SA3Det++: Side-Aware Quality Estimation for Semi-Supervised 3D Object Detection

Wenfei Yang, Chuxin Wang, Tianzhu Zhang, Yongdong Zhang, and Feng Wu

Abstract—Semi-supervised 3D object detection from point cloud aims to train a detector with a small number of labeled data and a large number of unlabeled data. Among existing methods, the pseudo-label based methods have achieved superior performance, and the core lies in how to select high-quality pseudo-labels with the designed quality evaluation criterion. Despite the success of these methods, they all consider the localization and classification quality estimation from a global perspective. For localization quality, they use a global score threshold to filter out low-quality pseudo-labels and assign equal importance to each side during training, ignoring the fact that sides with different localization quality should not be treat equally. Besides, a large number of pseudo-labels are discarded due to the high global threshold, which may also contain some correctly predicted sides that are helpful for model training. For the classification quality, they usually combine the objectness score and classification confidence score to filter out pseudo-labels. The main focus of them is designing effective classification confidence evaluation metrics, neglecting the importance of predicting better objectness score. In this paper, we propose SA3Det++, a side-aware quality estimation method for semi-supervised object detection, which consists of a probabilistic side localization strategy, a side-aware quality estimation strategy, and a soft pseudo-label selection strategy. Extensive results demonstrate that the proposed method consistently outperforms the baseline methods under different scenes and evaluation criterions. Code is avaliable at: https://github.com/OpenSpaceAI/Nesie.

Index Terms—Semi-supervised, 3D object detection, side-aware quality estimation.

1 Introduction

3D object detection in point clouds aims at estimating oriented 3D bounding boxes as well as category labels of objects. As a fundamental task in the computer vision area, it has significant application values in autonomous driving [1], [2], [3] and other areas. In the past few decades, many fully supervised 3D object detection methods have been proposed [4], [5], [6], [7]. However, these methods rely heavily on a large amount of fully annotated data with instance-level bounding boxes and category labels, which are expensive and time-consuming to collect. For example, it takes more than 100 seconds to annotate an object in 3D point cloud [8].

To reduce the high annotation cost associated with fully supervised methods, semi-supervised methods [9], [10], [11], [12] that use a combination of labeled data and a large amount of unlabeled data for model training have gained increasing attention. According to the technique design, existing semi-supervised 3D methods can be broadly divided into two categories, consistency-based methods [9] and pseudo-label based methods [10], [11], [12]. For consistency-based methods, the core idea is to apply different augmentations on unlabeled data and encourage the predictions of them to be consistent. For example, SESS [9] adopts the mean-teacher framework and designs three consistency losses between the student and teacher model to enforce the consensus of object locations, semantic categories and sizes. For pseudo-label based methods, they aim to select high-quality pseudo-labels for unlabeled data to train the model. Currently, the pseudo-label based methods have achieved better performance than consistency based methods, and the key lies in designing effective strategy to select high quality pseudo-labels for unlabeled data. Each pseudo label consists of a pseudo category label and a pseudo bounding box, thus the quality estimation can be divided into classification quality estimation and localization quality estimation. For example, HPCA [10] introduces an adaptive class confidence selection module to select pseudo-labels with confidence above the threshold, while 3DIoUMatch [11] equips the detector with a 3D Intersection over Union (IoU) estimation module to estimate the global localization quality and proposes lower-half suppression strategy for pseudo-label filtering.

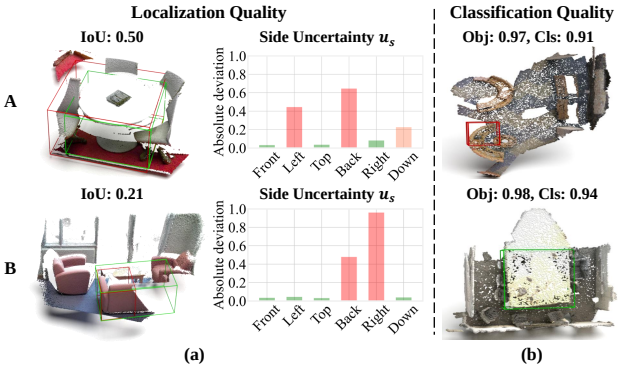

Fig. 1. **The motivation of this paper.** (a): Localization quality of pseudo-labels. Red color represents pseudo-labels and green color represents ground-truth. Pseudo-labels with high IoU may be incorrect on some sides, while pseudo-labels with low IoU may be correct on some sides. (b): Classification quality of pseudo-labels. Existing methods tend to predict high objectness score for pseudo-label that only covers a part of object, and this false confidence can not be suppressed by the proposal classification branch. All these examples are collected from the predictions of VoteNet on unlabeled data.

• *Wenfei Yang, Chuxin Wang, Tianzhu Zhang, Yongdong Zhang and Feng Wu are with school of information science, University of Science and Technology of China, Hefei 230027, China (e-mail: yangwf@ustc.edu.cn; wcx0602@mail.ustc.edu.cn; tzzhang@ustc.edu.cn; zhyd73@ustc.edu.cn; fengwu@ustc.edu.cn).*

Despite the success of previous pseudo-label based methods, they all consider the localization and classification quality estimation from a global perspective. **For the localization quality estimation**, they use global metrics (such as IoU) to measure the quality of pseudo bounding boxes. However, high global localization quality can not guarantee the quality of each side, while low global localization quality do not indicate all predicted sides are incorrect. As shown in Figure 1 (a), the IoU of pseudo-label A is high, but there are significant errors in the left and back sides. Using these incorrect sides to supervise the model training is harmful. On the contrary, although the IoU of pseudo-label B is relatively low, the predictions of four sides (front, left, top and down sides) are accurate, which are valuable for model training but are discarded due to the the low global quality. **For the classification quality estimation**, existing methods usually use the objectness score and classification confidence score (entropy, maximum value of the classification score, et al.) as quality metric, and the main focus of them lies in designing effective confidence metrics. However, as an important component of the classification quality, how to predict better objectness score has been neglected. As shown in Figure 1 (b), existing methods adopt the proposal feature to predict the objectness score, which may assign high objectness score to proposals that only covers the most discriminative part of an object. Although some part-based localization and classification techniques have advanced the performance of fully supervised 3D object detection [13], [14], [15], they are either designed to aggregate part-aware global features for box scoring and location refinement or to obtain part-sensitive classification scores, which can not be directly used for the side quality estimation and classification quality estimation.

Based on the above observations, we aim to explicitly decouple the pseudo-labels into independent sides and extract side-aware features for localization and classification quality estimation. The key idea is to model the location of each side as independent probabilistic distributions, and then extract side-aware geometric features and probabilistic properties to estimate the quality scores. To achieve this, three key issues need to be considered: 1) **How to estimate the location of each side?** Existing methods [10], [11], [12] predict the object center and size to decode the location of each side in a deterministic and coupled way, which makes them unable to reflect the localization quality of each side independently. To deal with this problem, a natural idea is to regard the location of each side as a gaussian distribution and predict the mean and variance of it, as commonly used in 2D area. However, the size of different object varies significantly, the variance cannot reflect the localization quality well since it is natural for objects with larger sizes to have larger variances. 2) **How to extract side-aware features for quality estimation?** Our target is to utilize side-aware features for quality estimation, involving the quality of each side, the quality of rotation angle and the objectness score, and different quality estimation target requires different features. For example, the quality estimation of each side should consider the side-aware local geometric feature, the estimation of the objectness score should consider the surrounding feature to determine how likely this is a positive object. 3) **How to select pseudo-labels for model training?** An effective pseudo-label selection strategy should suppress the interference of pseudo labels with low quality while retaining pseudo-labels with high quality to supervise model training. A high threshold can select pseudo-labels with high accuracy but degraded recall, and a low threshold can select pseudo-labels with high recall but degraded accuracy. How to obtain pseudo-labels with high recall and eliminate the impact of noisy pseudo-labels is very important.

Motivated by the above discussion, we propose a novel side-aware quality estimation method for semi-supervised object detection, which consists of a probabilistic side prediction strategy, a side-aware quality estimation strategy, and a side-aware soft pseudo-label selection strategy. **In the probabilistic side prediction strategy**, we decouple the localization task by dividing the object bounding box into independent sides and rotation angle. The position of each side is predicted as a probability distribution over an interval. The distribution can determine the location of each side and can be used in subsequent modules for localization quality estimation. **In the side-aware quality estimation strategy**, we first extract geometric-aware features for each side by meticulously analyzing its local structure. The properties of the predicted distribution are then combined with the side-aware geometric features, and the resulting concatenated features are processed by a Multi-Layer Perceptron (MLP) to predict the localization quality of each side. In the meantime, the side-aware geometric features of all sides are concatenated together and then fed into a MLP to predict the objectness score. During the training stage, a quality regression loss is designed to guide the quality estimation of each side and rotation angle. As shown in Figure 1, the proposed strategy can effectively estimate the quality of different sides and predict better objectness score. **In the side-aware soft pseudo-label selection strategy**, we first use a low threshold to filter out pseudo labels to keep the recall rate. Then we use the objeceness score as the weight of the pseudo-label classification loss, and use the side quality as the weight of different side in the bounding box regression loss. This strategy can suppress the interference of low-quality sides and fully utilize sides with higher quality in the pseudo-labels, as well as suppress pseudo-labels with low classification quality.

In summary, the contributions of this paper are as follows: 1) We introduce a novel side-aware quality estimation method for semi-supervised 3D object detection, which includes a probabilistic side prediction strategy, a side-aware quality estimation strategy, and a side-aware soft pseudo-label selection strategy. The proposed method allows for effective side-aware localization and classification quality estimation, enabling us to mine as much useful information as possible from the pseudo-labels for model training. 2) Extensive experimental results show that the proposed method can consistently improve the performance of different baseline detectors under different scenes and evaluation criteria, verifying the effectiveness of our method.

Our preliminary work is introduced in [16]. In this journal version, we have the following extensions. First, we modify the side localization strategy with an extra scale prediction module, which can help to better predict the location of sides. Second, we also add a side-aware rotation quality estimation branch, which can better estimate the localization quality for bounding boxes with rotations. Third, we design an extra side-aware objectness score estimation strategy, which can help to suppress pseudo-labels that may belong to background. Besides, we extend our method to two other baseline detectors and conduct more extensive experiments to show the advantage and limitations of our method.

2 Related Work

In this section, we briefly review semi-supervised object detection and uncertainty estimation methods.

2.1 Semi-supervised object detection

Semi-supervised 2D object detection To reduce the annotation burden of fully supervised object detection, many 2D object detection methods that utilize weakly labeled data for model training have been proposed [17], [18]. Among them, the semi-supervised methods that use a combination of labeled data and unlabeled data have received great attention. In [19] and [20], the ensembles of predictions on data with different augmentations are used to form the pseudo-labels for unlabeled images. Inspired by Fixmatch [21], STAC [22] proposes to combine pseudo-label based self-training and weak/strong data augmentations consistency regularization. A detector is trained on labeled data first and then used to generate pseudo-labels for unlabeled data, which are then used with labeled data to re-train the detector. To deal with the confirmation bias of self-training, some works [23], [24] turn to generate pseudo labels through the mean teacher framework [25], in which the teacher model is updated in an exponential moving average way. Recently, the methods that combine pseudo labels and mean teacher framework have achieved dominant performance in this area. ISTM [26] proposes to fuse the pseudo labels of teacher models from different iterations to generate more accurate pseudo labels. In SoftTeacher [27], they propose the first end-to-end pseudo-labeling framework by utilizing a teacher model to avoid the complicated multi-stage training process. In MixTeacher [28], they proposed to generate high-quality pseudo labels from a mixed scale feature pyramid. In ConsistentTeacher [29], they propose an adaptive anchor assignment (ASA) to replace the IoU-based pseudo label assignment strategy, which can mitigate the negative impact of noisy pseudo bounding boxes. Despite the success of these methods, they all use sparse pseudo labels after the non-maximum suppression process to supervise the training of student. Different from them, the Inverse NMS Clustering (INC) and Rank Matching (RM) module is proposed in DenseTeacher [30] to conduct dense-to-dense learning.

