

Dual-Perspective Knowledge Enrichment for Semi-supervised 3D Object Detection

Yucheng Han¹, Na Zhao^{2*}, Weiling Chen³, Keng Teck Ma³, Hanwang Zhang¹

¹Nanyang Technological University

²Singapore University of Technology and Design

³Hyundai Motor Group Innovation Center in Singapore

yucheng002@ntu.edu.sg, na_zhao@sutd.edu.sg, {weiling.chen, kengteck.ma}@hmgics.com, hanwangzhang@ntu.edu.sg

Abstract

Semi-supervised 3D object detection is a promising yet under-explored direction to reduce data annotation costs, especially for cluttered indoor scenes. A few prior works, such as SESS and 3DIoUMatch, attempt to solve this task by utilizing a teacher model to generate pseudo-labels for unlabeled samples. However, the availability of unlabeled samples in the 3D domain is relatively limited compared to its 2D counterpart due to the greater effort required to collect 3D data. Moreover, the loose consistency regularization in SESS and restricted pseudo-label selection strategy in 3DIoUMatch lead to either low-quality supervision or a limited amount of pseudo labels. To address these issues, we present a novel Dual-Perspective Knowledge Enrichment approach named DPKE for semi-supervised 3D object detection. Our DPKE enriches the knowledge of limited training data, particularly unlabeled data, from two perspectives: data-perspective and feature-perspective. Specifically, from the data-perspective, we propose a class-probabilistic data augmentation method that augments the input data with additional instances based on the varying distribution of class probabilities. Our DPKE achieves feature-perspective knowledge enrichment by designing a geometry-aware feature matching method that regularizes feature-level similarity between object proposals from the student and teacher models. Extensive experiments on the two benchmark datasets demonstrate that our DPKE achieves superior performance over existing state-of-the-art approaches under various label ratio conditions. The source code will be made available to the public.

Introduction

3D object detection, which can localize and categorize objects of interest in a 3D scene, is a crucial technology for numerous applications, including domestic robotics, autonomous driving, and augmented reality. Despite the effectiveness of modern 3D object detection approaches (Yang et al. 2023; Wang et al. 2022; Rukhovich, Vorontsova, and Konushin 2023; Zhang et al. 2022a; Xu, Zhong, and Neumann 2022; Zheng et al. 2021), their performance heavily relies on the availability of a large amount of well-annotated 3D data. However, compared to its 2D counterpart, annotating 3D data is more time-consuming and expensive. To

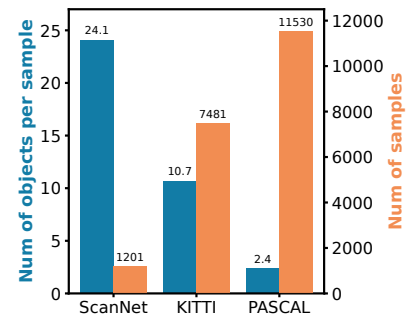


Figure 1: The dataset statistics. The orange bars represent the number of samples in the corresponding dataset, while the blue bars represent the number of objects per sample. The 3D indoor dataset (ScanNet) contains more objects per scene than the 3D outdoor dataset (KITTI) or the 2D dataset (Pascal).

be more specific, the process of annotating a single image takes approximately 0.7 minutes (Papadopoulos et al. 2016), whereas annotating an outdoor 3D scene can require up to 1.9 minutes (Lee et al. 2018). The task becomes even more arduous in the case of cluttered indoor 3D scenarios, such as annotating a single scan in the 3D indoor benchmark dataset ScanNet (Dai et al. 2017), which can take up to 22.3 minutes. The excessively high annotation costs make the scale of 3D indoor datasets significantly smaller than that of 3D outdoor datasets and 2D datasets as shown in Figure 1. In order to alleviate the burden of labor-intensive data annotation, our research delves into semi-supervised 3D object detection, which aims to achieve comparable performance to fully supervised approaches while reducing the amount of required annotated data by exploring knowledge from unlabeled data.

A few prior works (Zhao, Chua, and Lee 2020; Wang et al. 2021a; Chen et al. 2023) on semi-supervised 3D object detection adopt the Mean Teacher paradigm (Tarvainen and Valpola 2017), which involves a student and a teacher model, with the teacher model providing supervision to the student outputs of the unlabeled data. Specifically, SESS (Zhao, Chua, and Lee 2020) considers all outputs of the teacher model as pseudo labels and enforces the consis-

*Corresponding author.

