.

3.4. Hierarchical supervision generation

As shown in Fig. 2, with the three dual-thresholds for each class, we divide the mined object pseudo labels from unlabeled scenes into hierarchical supervision which includes (1) high-confidence level pseudo labels, (2) ambiguous level pseudo labels and (3) low-confidence level pseudo labels. Specifically, the high-confidence level pseudo label $\tilde{y}_i$ is chosen when the predicted result r_i^u of unlabeled scenes from the teacher network meets all three following inequalities simultaneously, *i.e.*, $s_{cls}^u > \tau_{cls}^{high}$, $s_{obj}^u > \tau_{obj}^{high}$ and $v^u > \tau_{iou}^{high}$, where s_{cls}^u and s_{obj}^u are the confidence score and the objectness score of the chosen predicted result, respectively, and v^u is the consistency IoU between chosen r_i^u and $\tilde{r}_i^u$. Besides, we group the rest of the predictions as ambiguous level pseudo labels, which meet all three following inequalities simultaneously, *i.e.*, $s_{cls}^u > \tau_{cls}^{low}$, $s_{obj}^u > \tau_{obj}^{low}$ and $v^u > \tau_{iou}^{low}$. As for those predictions that neither belong to the high-confidence level nor ambiguous level, we mark them as low-confidence level predictions.

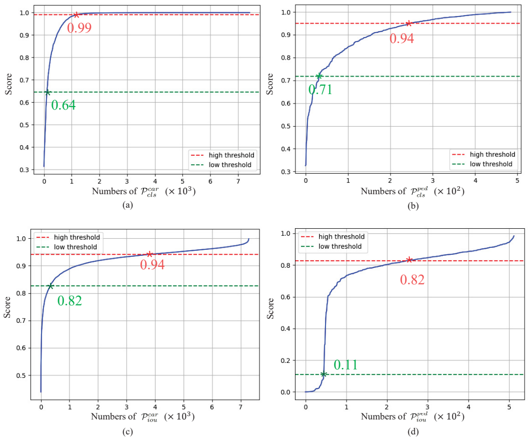

Figure 3. Selection of the dual-threshold based on the Jenks Natural Breaks [12]. In each subplot, the blue line represents the sorted scores from the sequence pool $\mathcal{P}$ and the red and green dash lines indicate the high threshold and low threshold, respectively. (a) and (b) show the process of dynamic confidence IoU selection for 'Car' and 'Pedestrian'. Similarly, (c) and (d) show the selection of dynamic consistency IoU threshold.

As mentioned previously, we also add the ground-truth from labeled scenes into the group of high-confidence level pseudo labels which provide strong object label supervision for the student network, while the ambiguous level group will supervise the student network training through a soft-weight way which will be introduced in Sec. 3.6. Following [16], we leverage the points removal strategy to eliminate noise information based on the group of low-confidence level pseudo labels.

3.5. Shuffle Data Augmentation

Weak-strong data augmentation plays a significant role in the teacher-student style framework, which guides the model to learn strong feature representation. As mentioned previously, the data augmentation of the student network of current methods is too weak to learn the strong ability of feature representation. However, due to the huge modality difference between 2D images and 3D point cloud, it is not feasible to directly apply the widely-used data augmentation strategies, *e.g.*, color transformation or Mixup [48] to point cloud for object detection. Thus, we propose a shuffle data augmentation strategy for the student network. Specifically, as shown in Fig. 2, given a scene $\hat{p}_i^u$ from the teacher network, we first clip the range of point cloud scene into $[x_1, x_2]$ for the X axis, $[y_1, y_2]$ for the Y axis and compress the Z axis to form BEV (bird-eye-view) grid. Then we split each scene into $\mathcal{R} \times \mathcal{C}$ (*e.g.*, 2×2) patches and generate the same shaped $\mathcal{R} \times \mathcal{C}$ position information to shuffle each scene patch. By feeding the shuffled patches to the detector backbone for feature extraction, the teacher network with a weak branch and the student network with a strong branch can achieve obvious differences, which is beneficial for the student network to learn more complex and peculiar information from hierarchical supervision. Then we leverage the position information to unshuffle the patch features for further regression and classification in the detection head. Although shuffled patches make it difficult to distinguish the edges or parts of objects in scenes, the unshuffle operation before the detection head restores the original location from the feature space. Hence, our student network delivers more effort into learning with weaker features, which would strengthen the ability of feature representation, and especially benefit to detect objects with weak feature due to small sizes, *e.g.*, 'Pedestrian', 'Cyclist', which will be shown in experiments.