Semi-supervised 3D object detection. In the 3D area, SESS [9] is the first work for semi-supervised point-based 3D object detection, which consists of a student model and an EMA teacher model. It applies two different augmentations to the unlabeled data and then feeds them into the student model and teacher model. By constraining the outputs of these two models to be consistent, it can effectively learn from these unlabeled data. However, it is suboptimal to enforce all predictions of the student and teacher model to be consistent, because there are many false predictions. To mitigate this issue, 3DIoUMatch [11] introduces the confidence-based filtering strategy and IoU prediction strategy to select pseudo-labels with high quality from the predictions of the teacher model. In line with this work, the contemporary work HPCA [10] designs an adaptive class confidence selection scheme for pseudo-label filtering, which is inspired by the FlexMatch [31] strategy for semi-supervised image classification. Recently, DQS3D [32] introduces a voxel-based quantization-aware framework to achieve dense supervision instead of sparse proposal matching, which has achieved significant improvements on indoor datasets. Diffusion-SS3D [33] has achieved superior performance by using the denoised diffusion process to generate pseudo-labels with better quality. Besides from pseudo-label selection, some methods propose to explore better data augmentation strategy. For example, a shuffle data augmentation strategy is designed in HSSDA [34] to strengthen the feature representation ability of the student network. Similar to the 2D cut-paste augmentation, DPKE [35] proposes to crop bounding boxes from labeled scenes and then randomly paste them into other labeled and unlabeled scenes, along with a collision detection strategy. Instead of directly applying pseudo-labels, NoiseDet [36] and Reliable-Student [37] propose to treate the semi-supervised problem as learning from noisy pseudo-labels. Specifically, NoiseDet [36] proposes to soften the categorical label into a value ranging from 0 to 1 with the guidance of confidence score and IoU score. In Reliable-Student [37], the reliability weights are determined by querying the teacher network for confidence scores of the student-generated proposals. Different from existing methods that treat each bounding box as a whole and measure the pseudo-label quality from the global perspective, the proposed method aims to evaluate the localization quality of each side and treat them with different importance during the training stage.

2.2 Uncertainty estimation.

Uncertainty estimation aims to produce a measure of confidence for model predictions, which is usually used for pseudo-label quality estimation and plays an important role in many artificial intelligence systems, such as autonomous driving. Traditional deep neural networks are deterministic models, and the bayesian deep learning makes it possible to estimate the uncertainty of deep model predictions [38]. In [39], a bayesian deep learning framework is proposed to estimate the aleatoric uncertainty about data and epistemic uncertainty about model, respectively. Inspired by this work, many subsequent works have been proposed for various computer vision tasks, such as object detection [40], [41], medical image segmentation [42], person re-identification [43]. In [44] and [40], they introduce the idea of uncertainty estimation into the object detection area by modeling the bounding box coordinates as the gaussian parameters. Different from these works that treat the bounding box as gaussian distribution, Li et al. [45] propose to represent the bounding box locations as arbitrary distribution by learning a discrete probability vector over the continuous space. For 3D object detection, a probabilistic detector is proposed in [41] to quantify uncertainty for LiDAR point cloud vehicle detection. Based on this work, Meyer et al. [46] propose to improve the probability distribution learning ability by considering the potential noise in the ground-truth labeled data. In [47] and [48], geometry information is used to predict the uncertainty of the detection bounding box. Different from previous methods, we design a side-aware quality estimation method by incorporating the spatial distribution properties of the sides and the nearby geometric properties. To the best of our knowledge, no previous works have attempted to evaluate the localization quality of each side in the semi-supervised 3D object detection area.

3 METHOD

3.1 Overview

Given a 3D point cloud, the purpose of 3D object detection is to localize the position of each object and identify its semantic category label. Denote l and u as the indicators for labeled data and unlabeled data, semi-supervised 3D object detection aims to train a detector with a small labeled dataset $\{X^l, Y^l\}$ and a large unlabeled dataset $\{X^u\}$. For labeled data, the center point position, scale in each direction, orientation along the vertical axis and category label are given. For unlabeled data, the target is to generate pseudo-labels for them to train the model. In this

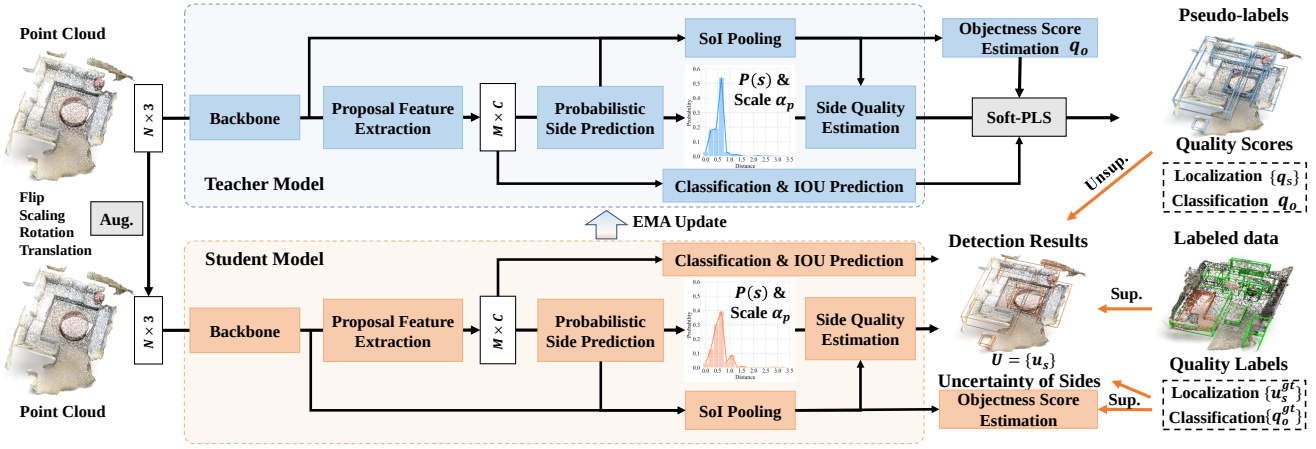

Fig. 2. **The overview of the proposed side-aware quality estimation method for semi-supervised object detection.** The proposed method follows the mean-teacher framework. We start by augmenting the input data and feeding it into the student model to obtain both the detection result and the deviation of the sides. For labeled data, we directly use the ground-truth to supervise the predicted results. For unlabeled data, we apply the proposed method to filter the predicted results of the teacher model to generate pseudo labels and quality scores, which are used to supervise the student model. To update the parameters of the teacher model, we employ the Exponential Moving Average (EMA) strategy.

paper, we propose a side-aware method for pseudo-label quality estimation, and the pipeline of the proposed method is illustrated in Figure 2. Given the input point cloud, we first use the 3D backbone model to extract point cloud features P_{seed}, which are then used to obtain the candidate proposal features through the proposal feature extraction module. Then, we feed the proposal feature into the probabilistic side localization module to obtain the probablistic distribution of each side. By integrating the spatial distribution of each side, we can then determine the position of each side. Although the distribution properties (e.g., variance) can partly reflect the localization quality of each side, they are based on the global proposal feature and neglect the side-aware local features. Consequently, we design a Side of Interest Pooling (SoI Pooling) module to extract the geometric features of each side and fuse them with distribution properties. The fused features are then input into the quality estimation head to estimate the quality of the sides. Besides, we also fuse the features of all sides for rotation angle and objectness score estimation. Lastly, we employ soft Pseudo-Label Selection (soft-PLS) to filter the predicted results of the teacher model, and then use them to supervise the student model.

3.2 Probabilistic side localization

In existing semi-supervised 3D object detection methods, the positions of sides are jointly determined by the center position and size predictions, which cannot reflect the localization quality of each side. In this paper, we decouple the bounding box into independent side and design a probabilistic side localization strategy to predict the side location. Specifically, given the candidate proposal features and locations of candidate points, we directly predict the distance from the candidate points to each side in a probabilistic way. We denote a bounding box as $B = \{t, d, l, r, f, b, \theta\}$, here t, d, l, r, f, b indicates the top, down, left, right, front and back sides, θ indicates the orientation angle, respectively. It should be noted that the rotation angle is only predicted for datasets with rotation annotations (SUNRGB-D and KITTI in this work). Instead of predicting the value of $s \in B$ as a Dirac distribution $\delta(s - \hat{s})$, we aim to predict the value of s as a probabilistic distribution over an interval. For the rotation angle, we discrete

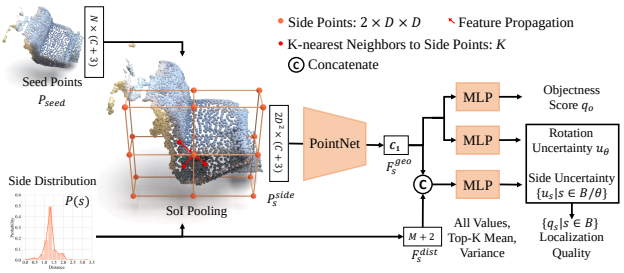

Fig. 3. **The side-aware quality estimation module.** The geometric features and distribution properties are extracted to estimate the pseudo-label quality. Here N represents the number of seed points, D represents the number of segments when generating side points, and M represents the number of bins in the side distribution.

the $[-\pi, \pi]$ interval into N_θ bins. For the boundary side, a fixed interval $[s_{min}, s_{max}]$ cannot suite all objects well because the side distance varies significantly for object with different sizes. Consequently, we use the proposal feature to predict a scaling parameter α_p to get more accurate interval for each candidate proposal, and then predict a probability distribution $P(s)$ defined over $[\alpha_p * s_{min}, \alpha_p * s_{max}]$. For convenience of implementations, we discrete the continuous distribution by dividing the interval into M bins $(s_0, s_1, \cdots, s_{M-1})$, where $s_i = \alpha_p * (s_{min} + i * \frac{s_{max} - s_{min}}{M-1})$. In this formulation, the expected value $\overline{s}$ of s can be calculated as Equation (1).

$$\overline{s} = \sum_{i=0}^{M-1} P(s = s_i) s_i. \tag{1}$$

3.3 Side-aware quality estimation

The side-aware quality estimation consists of two parts, the localization quality estimation (including side and rotation angle quality) and the objectness score estimation.

For the localization quality estimation, we first predict the uncertainty measure and then use it to calculate the quality score. Given the predicted distribution of each side, one naive approach

is to directly utilize statistical measures such as variance to assess the uncertainty of each side. However, this approach has certain limitations. Firstly, the relationship between distribution properties and the localization uncertainty of sides is complex, and cannot be accurately measured by simple statistics. Secondly, it's challenging to determine the uncertainty of each side without considering its geometric feature. Therefore, we propose to combine the distribution properties and geometric features of the side to predict its localization uncertainty through MLP, as shown in Figure 3. Specifically, for each side $s \in \{t, d, l, r, b, f\}$, we first generate a series of side points $P_s^{side} \in R^{2D^2 \times (C+3)}$ near its predicted position $\overline{s}$. We take the front side of a predicted 3D box with side location $\{\hat{t}, \hat{d}, \hat{l}, \hat{r}, \hat{b}, \hat{f}, \hat{\theta}\}$ as an example to introduce how to generate side points. Firstly, we generate two sibling planes around each side, which can be formally represented as $\{\hat{l} \leq x \leq \hat{r}, y = \hat{f} - \delta_y, \hat{d} \leq z \leq \hat{t}\}$, $\{\hat{l} \leq x \leq \hat{r}, y = \hat{f} + \delta_y, \hat{d} \leq z \leq \hat{t}\}$, where $\delta_y = 0.1 * (\hat{b} - \hat{f})$. Then, we rotate these two planes according to the predicted rotation angle. Secondly, for each rotated plane, we uniformly sample $D \times D$ side points to model the geometric structure of each side. Since the side points are virtual points, we find the k-nearest neighbors for each side point and apply a distance-weighted interpolation for feature propagation from the seed points P_{seed} to the side point. Lastly, we input all the side points into a PointNet to obtain the geometric features $F_s^{geo} \in \mathcal{R}^{C_1}$. Simultaneously, we concatenate the distribution of the side and corresponding statistics (Top-k value mean and variance) to obtain the distribution properties $F_s^{dist} \in \mathcal{R}^{M+2}$. Finally, we fuse the geometric features and distribution properties into a MLP to obtain the uncertainty measure $\{u_s | s \in \{t, d, l, r, f, b\}\}$ of sides as follows:

$$u_s = \text{Sigmoid}(\text{MLP}(\text{Cat}(F_s^{geo}, F_s^{dist}))). \tag{2}$$

For the uncertainty u_θ of the rotation angle, we concatenate $\{F_s^{geo}\}_{s \in \{t,d,l,r,b,f\}}$ and $\{F_\theta^{dist}\}$ to predict because all sides are rotated according to rotation angle θ and can reflect its quality.

$$u_\theta = \text{Sigmoid}(\text{MLP}(\text{Cat}(\{F_s^{geo} | s \in B/\theta\}, F_\theta^{dist})))). \tag{3}$$

By combining the uncertainty of all sides and the rotation angle, we can obtain the localization uncertainty measure $U = \{u_s | s \in B\}$. With the uncertainty measure, the localization quality score Q can be directly calculated as follows,

$$Q = \{q_s | q_s = e^{-\alpha_q u_s}, s \in B\}, \tag{4}$$

where α_q is a scaling parameter and is set to be 5 in this paper.