3.6. Training Objective Function

Following [17, 36], we freeze the optimization of the teacher detector, and the student detector is trained on both unlabeled scenes with the hierarchical supervision and labeled scenes with the ground-truth. More specifically, the training objective consists of a supervised loss for labeled and unlabeled scenes.

$$\mathcal{L}_s = \sum_i \mathcal{L}_{cls}\left(p_i^s, y_i^s\right) + \mathcal{L}_{reg}\left(p_i^s, y_i^s\right), \tag{2}$$

$$\begin{aligned} \mathcal{L}_u = & \sum_i \mathcal{L}_{cls}\left(\hat{p}_i^u, \tilde{y}_{ij}^u\right) + \mathcal{L}_{reg}\left(\hat{p}_i^u, \tilde{y}_{ij}^u\right) \\ & + w_{ij}\mathcal{L}_{cls}\left(\hat{p}_i^u, \hat{y}_{ij}^u\right) + w_{ij}\mathcal{L}_{reg}\left(\hat{p}_i^u, \hat{y}_{ij}^u\right), \end{aligned} \tag{3}$$

where $\mathcal{L}_{cls}$ is the classification loss, $\mathcal{L}_{reg}$ is the regression loss, $\hat{y}_{ij}^u$ and $\tilde{y}_{ij}^u$ are the j-th ambiguous level pseudo labels and high-confidence level pseudo labels generated by the teacher detector in the i-th scene, and $w_{ij} = r_{ij}^{cls} \cdot r_{ij}^{obj}$ denotes the soft-weight for ambiguous level pseudo labels $\hat{y}_{ij}^u$, which is determined by a combination of predicted confidence score and predicted objectness reliability score. Thanks to the clean scenes generated by the noise points removal operation and further obtaining complex scenes by GT sampling data augmentation [44], we do not force each training batch to contain a mixture of labeled scenes p_i^s and unlabeled scenes p_i^u with a certain ratio (*e.g.*, $1:1$ in [39]), but randomly sample each batch from the entire dataset for training. Hence, the training loss is defined as follows:

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_u. \tag{4}$$

Thus, we remove the hyper-parameter to trade-off between $\mathcal{L}_s$ and $\mathcal{L}_u$ as used in the common teacher-student framework [17, 34, 39].

4. Experiments

4.1. Datasets and Evaluation Metrics

KITTI Dataset. Following the state-of-the-art methods [23, 39], we evaluate our HSSDA on the KITTI 3D detection benchmark [7], and we use the divided *train* split of 3,712 samples and *val* split of 3,769 samples as a common practice [29]. Then we sample three different 1% and 2% labeled scenes over *train* split based on the released 3DIoUMatch [39] splits. The reported results are averaged over model training on three sampled splits and evaluated on the *val* split. In addition, the KITTI benchmark has three difficulty levels (easy, moderate, and hard) due to the occlusion and truncation levels of objects. For fair comparisons, we report the mAP with 40 recall positions, with a 3D IoU threshold of 0.7, 0.5, and 0.5 for the three classes: car, pedestrian, and cyclist, respectively.

Waymo Open Dataset. We also evaluate our HSSDA on the Waymo Open Dataset [35], which is one of the biggest autonomous driving datasets, containing 798 sequences (approximately 158k point cloud scenes) for training and 202 sequences (approximately 40k point cloud scenes) for validation, whilst the view of annotations is in full 360° field. We find that even only 1% of the labeled Waymo scenes contain approximately 3 times as many object annotations as the full KITTI *train* split. Thus, we sample 1% of the 798 training sequences (approximately 1.4k point cloud scenes) and report the standard mean average precision (mAP) as well as mAPH, which represent the heading angle factors. In addition, the prediction results are split into LEVEL_1 and LEVEL_2 for 3D objects including more than five LiDAR points and one LiDAR point, respectively.

4.2. Implementation Details

At the training stage, the student network of our HSSDA is end-to-end optimized with the ADAM optimizer and a cosine annealing learning rate [18]. As for the weak augmentation for the teacher network, we randomly flip each scene along X-axis and Y-axis with 0.5 probability, and then scale it with a uniformly sampled factor from $[0.91, 1.12]$. Finally, we rotate the point cloud around Z-axis with a random angle sampled from $\left[-\frac{\pi}{4}, \frac{\pi}{4}\right]$. For

Model	Modality	1%				2%				20%			
		Car	Ped.	Cyc.	Avg.	Car	Ped.	Cyc.	Avg.	Car	Ped.	Cyc.	Avg.
PV-RCNN [29]	LiDAR	73.5	28.7	28.4	43.5	76.6	40.8	45.5	54.3	77.9	47.1	58.9	61.3
3DIoUMatch [39] (PVR.-based)	LiDAR	76.0	31.7	36.4	48.0	78.7	48.2	56.2	61.0	-	-	-	-
DetMatch [23] (PVR.&FR. [27]-based)	LiDAR + RGB	77.5	**57.3**	42.3	59.0	78.2	54.1	64.7	65.6	78.7	57.6	69.6	68.7
Our HSSDA (PVR.-based)	LiDAR	**80.9**	51.9	**45.7**	**59.5**	**81.9**	**58.2**	**65.8**	**68.6**	**82.5**	**59.1**	**73.2**	**71.6**

Table 1. Experimental results on KITTI dataset compared with recent state-of-the-art methods. For fair comparison, the results are reported with 40 recall positions, under IoU thresholds 0.7, 0.5, and 0.5 for ‘Car’, ‘Pedestrian’, and ‘Cyclist’, respectively.

1% Data (~ 1.4k scenes)	Modality	Veh. (LEVEL 1)		Veh. (LEVEL 2)		Ped. (LEVEL 1)		Ped. (LEVEL 2)		Cyc. (LEVEL 1)		Cyc. (LEVEL 2)	
		mAP	mAPH	mAP	mAPH	mAP	mAPH	mAP	mAPH	mAP	mAPH	mAP	mAPH
PV-RCNN [29]	LiDAR	47.3	45.6	43.6	42.0	28.9	15.6	26.2	14.1	-	-	-	-
DetMatch [23] (PVR.& FR. [27]-based)	LiDAR+RGB	52.2	51.1	48.1	47.2	39.5	18.9	**35.8**	17.1	-	-	-	-
Improvement	-	+4.9	+5.5	+4.5	+5.2	+10.6	+3.3	+9.6	+3.0	-	-	-	-
PV-RCNN [29] (Reproduced by us)	LiDAR	48.5	46.2	45.5	43.3	30.1	15.7	27.3	15.9	4.5	3.0	4.3	2.9
Our HSSDA (PVR.-based)	LiDAR	**56.4**	**53.8**	**49.7**	**47.3**	**40.1**	**20.9**	33.5	**17.5**	**29.1**	**20.9**	**27.9**	**20.0**
Improvement	-	+7.9	+7.6	+4.2	+4.0	+10.0	+5.2	+6.2	+1.6	+24.6	+17.9	+23.6	+17.1

Table 2. Performance comparison on the Waymo Open Dataset with 202 validation sequences for the 3D detection.

Model	Data	3D Detection (Car)			3D Detection (Ped)			3D Detection (Cyc)		
		Easy	Mod	Hard	Easy	Mod	Hard	Easy	Mod	Hard
Voxel-RCNN [5]	1%	87.9	74.0	67.1	23.7	19.0	17.4	44.8	37.0	25.5
Ours (Voxel-RCNN-based)	1%	**92.5**	**81.7**	**77.5**	**50.7**	**43.9**	**42.4**	**65.2**	**48.3**	**42.5**
Voxel-RCNN [5]	2%	89.2	76.5	71.5	44.2	40.2	34.4	56.7	39.9	37.4
Ours (Voxel-RCNN-based)	2%	**91.6**	**82.0**	**77.9**	**64.9**	**58.3**	**50.9**	**88.0**	**65.7**	**60.9**

Table 3. Experimental results on KITTI dataset based on the Voxel-RCNN detector, where the metrics are the same as Tab. 1.

the KITTI Dataset, the X-axis and Y-axis are limited in $[0, 70.4]m$ and $[-40, 40]m$ in the shuffle data augmentation, and our HSSDA (PV-RCNN-based) is trained for 80 epochs with the batch size 50. For the Waymo Open Dataset, the point cloud scene is clipped into $[-75.2, 75.2]m$ for X and Y axes, and training with the batch size 30 for 10 epochs. We set the value of τ_{pair} to 0.5 in all experiments according to the evaluation metric in the public datasets.