For the objectness score estimation, existing methods usually adopt the proposal feature to predict it, which may predict high objectness score for pseudo-label that only covers a small discriminative part of an object. To deal with this problem, we propose to use the side-aware features for objectness estimation by carefully considering the surrounding context information. Specifically, we concatenate the geometric feature $F_{geo} \in \mathcal{R}^{C_1}$ of six sides together and then feed the feature into a MLP layer to predict the objectness score of the proposal.

$$q_o = \text{Sigmoid}(\text{MLP}(\text{Cat}(\{F_s^{geo} | s \in B/\theta\}))). \tag{5}$$

Compared with the proposal feature, the side feature can carefully consider the boundary structure of the object, which can help to suppress pseudo-labels that only covers the most discriminative part of an object. It should be noted that the side-aware quality estimation design is only used in the training stage, thus the proposed method does not introduce any extra computation cost in the inference stage.

To guide the training of the side-aware quality estimation module, we introduce the quality regression loss into our method. For the localization quality, we ultize the absolute deviation of the predicted side value $\overline{s}$ and ground-truth s^{gt} to compute the groundtruth localization uncertainty u_s, as follows:

$$u_s^{gt} = \text{MIN}(\alpha_u |\overline{s} - s^{gt}|, 1.0), s \in \{t, d, l, r, b, f\}, \tag{6}$$

$$u_\theta^{gt} = \text{MIN}(|sin(\theta^{gt}) - sin(\overline{\theta})| + |cos(\theta^{gt}) - cos(\overline{\theta})|, 1), \tag{7}$$

where α_u is the scaling factor for uncertainty and is set to be 4 in this paper. Then the quality regression loss can be computed as the mean square error between u_s^{gt} and u_s:

$$\mathcal{L}_{uncer.} = \frac{1}{|U|} \sum_{s \in B} (u_s^{gt} - u_s)^2 \tag{8}$$

where $U = \{u_s | s \in B\}$ is the uncertainty estimation of each side of the pseudo-label. With this loss constraint, the localization quality estimation module can learn to measure the localization uncertainty of each side from labeled data, which can then be transferred to unlabeled data for model training. For the objectness score estimation, we directly use the objectness label to supervise the predicted objectness score from the side features.

3.4 Soft pseudo-Label selection

In the semi-supervised setting, the performance of the detector depends heavily on the quality of the pseudo-labels. Thus, we propose the Soft Pseudo-Label Selection (soft-PLS), which consists of three components: category-specific filter, IoU-guided NMS with low-half strategy, side-aware weight assignment.

For category-specific filter, we select pseudo-labels by jointly considering the classification score, objectness score q_o and IoU score. In this area, a portion of training data is randomly sampled to train the model. Consequently, the training object numbers of different categories may be very unbalanced, especially for training with a very small portion of the train dataset (e.g., 5%). The imbalance issue can result in different learning difficulties for different categories. To deal with this issue, we take inspiration from FlexMatch [31] and compute a scale factor $y_t(c)$ for each category. The main intuition is that categories with smaller number of objects tend to have lower prediction confidence, thus we need to decrease the confidence thresholds to select pseudo-labels for these categories. Denote the threshold for classification score, objectness score and IoU score as τ_{cls}, τ_{obj} and τ_{iou}, we first calculate a category-specific scale factor $\gamma_t(c)$ at time stamp t. Specifically, we define the learning progress of a category as the number of pseudo-labels used in the semi-supervised training process, as follows:

$$N_t(c) = \sum_{i=0}^{t} \text{Count}(i, c), \tag{9}$$

where the $\text{Count}(i, c)$ is the number of pseudo-labels for the category c in the iteration i. By applying the normalization to $N_t(c)$, we obtain the relative learning progress of each category, the formula is as follows:

$$\sigma_t(c) = \frac{N_t(c)}{\max_c \{N_t(c)\}}, \tag{10}$$

Due to the instability in the early stage of model training, we introduce a warm-up strategy in the above equation, as follows:

$$\beta_t(c) = \frac{N_t(c)}{\max\{\max_c\{N_t(c)\}, N - \sum_c N_t(c)\}}, \tag{11}$$

where N is the hyper-parameter of the warm-up process and we set N to be four times the quantity of unlabeled data. Finally, a convex function $\mathcal{M}(x) = \frac{x}{2-x}$ is applied to generate the category-specific scale factor $\gamma_t(c)$ as in Equation (12),

$$\gamma_t(c) = \mathcal{M}(\beta_t(c)). \tag{12}$$

Then, the adaptive threshold $\tau_t(c)$ for category c can be calculated as follows:

$$\tau_t(c) = \tau_{min} + (\tau_{max} - \tau_{min})\gamma_t(c), \tag{13}$$

where τ_{min} and τ_{max} control the range of the adaptive threshold. Since τ_{obj} is class-agnostic, we only apply this adaptive threshold strategy on τ_{cls} and τ_{iou}.

After obtaining pseudo-labels through the category-specific thresholds, we need to suppress noise in pseudo-labels caused by duplicated bounding box predictions. Therefore, we utilize the IoU-guided non-maximal suppression with low-half keeping strategy [11] to eliminate redundant pseudo-labels. Specifically, for a bunch of highly-overlapped pseudo-labels, we only discard half of the proposals with lower predicted IoU. Although better pseudo-labels can be obtained through the above modules, there are many pseudo-labels with poor localization quality and classification quality, which is detrimental to model training. To mitigate this issue, we use the quality score Q of the pseudo-label to weight the loss function as follows:

$$\mathcal{L}^u_{box} = q_B\mathcal{L}_{iou}(B) + \sum_{s\in B}(q_s\mathcal{L}_{reg}(s)), \tag{14}$$

$$\mathcal{L}^u_{cls} = q_o * (\mathcal{L}_{obj}(X^u, \hat{Y}^u) + \mathcal{L}_{cls}(X^u, \hat{Y}^u)). \tag{15}$$

Here q_B is the mean value of q_s and reflects the global localization quality of the bounding box. In this way, we can reduce the interference of pseudo-labels with poor quality in model training.

3.5 Model training

Our approach follows the Mean Teacher paradigm [25]. In the pre-training stage, we train the side-aware model on the labeled dataset $\{X^l, Y^l\}$. The training loss is the same as the original detector except that we add the uncertainty regression loss in Equation (8) and replace the regression with the side-aware regression loss in Equation (16).

$$\mathcal{L}^l_{box} = \mathcal{L}_{iou}(B) + \sum_{s\in B}\mathcal{L}_{reg}(s), \tag{16}$$

$$\mathcal{L}^l_{cls} = \mathcal{L}_{obj}(X^l, Y^l) + \mathcal{L}_{cls}(X^l, Y^l). \tag{17}$$

Here $\mathcal{L}_{iou}$ indicates the rotated IoU loss [49] and $\mathcal{L}_{reg}$ indicates the side-based smooth L1 loss. For the rotation angle θ, we first compute $sin(\theta), cos(\theta)$ and use them to compute the smooth L1 loss, because θ is periodic. For the six sides, the side-based smooth-L1 loss $\mathcal{L}_{reg}(s)$ focuses on the localization of each side, the rotated IoU loss $\mathcal{L}_{iou}(B)$ focuses on the global localization of the bounding box, which is robust to shape and scale variations. To help the training of the side quality estimation module, we also add random noise to the predicted bounding boxes to increase the prediction diversity. Once converged, we clone the model to create a pair of student and teacher models.

In the semi-supervised training stage, labeled data $\{X^l, Y^l\}$ and unlabeled data $\{X^u\}$ are randomly sampled according to a predefined sampling ratio r in each training iteration. For unlabeled data, we apply two different augmentations to the input of the teacher model and student model, respectively. The teacher model outputs the pseudo-labels $\{\hat{Y}^u\}$ and corresponding quality scores Q, which are then used together with the labeled data $\{X^l, Y^l\}$ to train the student model. For the supervised loss, $\mathcal{L}(X^l, Y^l)$ is computed in the same way as the pre-training stage. For the unsupervised loss, $\mathcal{L}(X^u, \hat{Y}^u)$ is defined as the sum of Equation (14) and Equation (15). The overall loss is defined as the weighted sum of the supervised loss and unsupervised loss:

$$\mathcal{L} = \mathcal{L}(X^l, Y^l) + \beta\mathcal{L}(X^u, \hat{Y}^u), \tag{18}$$

where $\mathcal{L}(X^l, Y^l)$ and $\mathcal{L}(X^u, \hat{Y}^u)$ denote the loss on labeled data and unlabeled data, respectively. During the training stage, the teacher model is updated by the student model with an Exponential Moving Average (EMA) strategy [25].

3.6 Discussions

In this section, we briefly discuss the difference with several works that also use probabilistic models and part-based assessment techniques. Firstly, probabilistic models are also used in GFLV2 [50] and LaserNet [41], the differences are discussed as follows: (1) In GFLV2, the main goal is to predict the global localization quality based on the spatial distribution of 2D bounding boxes. Specifically, GFLV2 employs a distribution-guided quality predictor to estimate a joint representation of classification and IoU. While in our method, the main goal is to estimate the quality score of each side individually and treat them with different importance during the training stage. We utilize the distribution properties and the geometric features of each side to estimate the quality score of each side individually and propose the side-aware quality regression loss to guide the training of the model. (2) LaserNet [41] models the global uncertainty of the bounding box based on pixel predictions within the same cluster. Specifically, LaserNet predicts the shared variance of the bounding box corners for each pixel and use mean-shift clustering to obtain the bounding box with the distribution variance. Different from LaserNet, our approach predicts the spatial distribution of each side and evaluates the quality of each side independently.

Secondly, part-based assessment techniques have been also used in several fully supervised 3D object detection methods, Part-A^2net [13], CasA [14] and SA-SSD [15], the differences are discussed as follows: (1) In Part-A^2net [13], they utilize the property of 3D box annotations to learn intra-object part location of the foreground point and then aggregate part-aware geometric features for box scoring and location refinement. (2) In CasA [14] and SA-SSD [15], they take inspiration from the PSRoIAlign operation in the 2D object detection area and propose its 3D variant to align the classification confidences with the predicted bounding boxes by performing spatial transformation on the feature maps. In these methods, the main goal is to learn robust global features by incorporating part-sensitive features, while our goal is to extract side-aware features for side quality estimation.

4 EXPERIMENTS

4.1 Experimental setup

Dataset. We evaluate the proposed method on two indoor datasets and one outdoor dataset, including ScanNet [51], SUNRGB-D [8] and KITTI [1]. For the indoor datasets, **ScanNet** [51] dataset contains 1.2K training samples and 312 validation samples annotated with per-point instances, semantic labels, and axis-aligned 3D bounding boxes belonging to 18 categories. **SUNRGB-D** [8] dataset consists of 5K single-view indoor RGB-D images annotated with per-point semantic labels, and oriented 3D bounding boxes belonging to 37 categories. For the outdoor dataset, **KITTI** [1] dataset contains 7481 outdoor scenes for training and 7518 scenes for testing, and the training samples are generally divided into a training split of 3712 samples and a validation split of 3769 samples. For all datasets, we follow [9] to evaluate on different proportions of labeled data randomly sampled from all the training data. We keep the remaining data as unlabeled data for training in our semi-supervised framework. As for 100% labeled data, we simply make a copy of the full dataset as unlabeled data.