4.3. Main Results

KITTI Dataset. We first evaluate our proposed model on the popular KITTI dataset. Tab. 1 lists the results of different methods. From this table, we can observe that our approach significantly outperforms the state-of-the-art methods. Specifically, our approach has a remarkable boost in the ‘Car’ class for all settings, which has improvements of 7.4, 5.3, and 4.6 points compared to the PV-RCNN baseline for 1%, 2%, and 20%, respectively. Even compared to the recently proposed DetMatch [23] which uses two modalities of LiDAR and RGB, our methods just with LiDAR still have better results for most of the settings. Besides, we replace the point-voxel-based PV-RCNN 3D detector with a representative voxel-based Voxel-RCNN [5] 3D detector. The similarly impressive experimental results in Tab. 3 demonstrate the effectiveness of our HSSDA.

Waymo Dataset. For the more challenging Waymo Dataset, our approach still has a significant improvement in performance compared to the state-of-the-art methods. As shown in Tab. 2, our approach surpasses DetMatch [23]. It is worth mentioning that the proposed method achieves 29.1 mAP for ‘Cyclist’, which far exceeds the baseline.

4.4. Ablation Study

In this section, we present a series of ablation studies to analyze the effect of our proposed strategies in HSSDA. All the experiments are conducted based on the Voxel-RCNN detector with the 2% KITTI split and evaluated on *val* split due to its fast training speed. Tab. 4 summarizes the ablation results on our shuffle data augmentation (SDA) and hierarchical supervision (HS) of the teacher network, which provides three levels of supervision: high-confidence level (H_LEV) pseudo labels, ambiguous level (A_LEV) pseudo labels, and low-confidence level (L_LEV) pseudo labels.

Effect of the hierarchical supervision. It can be found that only considering the high-confidence level pseudo labels will perform better than the baseline as shown in Exp.2 in Tab. 4, but the improvements are limited by the confused background supervision. The introduction of ambiguous level supervision information can lead to further performance improvements which can be seen in Exp.3. Furthermore, we can observe that from Exp.4 generating clean scenes via low-confidence supervision can significantly improve the detection accuracy, which indicates the effectiveness of the points removal operation. Besides, the collaboration of three different levels of supervisions can greatly improve performance, as shown in Exp.5 of Tab. 4. Those results mean that all hierarchical supervisions have contributions to final performance when they work together.

Effect of the shuffle data augmentation. Exp.5 in Tab. 4 shows the effect of our shuffle data augmentation strategy. The classes of ‘Pedestrian’ and ‘Cyclist’ have very weak original features due to their small sizes. Both of them usually are very hard to detect. However, our shuffle data augmentation strategy can significantly improve their performance. It can be also observed that a slight drop for ‘Car’ may be due to the shuffle data augmentation splitting the original objects, leading to blurred edges for locating.

Exp.	Hierarchical Supervision (HS)			SDA	3D Detection			mAP
	H_LEV	A_LEV	L_LEV		Car	Ped.	Cyc.	
1	-	-	-	-	76.5	40.2	39.9	52.2
2	✓	-	-	-	75.9	51.8	51.5	59.7
3	✓	✓	-	-	78.8	51.1	52.2	60.7
4	✓	✓	✓	-	**82.4**	56.0	61.3	66.5
5	✓	✓	✓	✓	82.0	**58.3**	**65.7**	**68.6**

Table 4. Ablation study of different components in HSSDA.

Method	$\mathcal{R}$	$\mathcal{C}$	3D Detection			mAP
			Car	Ped.	Cyc.	
HS	1	1	82.4	56.0	61.3	66.5
HS + SDA	1	2	**82.5**	57.1	64.3	67.9
	2	2	82.0	**58.3**	65.7	**68.6**
	2	4	81.2	56.8	65.5	67.8
	4	4	81.1	54.8	**65.8**	67.2

Table 5. Results of various combinations of $\mathcal{R}$ and $\mathcal{C}$.