Evaluation metrics. We follow the standard evaluation protocol [52] of the indoor datasets to evaluate the performance on the Val set with mAP values under IoU thresholds 0.25 and 0.5, denoted as mAP_{25} and mAP_{50}. For the outdoor dataset, We follow [4] 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 car, pedestrian, and cyclist categories, respectively. Due to the randomness of the data splits, we report the results as mean $\pm$ standard deviation across 3 runs.

Hyperparameter settings. For indoor datasets, we set the ranges of front, back, left and right sides as [0, 3.5], while the ranges of top and bottom sides are [0, 2.0]. For outdoor dataset, we set the ranges of front and back sides as [0, 0.4], while the ranges of left, right, top and bottom sides are [0, 0.3]. Since there are only three categories in this dataset, we use fixed IoU thresholds $\tau_{car} = 0.7, \tau_{ped} = 0.3, \tau_{cls} = 0.15$ in the category-specific filter of soft-PLS. The remaining hyper-parameters are the same for two datasets. For the probabilistic side localization, we split the range into $M = 32$ bins for six sides and $M = 20$ bins for the rotation angle. We consider a side with an error greater than 0.25 to be unreliable, thus we set $\alpha_u = 4$ in Equation (6) and $\alpha_q = 5$ in Equation (4). For pseudo-label selection, the objectness score threshold is set to be $\tau_{obj} = 0.8$, the minimum threshold τ_{min} for the classification score and IoU score are set to be 0.7 and 0.15, and the maximum threshold τ_{max} is set to be 0.9 and 0.25.

4.2 Baseline detectors.

For the indoor datasets, we conduct experiments with both group-based detectors and group-free detectors. Specifically, we use VoteNet [52] as the group-based model to be in line with previous works [9], [11]. Besides, we also conduct experiments with Diffusion-SS3D [33], which also uses VoteNet as baseline but uses improved techniques. For group-free detectors, we conduct experiments with GF3D [53] to verify the effectiveness of the proposed method as it's the first work to adopt the DETR architecture for group-free 3D object detection. For the outdoor dataset, we use PV-RCNN [4] as the baseline models to be in line with previous works. To verify the generalization ability, we also conduct experiments with Voxel-RCNN [54], an improved detector over PVRCNN. Different from VoteNet [52], PV-RCNN and Voxel-RCNN are two-stage detectors, where a region proposal network (RPN) is used to generate candidate proposals. The objectness score is predicted by the RPN and used to pick the top-100 proposals for the Region of Interest (RoI) head in the second stage. Details about each detector are as follows:

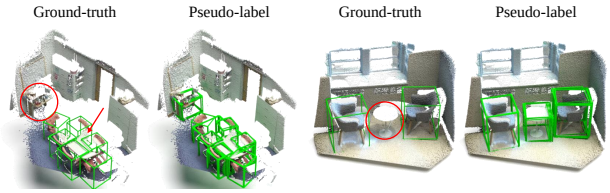

Fig. 4. **Annotations with low-quality on SUNRGB-D dataset.** We train the model on 20% labeled data and then generate pseudo-labels by the proposed soft-PLS.

VoteNet detector. During the pre-training stage, we use the Adam optimizer [55] with an initial learning rate of 0.008 and weight decay of 0.01. The batch size is set to 16, and we train the model for 360 epochs on a single NVIDIA GeForce RTX 3090 GPU, with the learning rate decayed by 0.1 at the 240th and 360th epochs. For the training stage, we form a batch by sampling 4 labeled samples and 8 unlabeled samples. The initial learning rate is set to 0.005, and we train the model for another 360 epochs with the same settings as the pre-training stage. The weight β for the unsupervised loss is set to 2.

GF3D detector. We use the standard version of GroupFree as the backbone, which has six layers of decoders and the feature dimension of 256. During the pre-training stage, we use the AdamW optimizer [55] with an initial learning rate of 0.006 and weight decay of 0.0005. The batch size is set to 8, and we train the model for 400 epochs on two NVIDIA GeForce RTX 3090 GPUs, with the learning rate decayed by 0.1 at the 280th and 340th epochs. For the training stage, we form a batch by sampling 8 labeled samples and 8 unlabeled samples. The initial learning rate is set to 0.004, and we train the model for another 400 epochs with the same settings as the pre-training stage. The weight β is set to 2.

PVRCNN detector. During the pre-training stage, we use the Adam optimizer with an initial learning rate 0.01 and weight decay 0.0025. We train the side-aware PV-RCNN [4] with a batch size of 32 for 80 epochs on 8 NVIDIA GeForce RTX 3090 GPUs. The learning rate is decayed by 0.1 at the 35th and 45th epochs. For the training stage, each batch consists of 8 labeled samples and 8 unlabeled samples. The weight β is set to 1.

Diffusion-SS3D detector. We use the official code of Diffusion-SS3D and combine the side-aware quality estimation method with it, the model is trained with the default parameter settings of the original paper [33]. It should be noted that the released code only provides the training code for the ScanNet dataset, we can only combine the proposed method with it on this dataset. The weight β is set to 1.

4.3 Comparison with state-of-the-art methods

Results on 3D ScanNet and SUNRGB-D dataset. For fair comparison with existing 3D based semi-supervised methods [9], [11], [52], we follow 3DIoUMatch [11] and train the proposed model on ScanNet and SUNRGB-D under different ratios of labeled data. Due to the randomness of the data splits, we report the mAP_{25} and mAP_{50} as mean $\pm$ standard deviation across

TABLE 1
Results on ScanNet Val dataset. Results are reported as mean ± standard deviation across 3 runs with random data splits. The rows with gray color represent the baseline model. * denotes baseline model with the proposed probabilistic side localization strategy. VoteNet* (Aug) and SA3Det++ (Aug) represent that we use the data augmentation strategy in DPKE to train the corresponding model.

Model	5%		10%		20%		50%		100%	
	mAP_{25}	mAP_{50}	mAP_{25}	mAP_{50}	mAP_{25}	mAP_{50}	mAP_{25}	mAP_{50}	mAP_{25}	mAP_{50}
VoteNet [52]	27.9±0.5	10.8±0.6	36.9±1.6	18.2±1.0	46.9±1.9	27.5±1.2	56.1±1.1	36.5±0.6	57.8	36.0
VoteNet* [52]	28.1±0.6	11.3±0.6	37.0±1.4	18.4±1.2	47.2±1.4	27.8±1.1	56.4±0.7	36.3±0.8	58.2	36.4
VoteNet* (Aug) [52]	32.3±0.9	17.1±0.8	44.2±1.4	22.8±1.0	47.9±1.3	30.1±1.0	59.2±1.3	40.8±0.7	64.5	46.2
SESS [9]	32.0±0.7	14.4±0.7	39.5±1.8	19.8±1.3	49.6±1.1	29.0±1.0	57.2±1.2	37.7±0.7	61.3	39.0
3DIoU [11]	40.0±0.9	22.5±0.5	47.2±0.4	28.3±1.5	52.8±1.2	35.2±1.1	59.8±0.7	41.2±0.5	62.9	42.1
DPKE [35]	44.0±1.1	27.0±1.9	51.9±1.4	34.1±0.7	57.6±0.8	41.4±1.1	-	-	65.3	48.7
SA3Det [16]	40.5±1.1	23.8±0.8	48.8±0.9	31.1±1.1	54.5±0	37.3±0.5	61.5±1.4	43.1±0.8	63.8	44.1
SA3Det++	41.8±1.6	25.2±1.3	49.5±1.4	32.8±1.2	54.8±0.7	38.3±0.8	62.4±0.4	44.5±0.5	64.2	44.9
SA3Det++ (Aug)	**45.9±1.2**	**28.4±1.5**	**52.6±1.6**	**36.6±1.1**	**59.4±0.9**	**43.7±0.9**	**64.1±0.9**	**46.8±0.8**	**65.6**	**48.9**
GF3D [53]	45.3±0.7	26.2±0.4	47.1±0.9	27.9±1.3	53.6±1.2	36.1±1.0	61.4±0.5	42.1±0.7	67.3	48.9
GF3D* [53]	45.4±0.4	26.4±0.5	47.3±1.2	27.8±1.1	53.8±1.4	36.2±1.3	61.2±0.6	42.2±0.5	67.4	49.1
3DIoU [11]	47.1±1.5	28.7±0.9	49.6±1.8	32.2±1.9	54.6±0.6	38.8±0.7	63.8±0.4	45.6±0.3	68.1	50.0
SA3Det [16]	48.3±1.0	30.5±0.8	51.2±1.4	33.4±1.2	56.6±1.5	40.3±0.7	65.7±0.6	47.4±0.4	68.4	50.3
SA3Det++	**49.8±1.2**	**31.9±0.9**	**52.4±1.1**	**34.7±1.3**	**57.8±0.8**	**41.7±0.5**	**66.7±0.7**	**48.9±0.4**	**68.6**	**50.8**
Diffusion-SS3D [33]	43.5±0.2	27.9±0.3	50.3±1.4	33.1±1.5	55.6±1.7	36.9±1.4	60.4±0.9	44.5±1.2	64.1	43.2
Diffusion-SS3D* [33]	43.7±0.5	28.0±0.8	50.6±1.2	33.4±1.7	55.6±1.4	37.2±1.8	60.6±0.9	44.8±1.4	64.4	43.5
SA3Det [16]	44.0±0.4	29.3±0.6	50.8±1.1	34.7±1.4	56.8±1.3	39.6±0.9	61.7±1.4	44.8±0.8	64.5	44.9
SA3Det++	**44.6±0.4**	**30.6±0.5**	**51.2±0.8**	**35.3±1.3**	**57.1±1.8**	**40.8±1.3**	**62.8±0.7**	**45.1±0.6**	**64.8**	**45.7**

TABLE 2
Results on SUNRGB-D Val dataset. Results are reported as mean ± standard deviation across 3 runs with random data splits. The rows with gray color represent the baseline model. * denotes baseline model with the proposed probabilistic side localization strategy. VoteNet* (Aug) and SA3Det++ (Aug) represent that we use the data augmentation strategy in DPKE to train the corresponding model.

Model	5%		10%		20%		50%		100%	
	mAP_{25}	mAP_{50}	mAP_{25}	mAP_{50}	mAP_{25}	mAP_{50}	mAP_{25}	mAP_{50}	mAP_{25}	mAP_{50}
VoteNet [52]	29.9±1.5	10.5±0.5	38.9±0.8	17.2±1.3	45.7±0.6	22.5±0.8	55.3±1.1	31.9±0.8	58.0	33.4
VoteNet* [52]	30.2±1.4	11.0±0.8	39.2±1.3	17.4±1.5	45.8±0.8	22.3±1.2	55.7±1.6	32.2±0.9	58.4	33.7
VoteNet* (Aug) [52]	32.8±1.8	15.6±1.2	41.3±1.4	22.6±0.8	48.2±0.7	28.8±0.4	57.4±0.5	36.3±0.6	61.7	40.8
SESS [9]	34.2±2.0	13.1±1.0	42.1±1.1	20.9±0.3	47.1±0.7	24.5±1.2	56.2±0.8	33.7±0.7	60.5	38.1
3DIoU [11]	39.0±1.9	21.1±1.7	45.5±1.5	28.8±0.7	49.7±0.4	30.9±0.2	58.3±0.9	35.6±0.4	61.5	41.3
Diffusion-SS3D [33]	**43.9±0.6**	24.9±0.3	49.1±0.5	30.4±0.7	51.4±0.8	32.4±0.6	-	-	-	-
DPKE [35]	41.5±1.0	25.0±1.2	49.9±1.0	32.5±0.4	53.3±0.2	35.0±0.2	-	-	63.9	46.9
SA3Det [16]	41.1±1.2	21.8±1.8	47.4±0.8	29.2±1.2	53.4±0.9	31.2±1.3	60.1±0.4	37.8±0.8	62.7	42.1
SA3Det++	42.5±1.5	25.1±1.6	48.2±0.8	29.8±1.2	52.9±0.8	32.6±1.3	60.1±0.4	40.3±0.7	62.8	42.6
SA3Det++ (Aug)	43.8±1.2	**27.6±1.0**	**51.7±1.4**	**33.8±0.6**	**55.7±1.4**	**36.7±0.5**	**63.2±0.5**	**44.1±0.7**	**64.4**	**47.3**
GF3D [53]	40.1±2.8	20.9±1.9	43.2±1.8	28.3±2.1	51.3±1.9	34.7±1.2	59.5±0.9	42.3±0.5	63.0	45.2
GF3D* [53]	40.4±2.3	21.1±2.1	43.3±1.9	28.5±2.4	51.5±1.7	34.9±1.1	59.4±0.6	42.6±0.7	63.2	45.2
3DIoU [11]	42.3±2.1	24.8±2.6	47.1±2.1	30.4±2.3	53.3±1.5	36.5±1.7	61.8±0.6	44.0±0.6	64.2	45.7
SA3Det [16]	43.8±2.1	26.2±2.3	48.8±1.9	31.5±2.3	54.5±1.8	37.7±1.6	62.2±0.6	44.9±0.4	64.6	46.1
SA3Det++	**44.5±2.4**	**27.1±2.1**	**49.6±1.6**	**32.7±1.9**	**55.6±1.8**	**38.8±1.3**	**62.5±0.8**	**45.3±0.3**	**64.8**	**46.2**

3 runs. For 100% labeled data, we make a copy of the full dataset as unlabeled data. As shown in Table 1 and Table 2, our approach shows significant performance improvements on both ScanNet and SUNRGB-D benchmark and achieves new SOTA performance under different ratios of labeled data. For ScanNet dataset, when adopting VoteNet as the baseline detector, the proposed SA3Det++ outperforms 3DIOUMatch by 4.5% and 3.1% for mAP_{50} on 10% and 20% labeled datasets, respectively. Although the performance of SA3Det++ falls behind DPKE (a contemporary work with SA3Det) in many metrics, SA3Det++ is complementary with DPKE because SA3Det++ focuses on pseudo-label quality estimation while DPKE focuses on data augmentation and feature-level enhancement. To verify this, we apply the data augmentation strategy of DPKE to train SA3Det++ and observe that SA3Det++ (Aug) significantly outperforms both DPKA and SA3Det++. When adopting GF3D as the baseline detector, the corresponding performance advantage is 2.5% and 2.9%, respectively. When adopting the Diffusion-SS3D as baseline detector, the corresponding performance advantage is 2.2% and 3.9%. For SUNRGB-D dataset, when adopting VoteNet as the baseline detector, SA3Det++ outperforms 3DIOUMatch by 2.7%/3.2% for mAP_{50} on 10%/20% labeled datasets, respectively. When adopting GF3D as the baseline detector, the corresponding performance advantage is 2.3%/2.3%. When further incorporated with the data augmentation strategy of DPKE, we observe similar phenomenon as that on ScanNet dataset, with 1.3%/1.7% improvement over DPKE for mAP_{50} on 10%/20% labeled datasets.

It should be noted that Diffusion-SS3D is designed to improve pseudo-labels quality through the denoising process, while our method is designed to evaluate the quality of pseudo-labels.

TABLE 3
Results on KITTI Val set. The results are reported as mean ± standard deviation across 3 runs with random data splits. The row with gray color represent the baseline model. * denotes baseline model with the proposed probabilistic side localization strategy.

Model	mAP(1%)			mAP(10%)			mAP(20%)			mAP(100%)		
	Car	Ped.	Cyc.	Car	Ped.	Cyc.	Car	Ped.	Cyc.	Car	Ped.	Cyc.
PV-RCNN [4]	73.1±0.2	21.4±11.1	28.0±6.0	80.7±1.0	50.0±3.2	60.5±4.7	82.4±0.2	52.4±1.5	65.8±1.3	82.5	58.1	73.5
PV-RCNN* [4]	73.3±0.8	21.3±13.2	28.6±6.8	80.9±0.8	50.3±3.8	60.8±4.1	82.5±0.4	52.2±1.2	66.2±1.1	82.7	57.8	73.8
3DIoU [11]	75.2±1.8	32.9±16.1	31.4±7.8	81.3±0.8	52.6±1.9	62.0±5.8	82.9±0.1	54.5±1.4	67.4±1.7	84.2	60.5	75.2
HSSDA [34]	**80.9**	**51.9**	**45.7**	-	-	-	82.5	**59.1**	**73.2**	-	-	-
Reliable Student [37]	76.5±0.4	42.3±7.7	36.2±11.6	80.3±0.4	49.8±3.5	66.9±3.3	53.2±10.9	32.6±6.0	67.5±0.1	-	-	
SA3Det [16]	76.3±1.0	33.1±13.6	33.6±5.2	83.1±0.5	54.2±1.9	65.3±3.6	84.1±0.2	57.8±1.5	70.8±0.5	80.3	60.9	**76.3**
SA3Det++	76.9±2.5	34.2±11.8	34.2±6.8	**83.8±0.6**	**54.4±2.9**	**65.8±5.2**	**85.2±0.7**	58.4±1.1	71.5±0.9	**85.5**	**61.1**	75.9

TABLE 4
Results on KITTI 1% labeled data. The results for all difficulty levels are evaluated by the mAP with 40 recall positions.

Model	Car			Pedestrian			Cyclist		
	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
PVRCNN [4]	87.4±0.0	73.1±0.2	66.9±0.8	30.2±14.9	21.4±11.1	19.2±9.5	47.1±10.1	28.0±6.0	26.1±5.8
3DIoU [11]	88.0±1.9	75.2±1.8	69.8±1.0	36.8±18.7	32.9±16.1	27.4±12.0	48.7±17.3	31.4±7.8	27.5±10.1
Reliable Student [37]	88.7±1.3	76.5±0.4	**71.7±0.4**	**47.3±9.9**	**42.3±7.7**	**39.0±6.5**	**58.8±18.9**	**36.2±11.6**	**34.1±11.3**
SA3Det	89.3±0.7	76.3±1.0	70.7±0.5	38.3±17.3	33.1±13.6	30.1±11.3	55.8±7.6	33.6±5.2	30.6±4.7
SA3Det++	**90.1±1.6**	**76.9±2.5**	71.2±1.4	38.8±17.9	34.2±11.8	31.0±16.9	56.3±7.2	34.2±6.8	31.6±5.5
Voxel-RCNN [54]	91.1±0.5	75.1±0.9	64.1±0.1	43.2±0.5	36.5±0.3	31.1±0.2	76.7±0.5	45.3±0.7	41.6±1.0
SA3Det++	**93.4±0.5**	**77.6±0.7**	**67.6±0.2**	**43.7±0.7**	**40.1±0.3**	**34.6±0.3**	**77.0±0.6**	**47.4±0.6**	**44.1±0.9**

Consequently, our method is complementary with Diffusion-SS3D and the results also verify this. In addition, with only 50% labeled data, our method achieves better performance than the fully supervised baseline model on both ScanNet and SUNRGB-D datasets. We notice that our method achieves performance gain when using 100% labeled data. The performance improvement may come from the suppression of dataset noise by the proposed semi-supervised framework. As shown in Figure 4, the pseudo-label generated by the teacher model is a correction for missing or incorrect annotations, providing more valuable information for student model training.

Results on KITTI dataset. To verify the adaptability of our method to object detection of different scenes, we conduct experiments on KITTI benchmark. Similar to the indoor datasets, we train the proposed model under different ratios of labeled data and report the results as mean ± standard. Different from indoor scenes, the training object number is very small due to the sparse property of outdoor scenes. For example, when using 1% labeled data for model training, the average labeled object numbers (averaged over three random splits) for car/pedestrian/cyclist is 170/24/8.7. In this situation, the pre-trained model is trained very insufficiently and tends to regard many pure background proposals as foreground objects and regard foreground proposals as backgrounds, where the side localization quality estimation is less important than improving the foreground/background discrimination ability. To deal with this problem, existing methods usually design data augmentation strategy (such as HSSDA [34]) to increase the labeled object numbers, or use global quality estimation strategy (such as HSSDA [34] and Reliable Student [37]) to improve the foreground/background classification ability. Although the proposed method can improve the performance by designing side-aware quality estimation strategy, its performance are inferior than methods (HSSDA [34] and Reliable Student [36]) that are specifically designed for outdoor dataset under severe data hungry situations, as shown in Table 3 and Table 4. However, when increasing the labeled data ratio, the foreground/background discrimination ability of pretrained model becomes stronger, the side localization quality becomes more and more important. As shown in Table 4, HSSDA outperforms SA3Det++ by 17.7% for the pedestrian category when using 1% labeled data for model training, but it outperforms SA3Det++ by only 0.7% when using 20% labeled data. The same phenomenon can be also observed on Reliable Student, where SA3Det++ even outperforms Reliable Student by using 10% and 20% labeled data. These results indicate that SA3Det++ and HSSDA, Reliable Student are suited for different situations, and inspires us that how to combine the advantage of these methods may be our further exploration direction.

Per-category results comparison. We present per-category results using 50% labeled data on both ScanNet and SUNRGB-D datasets. Table 5 shows the mAP@IoU=0.25 and mAP@IoU=0.5 for each category on the ScanNet dataset with 50% labeled data. Similarly, Table 6 shows the mAP@IoU=0.25 and mAP@IoU=0.5 for each category on the SUNRGB-D dataset with 50% labeled data. These results indicate that the proposed method achieves promising results for most categories under both datasets. Our method shows superior performance in detecting objects belonging to certain challenging categories such as tables, pictures, windows, and bookshelves. For the outdoor datset, we divide all objects into three difficulty levels according to the height range, the occlusion level and the truncation of the bounding box.

4.4 Ablation Studies

In this section, we perform a series of ablation studies to evaluate the effectiveness of each design. It's worth noting that all experiments are conducted on 3D object detection unless specified.
Evaluation of the model with different designs. In this section, we present our extensive ablation studies on the ScanNet and SUNRGB-D datasets to evaluate the effectiveness of different

TABLE 5
Pre-category mAP on ScanNet with 50% labeled data. Results are reported as mean ± standard deviation across 3 runs with random splits.