Our shuffle data augmentation has two hyperparameters: $\mathcal{R}$ and $\mathcal{C}$, which decide the number of scene patches to shuffle. To evaluate the effect of two hyperparameters, we investigate the performance of the proposed HSSDA with different combinations of $\mathcal{R}$ and $\mathcal{C}$ in Tab. 5. We can observe that the model achieves the best result when $\mathcal{R} = \mathcal{C} = 2$ (*i.e.,* the scene grids are split into 4 patches and perform random shuffle). Hence, we set $\mathcal{R}$ and $\mathcal{C}$ to 2 in all our experiments.

4.5. Quality Analysis

In this section, we analyze the quality of the generated pseudo labels which play a key role during model training. First of all, if the 3D IoU between pseudo labels and ground-truth boxes of labeled scenes is bigger than 0.5 with the same class, we regard the pseudo-label as a correctly mined object. From Tab. 6, we can see the final precision of our high-confidence level pseudo labels for each class on the KITTI dataset is particularly accurate, which indicates the effectiveness of our dual-threshold strategy. Besides, we provide qualitative results of wrong high-confidence level pseudo labels in Fig. 4. For ease of viewing, we only show the object of one failure case in each scene. (a) and (b) in Fig. 4 show that the common failures for 'Car' usually occur with similar classes (such as vans and trucks). Interestingly, our method can reliably mine some real objects which were not annotated in the dataset (see Fig. 4 (c) and Fig. 4 (d)). Additionally, due to the small sizes of 'Pedestrian', most of the failure examples are caused by inaccurate localization, as shown in Fig. 4 (e) and (f) (the ground truth and pseudo-labeling 3D bounding box are drawn in red and cyan.)

5. Conclusion

In this paper, we propose a novel teacher-student-based method for 3D semi-supervised object detection, called HSSDA. Through the dual-threshold strategy in the teacher network, we can provide hierarchical supervision to effectively train the student network, while eliminating the negative impact of missing-mined objects of unlabeled scenes. In addition, the shuffle data augmentation strategy shuffles the input and unshuffles the feature blocks to strengthen the feature representation ability of the student network. Extensive experiments validate the superiority of our method in challenging datasets.Our HSSDA can train any 3D detector which consists of a backbone and detection head.

Category	Split setting on KITTI		
	1%	2%	20%
Car	96.73 (4627/4783)	98.69 (4476/4535)	98.88 (4508/4559)
Pedestrian	85.58 (204/239)	92.29 (273/296)	93.67 (74/79)
Cyclist	95.53 (107/112)	95.00 (114/120)	96.33 (105/109)

Table 6. Final precision of high-confidence level pseudo labels on 'Car', 'Pedestrian' and 'Cyclist' classes with different SSL settings. The blue and red numbers represent the total and correctly number of mined pseudo labels, respectively.

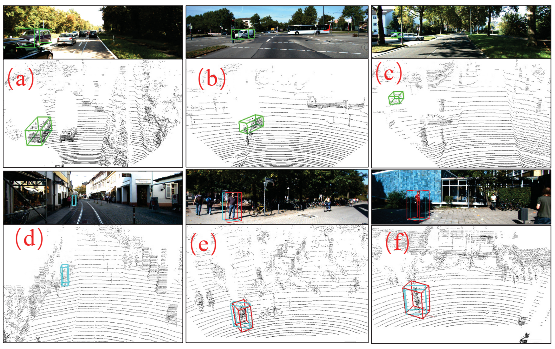

Figure 4. Qualitative analysis of pseudo labels on KITTI. For a better view, we only show the objects we are interested in and set the pseudo-labeling car, pseudo-labeling pedestrian, and ground truth bounding box in green, cyan, and red, respectively, whilst projecting boxes in point cloud back to RGB images. (a) and (b) show the case of category errors, (c) and (d) show the missing-annotated instance of the dataset, and (e) and (f) show the case of poorly localized pseudo labels. Best viewed in color.

Limitations. Our HSSDA designs a dynamic dual-threshold strategy that determines the optimal threshold in a global view. So, we need to use the teacher network for additional validation, which requires more training time. In addition, the shuffle data augmentation may split the complete objects resulting in blurred edges and locating.

Acknowledgments. This work is supported in part by the National Key R&D Program of China (2022YFA1004100), and in part by the National Natural Science Foundation of China (No.62176035, 62201111, 62036007), the Science and Technology Research Program of Chongqing Municipal Education Commission under Grant (No.KJZD-K202100606), the Chongqing graduate Research Innovation Project (CYB22249), the CQUPT Ph.D. Innovative Talents Project (BYJS202105), and the Macao Science and Technology Development Fund (061/2020/A2).

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *IJCNN*, pages 1–8. IEEE, 2020. 3

[2] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *ICLR*, 2019. 2, 3

[3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *NeurIPS*, 32, 2019. 3

[4] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 2, 3

[5] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li. Voxel r-cnn: Towards high performance voxel-based 3d object detection. In *AAAI*, pages 1201–1209, 2021. 3, 7

[6] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 2, 3

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

[8] Chenhang He, Ruihuang Li, Shuai Li, and Lei Zhang. Voxel set transformer: A set-to-set approach to 3d object detection from point clouds. In *CVPR*, pages 8417–8427, 2022. 2, 3

[9] Chenhang He, Hui Zeng, Jianqiang Huang, Xian-Sheng Hua, and Lei Zhang. Structure aware single-stage 3d object detection from point cloud. In *CVPR*, pages 11873–11882, 2020. 2

[10] Jordan SK Hu, Tianshu Kuai, and Steven L Waslander. Point density-aware voxels for lidar 3d object detection. In *CVPR*, pages 8469–8478, 2022. 1

[11] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *CVPR*, pages 5070–5079, 2019. 3

[12] George F Jenks. Optimal data classification for choropleth maps. *Department of Geographiy, University of Kansas Occasional Paper*, 1977. 5

[13] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. *NeurIPS*, 32, 2019. 2, 3

[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 *CVPR*, pages 12697–12705, 2019. 2

[15] Zhaoqi Leng, Shuyang Cheng, Benjamin Caine, Weiyue Wang, Xiao Zhang, Jonathon Shlens, Mingxing Tan, and Dragomir Anguelov. Pseudoaugment: Learning to use unlabeled data for data augmentation in point clouds. In *ECCV*, pages 555–572. Springer, 2022. 3

[16] Chuandong Liu, Chenqiang Gao, Fangcen Liu, Jiang Liu, Deyu Meng, and Xinbo Gao. Ss3d: Sparsely-supervised 3d object detection from point cloud. In *CVPR*, pages 8428–8437, 2022. 5

[17] Yen-Cheng Liu, Chih-Yao Ma, and Zsolt Kira. Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In *CVPR*, pages 9819–9828, 2022. 2, 3, 6

[18] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *ICLR*, 2017. 6

[19] Jiageng Mao, Minzhe Niu, Haoyue Bai, Xiaodan Liang, Hang Xu, and Chunjing Xu. Pyramid r-cnn: Towards better performance and adaptability for 3d object detection. In *ICCV*, pages 2723–2732, 2021. 2

[20] Jiageng Mao, Yujing Xue, Minzhe Niu, Haoyue Bai, Jiashi Feng, Xiaodan Liang, Hang Xu, and Chunjing Xu. Voxel transformer for 3d object detection. In *ICCV*, pages 3164–3173, 2021. 2

[21] Peng Mi, Jianghang Lin, Yiyi Zhou, Yunhang Shen, Gen Luo, Xiaoshuai Sun, Liujuan Cao, Rongrong Fu, Qiang Xu, and Rongrong Ji. Active teacher for semi-supervised object detection. In *CVPR*, pages 14482–14491, 2022. 2

[22] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018. 3

[23] Jinhyung Park, Chenfeng Xu, Yiyang Zhou, Masayoshi Tomizuka, and Wei Zhan. Detmatch: Two teachers are better than one for joint 2d and 3d semi-supervised object detection. In *ECCV*, pages 370–389. Springer, 2022. 2, 3, 6, 7

[24] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *ICCV*, pages 9277–9286, 2019. 2

[25] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, pages 652–660, 2017. 3

[26] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, pages 5099–5108, 2017. 3

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015. 7

[28] Hualian Sheng, Sijia Cai, Yuan Liu, Bing Deng, Jianqiang Huang, Xian-Sheng Hua, and Min-Jian Zhao. Improving 3d object detection with channel-wise transformer. In *ICCV*, pages 2743–2752, 2021. 2

[29] 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 *CVPR*, pages 10529–10538, 2020. 2, 3, 6, 7