	IoU=0.25						IoU=0.5					
Method	**cabinet**	**bed**	**chair**	**sofa**	**table**	**door**	**cabinet**	**bed**	**chair**	**sofa**	**table**	**door**
VoteNet [52]	31.9±1.5	85.5±0.6	86.1±0.9	82.0±0.4	57.2±0.9	45.8±1.3	8.7±1.2	70.9±0.9	68.8±1.0	69.6±0.8	44.5±1.1	16.4±1.6
SESS [9]	41.0±1.6	86.4±1.2	88.2±1.1	88.7±0.8	59.8±1.1	49.5±1.5	12.3±0.8	75.7±0.4	71.9±0.4	74.1±0.7	51.2±0.8	18.6±1.3
3DIoU [11]	44.2±0.5	**87.3±0.7**	88.4±0.4	91.0±0.3	59.1±1.0	51.8±1.3	12.5±1.1	**76.7±0.5**	73.8±0.9	**81.7±0.4**	49.3±0.3	26.1±0.5
SA3Det [16]	46.3±1.1	86.8±0.5	89.1±0.3	**91.3±0.8**	64.4±0.9	50.9±1.3	18.6±2.1	74.1±0.6	76.5±0.6	81.3±0.9	56.7±1.5	26.4±1.1
SA3Det++	**50.0±1.4**	87.0±0.8	**89.9±0.6**	89.1±0.7	**65.4±0.5**	**52.3±1.1**	**22.8±1.1**	74.9±0.4	**79.9±0.8**	80.8±0.9	**59.4±1.2**	**29.0±1.4**
Method	**window**	**bkshf**	**picture**	**counter**	**desk**	**curtain**	**window**	**bkshf**	**picture**	**counter**	**desk**	**curtain**
VoteNet [52]	29.9±1.8	46.4±1.4	7.5±1.1	68.1±0.9	67.3±1.2	44.1±1.0	8.7±1.8	36.5±0.5	1.2±1.1	29.4±1.0	39.0±0.8	24.6±1.2
SESS [9]	35.7±1.9	52.8±0.5	10.6±0.6	60.9±1.1	67.9±1.1	36.7±1.9	9.6±1.5	43.2±0.6	2.2±0.8	19.7±1.2	38.6±0.6	25.0±0.9
3DIoU [11]	37.1±0.5	51.9±0.9	11.4±0.8	65.1±0.7	65.1±0.8	41.9±1.3	14.6±0.5	42.6±0.6	4.2±1.2	31.1±0.8	40.4±0.5	29.1±1.1
SA3Det [16]	39.6±1.8	**55.5±1.1**	**15.8±1.7**	69.1±0.8	**74.3±0.5**	44.8±1.2	13.6±0.8	**45.9±1.2**	**6.5±1.3**	**42.3±0.7**	**48.9±0.9**	**29.2±1.0**
SA3Det++	**40.6±1.4**	51.4±1.9	14.3±1.3	**69.1±0.4**	71.5±0.8	**51.2±1.4**	**18.4±1.5**	44.6±1.7	6.4±0.8	33.1±0.3	45.0±0.5	28.2±1.2
Method	**fridg**	**showr**	**toilet**	**sink**	**bathtub**	**ofurn**	**fridg**	**showr**	**toilet**	**sink**	**bathtub**	**ofurn**
VoteNet [52]	46.2±1.6	63.4±0.7	96.5±1.2	34.8±1.5	89.4±0.6	29.5±0.9	**35.2±0.7**	2.1±0.9	85.8±0.4	14.9±1.3	80.9±0.3	13.8±1.6
SESS [9]	44.5±0.8	64.1±0.4	98.8±0.3	32.9±1.6	92.1±0.8	37.5±1.7	33.4±0.7	3.7±0.8	**89.7±0.6**	15.3±1.7	89.6±0.5	19.7±1.2
3DIoU [11]	49.4±1.2	61.1±0.9	98.6±0.4	**43.3±0.8**	89.9±0.5	37.5±0.9	33.8±0.9	2.3±1.1	84.8±0.4	24.9±1.1	**89.7±0.8**	20.8±1.5
SA3Det [16]	**51.2±0.5**	54.9±2.2	**99.8±0.2**	42.8±1.5	**92.2±1.1**	38.6±0.6	33.3±1.7	2.1±1.8	85.8±0.5	25.5±1.1	82.9±0.6	26.1±1.2
SA3Det++	47.9±0.8	**75.4±1.4**	99.0±0.7	43.1±1.1	86.6±0.6	**38.9±0.7**	32.1±0.6	**12.0±1.2**	89.3±0.5	**28.0±1.0**	83.6±0.8	**28.5±0.9**

TABLE 6
Pre-category mAP on SUNRGB-D with 50% labeled data. Results are reported as mean ± standard deviation across 3 runs with random splits.

IoU	**Method**	**bed**	**table**	**sofa**	**chair**	**toilet**	**desk**	**dresser**	**nights**	**bkshf**	**bathtub**
0.25	VoteNet [52]	82.2±0.6	47.0±0.9	61.1±0.7	76.7±0.4	85.8±0.4	16.6±2.7	28.3±0.5	54.7±0.2	23.5±1.7	74.4±0.3
	SESS [9]	83.5±0.3	48.8±1.1	63.0±0.5	77.7±0.6	86.7±0.2	20.3±1.6	30.3±0.7	56.1±0.5	29.0±0.9	**79.8±1.2**
	3DIoU [11]	83.9±0.8	48.5±0.4	65.2±0.9	77.3±0.2	87.6±0.7	25.8±1.1	29.8±1.1	56.8±0.4	29.4±1.4	78.9±0.6
	SA3Det [16]	85.5±0.5	**54.1±0.8**	67.4±0.6	78.9±0.8	**90.6±0.4**	27.3±0.7	**31.4±0.9**	62.3±0.6	**32.3±0.2**	71.2±0.4
	SA3Det++	**87.6±0.7**	51.4±0.9	**68.1±0.7**	**79.0±0.9**	87.6±0.6	**28.3±1.6**	29.2±0.5	**68.6±0.6**	30.9±0.9	69.3±0.8
0.5	VoteNet [52]	46.1±0.4	19.5±0.7	46.2±0.5	57.8±0.4	52.6±0.3	3.1±1.1	14.2±0.6	30.8±0.3	2.1±0.6	47.3±0.5
	SESS [9]	41.7±0.9	20.2±0.5	48.4±1.3	58.3±1.2	57.2±0.5	4.7±0.6	15.6±0.2	38.7±0.6	3.3±0.5	**50.6±0.7**
	3DIoU [11]	53.0±0.7	22.9±0.6	50.1±0.7	59.5±0.5	59.0±0.4	7.1±1.3	17.5±0.7	36.8±0.4	5.4±0.8	45.3±0.2
	SA3Det [16]	55.3±0.3	**26.2±1.1**	51.3±0.6	60.0±0.8	62.7±1.2	**8.4±0.9**	**20.9±0.5**	42.3±0.2	**8.1±1.2**	42.9±0.4
	SA3Det++	**65.6±0.5**	25.3±0.8	**53.6±0.9**	**61.3±0.5**	**70.9±0.6**	8.3±0.8	16.7±0.3	**51.2±0.4**	7.2±0.7	43.6±0.6

TABLE 7
Evaluation of model with different designs. PSL is the Probabilistic Side Localization. Soft-PLS is the Soft Pseudo-Label Selection. SAQE is the Side Aware Quality Estimation Module. F_{geo} and F_{dist} are the geometric and distribution properties, respectively.

PSL	Soft-PLS	SAQE		ScanNet 20%		ScanNet 50%		SUNRGB-D 5%		SUNRGB-D 10%	
		F_{geo}	F_{dist}	mAP_{25}	mAP_{50}	mAP_{25}	mAP_{50}	mAP_{25}	mAP_{50}	mAP_{25}	mAP_{50}
✗	✗	✗	✗	50.35	30.21	58.23	38.26	32.53	12.51	40.52	19.94
✓	✗	✗	✗	50.52	30.77	58.25	38.48	33.07	13.29	41.81	20.26
✓	✓	✗	✗	51.38	33.49	58.97	40.24	37.87	16.22	43.97	26.31
✓	✓	✓	✗	54.39	37.94	61.88	43.38	41.85	24.38	47.14	28.68
✓	✓	✗	✓	53.86	37.42	61.47	42.64	41.23	23.89	46.56	27.95
✓	✓	✓	✓	**54.84**	**38.29**	**62.41**	**44.53**	**42.54**	**25.07**	**48.23**	**29.82**

designs. Table 7 reports the performance of our model under various settings. In the first line, we use VoteNet [52] as the backbone and adopt a fixed threshold setting for pseudo-label selection. In the second line, we modify the baseline model with our proposed probabilistic side localization. Although it does not yield a significant improvement in the model performance, it sets a solid foundation for subsequent investigations. In the third line, we utilize the distribution variance as the quality assessment of the sides and then filter the pseudo-labels by soft-PLS. The performance improvement indicates the effectiveness of our proposed pseudo-label filtering strategy. In the fourth line, we evaluate the quality of the sides using geometric features extracted from the SoI pooling. In the fifth line, we use distribution properties for the same purpose. Both methods achieve better performance than the distribution variance. Finally, in the last line, we combine the geometric and distribution properties to evaluate the quality of the sides and achieve the best results. The ablation studies reveal that each design is necessary.

Evaluation of quality estimation for different components. In Table 8, we report the results of quality estimation for different components. As indicated by the results, the quality estimation for each side can improve the mAP_{50} by 4.73% with 5% labeled data.

TABLE 8
Ablation studies on different quality estimation components.

Side	Ratation Angle	Objectness Score	SUNRGB-D 5% mAP_{25}	SUNRGB-D 5% mAP_{50}	SUNRGB-D 10% mAP_{25}	SUNRGB-D 10% mAP_{50}
✗	✗	✗	37.87	16.22	43.97	26.31
✓	✗	✗	40.82	20.95	46.65	28.46
✓	✓	✗	42.16	24.67	47.86	29.64
✓	✓	✓	**42.54**	**25.07**	**48.23**	**29.82**

TABLE 9
Results of different side localization methods. Naive parameterization is to predict the center and sizes, while side-aware parameterization is to predict six sides of the 3D bounding boxes. Probabilistic method means predicting a probabilistic distribution for each side.

Method	Prob.	ScanNet 20% mAP_{25}	ScanNet 20% mAP_{50}	ScanNet 50% mAP_{25}	ScanNet 50% mAP_{50}
w/ naive param.	✗	50.48	31.58	58.55	38.67
w/ naive param.	✓	51.64	32.66	59.05	39.13
w/ side. param.	✗	52.86	35.23	60.64	41.55
w/ side. param.	✓	**54.84**	**38.29**	**62.41**	**44.53**

By adding the rotation angle quality estimation, the performance can be improved by 3.72%. Finally, when incorporating the side-aware objectness score eatimation, the performance can be further improved by 0.4%. These results verify the effectiveness of each design.

Evaluation of different side localization methods. To demonstrate the effectiveness of our proposed parameterization method for semi-supervised 3D object detection, we conduct detailed ablation experiments, as presented in Table 9. In the first line, we use a naive parameterization method to estimate the uncertainty of centers and sizes in the uncertainty estimation module. Additionally, we change the weight assignment of each side to match the weight assignment of the corresponding regression values. In the second line, we make the parameterization approach probabilistic and use both distribution properties and geometric features to estimate the uncertainty of the regression values. However, both of the above methods result in severe performance degradation due to the strong correlation between individual regression values in the naive parameterization method. This correlation is difficult to accurately assess the localization quality of each regression values individually. Our results from the third and fourth lines further confirm the importance of using distribution properties in the uncertain estimation module.

Evaluation of the soft pseudo-label selection strategy. We present Table 10 to demonstrate the impact of each designed component on the pseudo-label selection strategy. In the first line, we utilize fixed thresholds to remove pseudo-labels with low-quality. As shown in the second line, we introduce the IoU-guided NMS with a low-half strategy to reduce the interference of repeated bounding boxes during model training and improve performance. In the third line, we adopt category-specific adaptive thresholds to address the long-tail effects and retain more pseudo-labels for model training. Finally, we ultize the side-aware weight assignment strategy based on the uncertainty of each side. This approach effectively suppresses the interference of false regression values on model training, resulting in a significant performance

TABLE 10
Results for different pseudo-label selection methods. LHS is the IoU-guided NMS with Low-Half Strategy. CSF is the Category-Specific Filter with adaptive threshold. SWA is the Side-aware Weight Assignment.

LHS	CSF	SWA	ScanNet 20% mAP_{25}	ScanNet 20% mAP_{50}	ScanNet 50% mAP_{25}	ScanNet 50% mAP_{50}
✗	✗	✗	50.52	30.77	58.25	38.48
✓	✗	✗	53.11	35.48	60.21	41.24
✓	✓	✗	53.59	36.22	61.03	41.88
✓	✓	✓	**54.84**	**38.29**	**62.41**	**44.53**
✗	✓	✓	54.15	37.85	61.92	43.75

improvement. To further demonstrate the effect of the side-aware weight assignment strategy, we remove the IoU-guided NMS with a low-half strategy. The results show that the method can still achieve good performance.

Evaluation of different bins number for side probability distributions and rotation angle estimation. Table 11 and Table 12 illustrates the effect of the number of bins in side probability distributions and rotation angle estimation on the model performance. Increasing the number of bins leads to a slight improvement in model performance, while decreasing the number of bins significantly reduces the model's performance. This indicates that the granularity of the distribution plays a critical role in uncertainty estimation and model performance. To balance computational efficiency and performance, we set the number of bins to 32 and 20 for side probability distributions and rotation angle estimation, respectively.

TABLE 11
Results for different number of bins for side probability distributions. N denotes the number of bins.

Number of side bins	ScanNet 20% mAP_{25}	ScanNet 20% mAP_{50}	ScanNet 50% mAP_{25}	ScanNet 50% mAP_{50}
N = 16	53.89	36.66	60.95	42.63
N = 24	54.21	37.22	61.64	43.37
N = 32	**54.84**	38.29	**62.41**	**44.53**
N = 64	54.65	**38.51**	62.28	44.41
N = 96	54.56	37.98	61.89	43.52

TABLE 12
Ablation studies about the bin numbers for rotation angle estimation.

Number of angle bins	SUNRGB-D 5% mAP_{25}	SUNRGB-D 5% mAP_{50}	SUNRGB-D 10% mAP_{25}	SUNRGB-D 10% mAP_{50}
12	41.33	23.83	47.7	28.98
16	42.16	24.46	48.13	29.57
20	42.54	**25.07**	**48.23**	29.82
24	**42.67**	24.83	47.85	29.63
28	42.44	24.96	48.05	**30.16**

Evaluation of different distribution properties for side quality estimation. Here we conduct experiments to investigate the effect of different distribution properties on the model performance. As shown in Table 13, besides using the distribution values, we introduce three statistical measures (top-k mean, variance, entropy) that reflect the flatness of the distribution. The

TABLE 13
Results of different distribution properties on quality estimation. "All Values" refers to using all distribution values as input. "Top-k" involves selecting the top-k values and computing their mean value as the property. "Variance" and "Entropy" correspond to calculating the distribution variance and entropy as properties, respectively.

All Values	Top-k	Variance	Entropy	ScanNet 20%		ScanNet 50%	
				mAP_{25}	mAP_{50}	mAP_{25}	mAP_{50}
✗	✗	✗	✗	52.66	35.51	60.84	42.26
✓	✗	✗	✗	53.22	36.08	61.57	43.52
✓	k=4	✗	✗	54.01	36.88	62.41	44.53
✓	k=8	✗	✗	54.65	37.34	61.89	43.79
✓	k=12	✗	✗	53.62	36.71	61.11	42.86
✓	k=8	✓	✗	**54.84**	**38.29**	**62.41**	**44.53**
✓	k=8	✓	✓	54.76	38.09	62.34	44.51

TABLE 14
Transductive learning results on ScanNet dataset under different ratios of labeled data. Results are reported as mean across 3 runs with random data splits.

Model	10%		20%		50%	
	mAP_{25}	mAP_{50}	mAP_{25}	mAP_{50}	mAP_{25}	mAP_{50}
VoteNet [52]	43.5	24.5	51.3	32.7	61.6	41.2
SESS [9]	48.8	28.7	56.6	36.9	67.6	44.2
3DIoU [11]	51.2	32.7	57.8	38.1	66.8	45.9
SA3Det [16]	53.8	34.6	60.2	40.4	68.5	48.5
SA3Det++	**55.6**	**36.1**	**61.8**	**42.1**	**69.4**	**49.7**

Top-k mean is insensitive to relative shifts over the distribution, resulting in a robust representation that is independent of the object scale. Based on the experimental results, we ultimately use the distribution values, top-k mean, and variance as distribution property inputs into the uncertainty estimation module.

4.5 Transductive Learning Comparison

Semi-supervised learning uses both labeled data and unlabeled data for model training. There are two categories of semi-supervised learning, based on the type of testing data: inductive learning and transductive learning. Inductive learning use the unseen data as testing data, while transductive learning use the testing data as unlabeled data in the training stage. To evaluate the effectiveness of our method in transductive learning, we conduct experiments on the ScanNet dataset under different ratios of labeled data. As shown in Table 14, our method outperforms all previous methods under different ratios of labeled data, which verifies the effectiveness of the proposed method for both inductive and transductive semi-supervised learning.

4.6 Sensitivity analysis

As shown in Table 15, we present the sensitivity analysis of the heuristics in the Soft Pseudo-Label Selection. By using different threshold settings to select pseudo-labels, we evaluate the robustness of our approach. The performance of the model only slightly decreased when using lower thresholds, which further confirms the effectiveness of our proposed soft-PLS in suppressing the interference of low-quality pseudo-labels on model training. However, setting higher thresholds results in a significant decrease in model performance. This is due to the fact that the high threshold filters out a large number of valuable pseudo-labels.

TABLE 15
Sensitivity analysis of the heuristics in the Soft Pseudo-Label Selection.

τ_{obj}	τ_{cls}^{min}	τ_{cls}^{max}	τ_{IoU}^{min}	τ_{IoU}^{max}	ScanNet 20%	
					mAP_{25}	mAP_{50}
0.60	0.50	0.80	0.10	0.20	53.95	37.25
0.70	0.60	0.85	0.15	0.25	**55.00**	37.98
0.80	0.70	0.90	0.15	0.25	54.84	**38.29**
0.90	0.70	0.95	0.20	0.30	53.78	37.13
0.90	0.80	0.95	0.25	0.35	52.64	35.61

TABLE 16
Memory usage and runtime comparison of different methods. Here we report the memory usage and time consumption for model training, including the pretraining stage and the training stage. The runtime consumption is calculated by performing a forward pass and a backward pass through the model.

	Method	ScanNet		SUNRGB-D	
		Mem. (GB)	RunTime (s)	Mem. (GB)	RunTime (s)
Pretrain	VoteNet [52]	16.573	0.322	16.527	0.301
	SESS [9]	16.573	0.322	16.527	0.301
	3DIoU [11]	17.133	0.382	17.078	0.375
	SA3Det [16]	18.253	0.391	18.211	0.381
	SA3Det++	23.090	0.422	23.052	0.409
Train	SESS [9]	12.281	0.403	12.264	0.391
	3DIoU [11]	16.909	0.915	16.868	0.901
	SA3Det [16]	17.789	0.951	17.767	0.919
	SA3Det++	18.766	0.987	18.708	0.938

4.7 Efficiency analysis

In Table 16, we report the memory usage and time consumption for model training, including the pretraining stage and the training stage. All results are produced with the same experiment setting on a single GTX 3090 GPU. The runtime consumption is computed with a forward pass and a backward pass for the both pretraining and training stage. The memory usage is computed with batch size 16 for the pretraining stage and batch size 12 for the training stage (4 labeled samples and 8 unlabeled samples). During the pretraining stage, since we have an extra uncertainty estimation network and IoU prediction network, the memory usage and time consumption are both slightly increased. During the training stage, our method is similar in speed to the pseudo-label based method 3DIoUMatch [11], but is slower than the consistency based method SESS [9]. This is because both our method and 3DIoUmatch need extra time to generate pseudo-labels.

4.8 Results visualization

In Figure 5, we present the visualization results obtained by different methods on the ScanNet 50% labeled data and the SUNRGB-D 50% labeled data, respectively. The results demonstrate that our method achieves superior performance with higher localization quality. This is attributed to the fact that we assign different weights to the sides with different localization qualities, which in turn helps to improve the localization ability by learning from sides with higher quality. In Figure 6, we provide visualization results of u_s on more samples, we can observe that the proposed method can consistently assign higher uncertainty score for sides with poor localization quality.

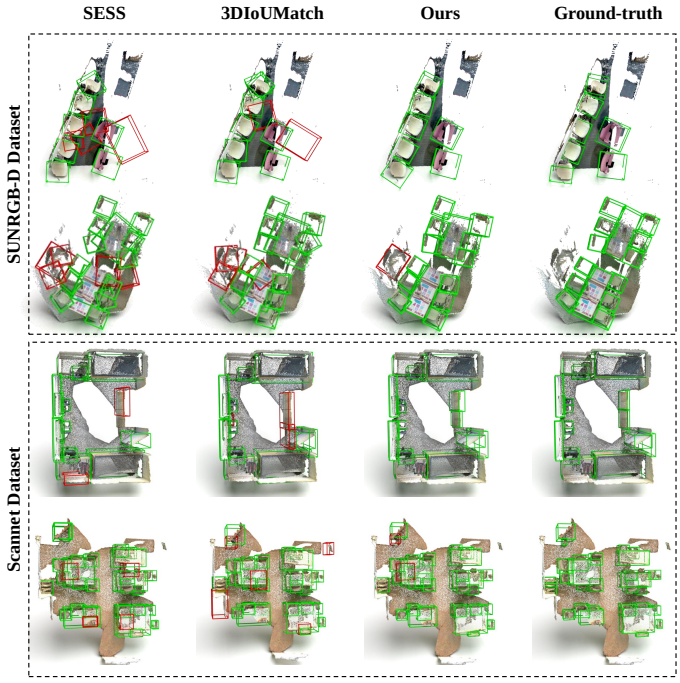

Fig. 5. **Visual comparisons of the detection results on ScanNet 50% labeled data.** Here green bounding boxes have $IoU \geq 0.25$ and red bounding boxes have $IoU < 0.25$.

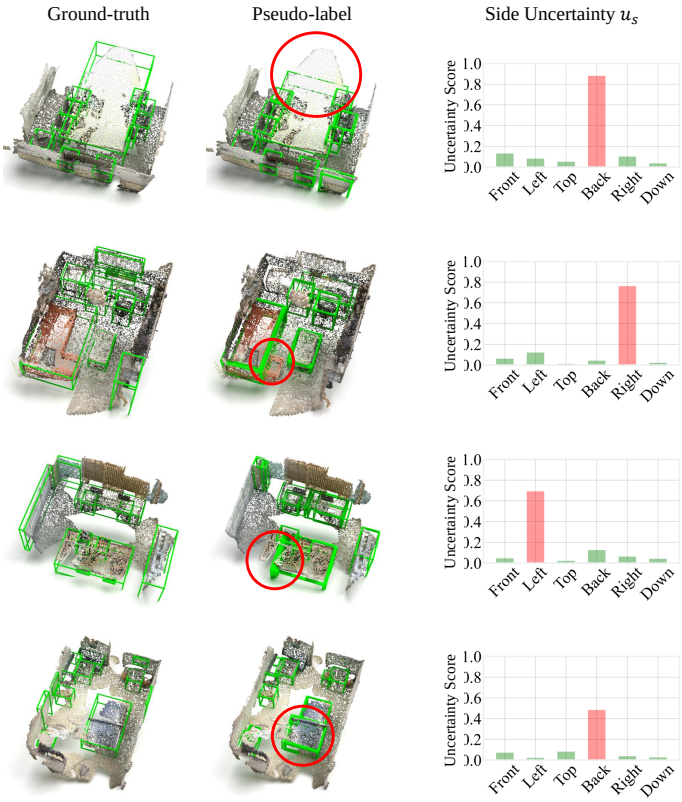

Fig. 6. Visualization results of side localization quality estimation results. Here we visualize the uncertainty score u_s, where lower value represents higher localization quality.

4.9 Failure case analysis

In Figure 7, we provide visualization results of two representative failure cases. The failure cases mainly come from two situations: (1) As illustrated in the first row of Figure 7, when numerous objects of the same category cluster together, the model tends to predict multiple overlapping detection results. This phenomenon occurs because the model struggles to extract discriminative features to delineate object boundaries under these scenarios, resulting in suboptimal suppression of erroneous pseudo-labels. (2) As demonstrated in the second row of Figure 7, the model often exhibits inaccurate boundary predictions for objects with significantly shape variations. For example, the model regard the sofa with L shape as two short sofas. This is because the model can hardly understand whether this is one object or the combination of two objects, since both situations are quite similar.

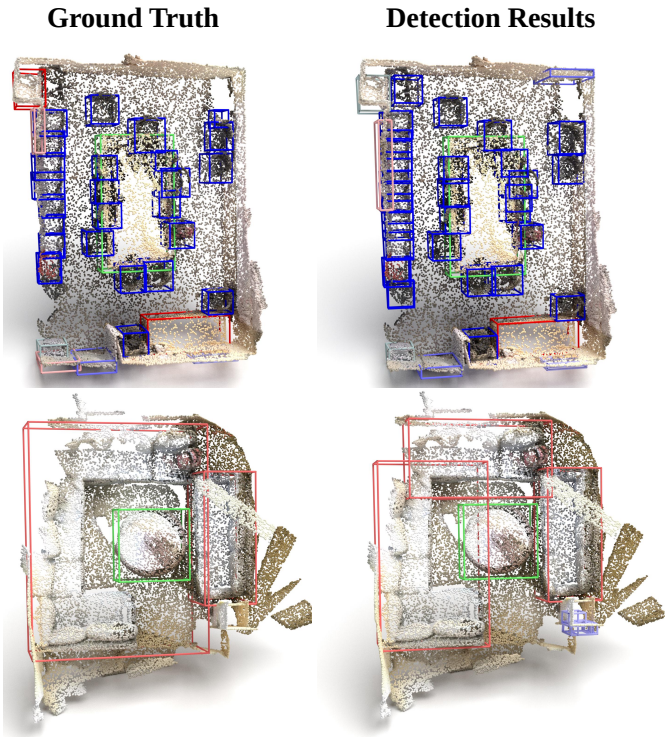

Fig. 7. Visualization of failure cases. Here the left column represents ground-truth, and the right column represents prediction results.