[30] Shaoshuai Shi, Li Jiang, Jiajun Deng, Zhe Wang, Chaoxu Guo, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn++: Point-voxel feature set abstraction with local vector representation for 3d object detection. *International Journal of Computer Vision*, 131(2):531–551, 2023. 1, 2, 3

[31] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *CVPR*, pages 770–779, 2019. 1, 2, 3

[32] Weijing Shi and Raj Rajkumar. Point-gnn: Graph neural network for 3d object detection in a point cloud. In *CVPR*, pages 1711–1719, 2020. 2

[33] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *NeurIPS*, 33:596–608, 2020. 2

[34] 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. 2, 3, 6

[35] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, pages 2446–2454, 2020. 6

[36] Yihe Tang, Weifeng Chen, Yijun Luo, and Yuting Zhang. Humble teachers teach better students for semi-supervised object detection. In *CVPR*, pages 3132–3141, 2021. 6

[37] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, pages 1195–1204, 2017. 4

[38] OpenPCDet Development Team. Openpcdet: An open-source toolbox for 3d object detection from point clouds. https://github.com/open-mmlab/OpenPCDet, 2020. 4

[39] He Wang, Yezhen Cong, Or Litany, Yue Gao, and Leonidas J Guibas. 3dioumatch: Leveraging iou prediction for semi-supervised 3d object detection. In *CVPR*, pages 14615–14624, 2021. 2, 3, 6, 7

[40] Zhenyu Wang, Yali Li, Ye Guo, Lu Fang, and Shengjin Wang. Data-uncertainty guided multi-phase learning for semi-supervised object detection. In *CVPR*, pages 4568–4577, 2021. 2, 3

[41] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *NeurIPS*, 33:6256–6268, 2020. 3

[42] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *CVPR*, pages 10687–10698, 2020. 3

[43] 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 *ICCV*, pages 3060–3069, 2021. 2

[44] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 2, 6

[45] Zetong Yang, Yanan Sun, Shu Liu, and Jiaya Jia. 3dssd: Point-based 3d single stage object detector. In *CVPR*, pages 11040–11048, 2020. 2

[46] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Std: Sparse-to-dense 3d object detector for point cloud. In *ICCV*, pages 1951–1960, 2019. 2

[47] Junbo Yin, Jin Fang, Dingfu Zhou, Liangjun Zhang, Cheng-Zhong Xu, Jianbing Shen, and Wenguan Wang. Semi-supervised 3d object detection with proficient teachers. In *ECCV*, pages 727–743. Springer, 2022. 2, 3

[48] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 2, 3, 6

[49] Yifan Zhang, Qingyong Hu, Guoquan Xu, Yanxin Ma, Jianwei Wan, and Yulan Guo. Not all points are equal: Learning highly efficient point-based detectors for 3d lidar point clouds. In *CVPR*, pages 18953–18962, 2022. 1

[50] Yifan Zhang, Qingyong Hu, Guoquan Xu, Yanxin Ma, Jianwei Wan, and Yulan Guo. Not all points are equal: Learning highly efficient point-based detectors for 3d lidar point clouds. In *CVPR*, pages 18953–18962, 2022. 3

[51] Yanan Zhang, Di Huang, and Yunhong Wang. Pc-rgnn: Point cloud completion and graph neural network for 3d object detection. In *AAAI*, number 4, pages 3430–3437, 2021. 2

[52] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Sess: Self-ensembling semi-supervised 3d object detection. In *CVPR*, pages 11079–11087, 2020. 2, 3

[53] Wu Zheng, Weiliang Tang, Sijin Chen, Li Jiang, and Chi-Wing Fu. Cia-ssd: Confident iou-aware single-stage object detector from point cloud. In *AAAI*, number 4, pages 3555–3562, 2021. 2

[54] Wu Zheng, Weiliang Tang, Li Jiang, and Chi-Wing Fu. Se-ssd: Self-ensembling single-stage object detector from point cloud. In *CVPR*, pages 14494–14503, 2021. 2

[55] Qiang Zhou, Chaohui Yu, Zhibin Wang, Qi Qian, and Hao Li. Instant-teaching: An end-to-end semi-supervised object detection framework. In *CVPR*, pages 4081–4090, 2021. 2, 3

[56] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *CVPR*, pages 4490–4499, 2018. 3