5 Conclusion

In this paper, we propose a side-aware method for semi-supervised 3D object detection, which includes a probabilistic side localization module, an side-aware quality estimation module, and a soft pseudo-label selection module. To the best of our knowledge, this is the first work to consider the pseudo-label quality from the side perspective for pseudo-label filtering, enabling full exploitation and utilization of valid information in the model prediction results for supervising student models. Experiment results indicate that our method can achieve consistent improvements over different baseline detectors on two indoor datasets and one outdoor dataset.

REFERENCES

[1] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[2] R. Gomez-Ojeda, J. Briales, and J. Gonzalez-Jimenez, "Pl-svo: Semi-direct monocular visual odometry by combining points and line segments," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 4211–4216.

[3] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby, and A. Mouzakitis, "A survey on 3d object detection methods for autonomous driving applications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3782–3795, 2019.

[4] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. 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*, 2020, pp. 10 529–10 538.

[5] D. Rukhovich, A. Vorontsova, and A. Konushin, "Fcaf3d: Fully convolutional anchor-free 3d object detection," *arXiv preprint arXiv:2112.00322*, 2021.

[6] L. Fan, Y. Yang, F. Wang, N. Wang, and Z. Zhang, "Super sparse 3d object detection," *IEEE transactions on pattern analysis and machine intelligence*, 2023.

[7] S. Shi, Z. Wang, J. Shi, X. Wang, and H. Li, "From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 8, pp. 2647–2664, 2020.

[8] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 567–576.

[9] N. Zhao, T.-S. Chua, and G. H. Lee, "Sess: Self-ensembling semi-supervised 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 079–11 087.

[10] H. Xu, F. Liu, Q. Zhou, J. Hao, Z. Cao, Z. Feng, and L. Ma, "Semi-supervised 3d object detection via adaptive pseudo-labeling," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 3183–3187.

[11] H. Wang, Y. Cong, O. Litany, Y. Gao, and L. J. Guibas, "3dioumatch: Leveraging iou prediction for semi-supervised 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 615–14 624.

[12] Z. Zhang, Y. Ji, W. Cui, Y. Wang, H. Li, X. Zhao, D. Li, S. Tang, M. Yang, W. Tan *et al.*, "Atf-3d: Semi-supervised 3d object detection with adaptive thresholds filtering based on confidence and distance," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 573–10 580, 2022.

[13] S. Shi, Z. Wang, J. Shi, X. Wang, and H. Li, "From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 8, pp. 2647–2664, 2020.

[14] H. Wu, J. Deng, C. Wen, X. Li, C. Wang, and J. Li, "Casa: A cascade attention network for 3-d object detection from lidar point clouds," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[15] C. He, H. Zeng, J. Huang, X.-S. Hua, and L. Zhang, "Structure aware single-stage 3d object detection from point cloud," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 873–11 882.

[16] C. Wang, W. Yang, and T. Zhang, "Not every side is equal: Localization uncertainty estimation for semi-supervised 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3814–3824.

[17] Y. Tang, J. Wang, X. Wang, B. Gao, E. Dellandréa, R. Gaizauskas, and L. Chen, "Visual and semantic knowledge transfer for large scale semi-supervised object detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 3045–3058, 2017.

[18] Y. Li, J. Zhang, K. Huang, and J. Zhang, "Mixed supervised object detection with robust objectness transfer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 3, pp. 639–653, 2018.

[19] I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari, and K. He, "Data distillation: Towards omni-supervised learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4119–4128.

[20] B. Zoph, G. Ghiasi, T.-Y. Lin, Y. Cui, H. Liu, E. D. Cubuk, and Q. Le, "Rethinking pre-training and self-training," *Advances in neural information processing systems*, vol. 33, pp. 3833–3845, 2020.

[21] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in neural information processing systems*, vol. 33, pp. 596–608, 2020.

[22] K. Sohn, Z. Zhang, C.-L. Li, H. Zhang, C.-Y. Lee, and T. Pfister, "A simple semi-supervised learning framework for object detection," *arXiv preprint arXiv:2005.04757*, 2020.

[23] Y.-C. Liu, C.-Y. Ma, Z. He, C.-W. Kuo, K. Chen, P. Zhang, B. Wu, Z. Kira, and P. Vajda, "Unbiased teacher for semi-supervised object detection," in *International Conference on Learning Representations*, 2020.

[24] Y. Tang, W. Chen, Y. Luo, and Y. Zhang, "Humble teachers teach better students for semi-supervised object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3132–3141.

[25] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[26] Q. Yang, X. Wei, B. Wang, X.-S. Hua, and L. Zhang, "Interactive self-training with mean teachers for semi-supervised object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5941–5950.

[27] M. Xu, Z. Zhang, H. Hu, J. Wang, L. Wang, F. Wei, X. Bai, and Z. Liu, "End-to-end semi-supervised object detection with soft teacher," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3060–3069.

[28] L. Liu, B. Zhang, J. Zhang, W. Zhang, Z. Gan, G. Tian, W. Zhu, Y. Wang, and C. Wang, "Mixteacher: Mining promising labels with mixed scale teacher for semi-supervised object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7370–7379.

[29] X. Wang, X. Yang, S. Zhang, Y. Li, L. Feng, S. Fang, C. Lyu, K. Chen, and W. Zhang, "Consistent-teacher: Towards reducing inconsistent pseudo-targets in semi-supervised object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3240–3249.

[30] G. Li, X. Li, Y. Wang, W. Yichao, D. Liang, and S. Zhang, "Dtg-ssod: Dense teacher guidance for semi-supervised object detection," *Advances in neural information processing systems*, vol. 35, pp. 8840–8852, 2022.

[31] B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinozaki, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[32] H.-a. Gao, B. Tian, P. Li, H. Zhao, and G. Zhou, "Dqs3d: Densely-matched quantization-aware semi-supervised 3d detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 905–21 915.

[33] C.-J. Ho, C.-H. Tai, Y.-Y. Lin, M.-H. Yang, and Y.-H. Tsai, "Diffusion-ss3d: Diffusion model for semi-supervised 3d object detection," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[34] C. Liu, C. Gao, F. Liu, P. Li, D. Meng, and X. Gao, "Hierarchical supervision and shuffle data augmentation for 3d semi-supervised object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 819–23 828.

[35] Y. Han, N. Zhao, W. Chen, K. T. Ma, and H. Zhang, "Dual-perspective knowledge enrichment for semi-supervised 3d object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 3, 2024, pp. 2049–2057.

[36] Z. Chen, Z. Li, S. Wang, D. Fu, and F. Zhao, "Learning from noisy data for semi-supervised 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 6929–6939.

[37] F. Nozarian, S. Agarwal, F. Rezaeianaran, D. Shahzad, A. Poibrenski, C. Müller, and P. Slusallek, "Reliable student: Addressing noise in semi-supervised 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4981–4990.

[38] N. Durasov, T. Bagautdinov, P. Baque, and P. Fua, "Masksembles for uncertainty estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 539–13 548.

[39] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *Advances in neural information processing systems*, vol. 30, 2017.

[40] J. Choi, D. Chun, H. Kim, and H.-J. Lee, "Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 502–511.

[41] G. P. Meyer, A. Laddha, E. Kee, C. Vallespi-Gonzalez, and C. K. Wellington, "Lasernet: An efficient probabilistic 3d object detector for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 677–12 686.
[42] Y. Shi, J. Zhang, T. Ling, J. Lu, Y. Zheng, Q. Yu, L. Qi, and Y. Gao, "Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation," *IEEE Transactions on Medical Imaging*, 2021.
[43] K. Zheng, C. Lan, W. Zeng, Z. Zhang, and Z.-J. Zha, "Exploiting sample uncertainty for domain adaptive person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 3538–3546.
[44] Y. He, C. Zhu, J. Wang, M. Savvides, and X. Zhang, "Bounding box regression with uncertainty for accurate object detection," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2019, pp. 2888–2897.
[45] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 002–21 012, 2020.
[46] G. P. Meyer and N. Thakurdesai, "Learning an uncertainty-aware object detector for autonomous driving," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 10 521–10 527.
[47] Y. Lu, X. Ma, L. Yang, T. Zhang, Y. Liu, Q. Chu, J. Yan, and W. Ouyang, "Geometry uncertainty projection network for monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 3111–3121.
[48] H. Pan, Z. Wang, W. Zhan, and M. Tomizuka, "Towards better performance and more explainable uncertainty for 3d object detection of autonomous vehicles," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, 2020, pp. 1–7.
[49] D. Zhou, J. Fang, X. Song, C. Guan, J. Yin, Y. Dai, and R. Yang, "Iou loss for 2d/3d object detection," in *2019 International Conference on 3D Vision (3DV)*. IEEE, 2019, pp. 85–94.
[50] X. Li, W. Wang, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 632–11 641.
[51] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5828–5839.
[52] C. R. Qi, O. Litany, K. He, and L. J. Guibas, "Deep hough voting for 3d object detection in point clouds," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9277–9286.
[53] Z. Liu, Z. Zhang, Y. Cao, H. Hu, and X. Tong, "Group-free 3d object detection via transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2949–2958.
[54] J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li, "Voxel r-cnn: Towards high performance voxel-based 3d object detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 2, 2021, pp. 1201–1209.
[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

Wenfei Yang received the bachelor's degree in Electronic Engineering and Information Science in 2017, and the Ph.D. degree in pattern recognition and intelligent systems from the department of Automation, University of Science and Technology of China, Hefei, China, in 2022. Currently, he is a post-doctor in Control Science and Engineering, University of Science and Technology of China. His current research interests include computer vision and machine learning, especially action detection and object detection.

Chuxin Wang received the bachelor's degree in Control Science and Engineering from University of Science and Technology of China, Hefei, China, in 2021. He is currently pursuing the Ph.D degree in pattern recognition and intelligent systems from the department of Automation, University of Science and Technology of China, Hefei, China. His research interests include computer vision and machine learning, especially 3D object detection, 3D instance segmentation, and point cloud completion.

Tianzhu Zhang (M'11) received the bachelor's degree in communications and information technology from Beijing Institute of Technology, Beijing, China, in 2006, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2011. Currently, he is a Professor at the Department of Automation, School of Information Science, University of Science and Technology of China. His current research interests include computer vision and multimedia, especially action recognition, object classification, object tracking, and social event analysis.

Yongdong Zhang (M'09-SM'13-F'25) received the Ph.D. degree in electronic engineering from Tianjin University, Tianjin, China, in 2002. He is currently a Professor at the University Of Science And Technology Of China. He has authored more than 100 refereed journal and conference papers. His current research interests include multimedia content analysis and understanding, multimedia content security, video encoding, and streaming media technology. Prof. Zhang serves as an Editorial Board Member of Multimedia Systems Journal and Neurocomputing. He was the recipient of the Best Paper Award in PCM2013, ICIMCS 2013, and ICME 2010, and the Best Paper Candidate in ICME 2011.

Feng Wu (M'99-SM'06-F'13) received the B.S. degree in Electrical Engineering from XIDIAN University in 1992. He received the M.S. and Ph.D. degrees in Computer Science from Harbin Institute of Technology in 1996 and 1999, respectively. Now he is a professor in University of Science and Technology of China and the dean of School of Information Science and Technology. Before that, he was principle researcher and research manager with Microsoft Research Asia. His research interests include image and video compression, media communication, and media analysis and synthesis. He has authored or co-authored over 200 high quality papers (including several dozens of IEEE Transaction papers) and top conference papers on MOBICOM, SIGIR, CVPR and ACM MM. He has 77 granted US patents. His 15 techniques have been adopted into international video coding standards. As a co-author, he got the best paper award in IEEE T-CSVT 2009, PCM 2008 and SPIE VCIP 2007. Wu has been a Fellow of IEEE. He serves as an associate editor in IEEE Transactions on Circuits and System for Video Technology, IEEE Transactions on Multimedia and several other International journals. He got IEEE Circuits and Systems Society 2012 Best Associate Editor Award. He also serves as TPC chair in MMSP 2011, VCIP 2010 and PCM 2009, and Special sessions chair in ICME 2010 and ISCAS 2013.