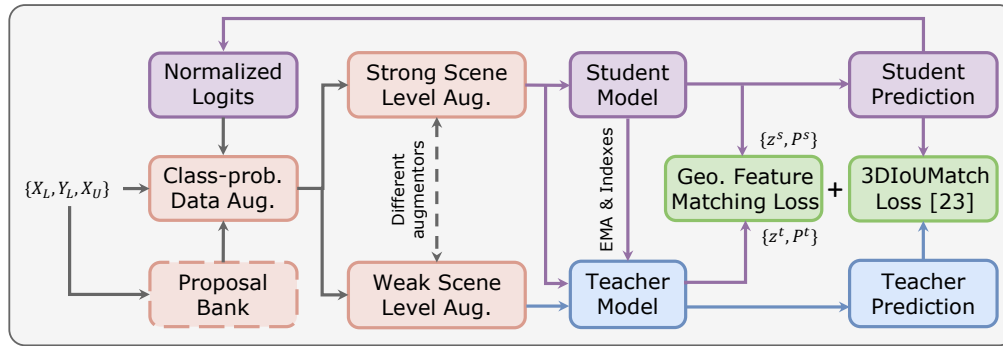


Figure 2: The overall framework of our proposed DPKE. Before the original scene-level augmentation operator, we introduce a new data augmentation module based on the normalized logits of the student model to increase the diversity of the input data. The student model is updated using the loss function combined by 3DIoUMatch Loss (Wang et al. 2021a) and our geometry-aware feature matching loss, which exploits knowledge from teacher predictions with lower confidence.

tency regularization between two model outputs in terms of center, size, and class. As a result, SESS tends to produce imprecise pseudo labels that exhibit high recall but poor quality. In contrast, methods like 3DIoUMatch (Wang et al. 2021a) apply a strict threshold strategy to select high-quality pseudo labels, which leads to a lower recall and thus fails to utilize the unlabeled data to its full potential. Finding a good balance between high recall and high-quality pseudo labels is crucial for the success of semi-supervised 3D object detection. In addition, compared to their 2D counterparts, 3D datasets are substantially smaller in scale, mainly due to the increased effort required to collect the data. For instance, even the largest 3D indoor benchmark dataset, SUN RGB-D (Song, Lichtenberg, and Xiao 2015), contains only 10,335 samples. This limited scale significantly restricts the data diversity and hinders the overall model performance.

To this end, we present a novel approach for semi-supervised 3D object detection named Dual-Perspective Knowledge Enrichment (DPKE). DPKE aims to enrich the knowledge of limited training data, particularly unlabeled data, from two perspectives: data-perspective and feature-perspective. Specifically, for the data-perspective, we adopt data augmentation inspired by semi-sampling (Wu et al. 2022) that randomly samples instances from labeled data and inserts them into the current training samples, to enhance the diversity of the input data, thereby enriching the knowledge of the data level. Unlike the uniform sampling for each class in (Wu et al. 2022), we design a class-probabilistic data augmentation method that selects sampled instances based on the varying distribution of class probabilities. These probabilities are computed from the model predictions, which are able to reflect the variance of class imbalance and learning difficulty. Our class-probabilistic data augmentation method can be viewed as a form of curriculum learning, as it gradually increases the sampling probability of instances in terms of classes with increasing difficulty during the training process as the model capacity improves.

Additionally, we introduce feature-perspective knowledge enrichment to compensate for the high-quality yet low-recall pseudo labels as in 3DIoUMatch. This is achieved

by proposing a geometry-aware feature matching approach that can effectively exploit knowledge from teacher proposals with lower confidence. Instead of applying consistency regularization on model outputs like SESS, DPKE enforces consistency regularization on proposal features, as they contain richer information regarding potential objects. Before matching proposal features, a lower threshold is adopted in terms of objectness score to filter out background proposals. Subsequently, the geometry similarity between the points from the bounding box of the two proposals is measured, and this similarity is used to weigh the feature matching. This is because proposals with similar geometry are more likely to represent the same object, and therefore their features should be matched more strongly during the knowledge enrichment process.

The **main contribution** of this paper can be summarized as: 1) We propose a novel Dual-Perspective Knowledge Enrichment approach to address the challenges in semi-supervised 3D object detection, such as limited data diversity and unsatisfied pseudo labels. 2) We design a class-probabilistic data augmentation method for the data perspective and geometry-aware feature matching method for the feature perspective. 3) Our proposed method achieves state-of-the-art performance on two benchmark datasets for semi-supervised 3D object detection, demonstrating the effectiveness of our approach in enriching the knowledge from limited data.

Related Work

Semi-supervised 2D object detection. Many recent studies (Jeong et al. 2019; Chen et al. 2022; Liu et al. 2023, 2021; Liu, Ma, and Kira 2022; Sohn et al. 2020; Tang et al. 2021; Xu et al. 2021; Zhang, Pan, and Wang 2021; Zhou et al. 2022) have explored the realm of semi-supervised 2D object detection. Prior works have drawn insights from traditional semi-supervised learning techniques in the context of image classification. For instance, STAC (Sohn et al. 2020) employs a two-stage training scheme for Faster-RCNN, where the first stage model provides pseudo labels for the subsequent stage. Subsequent works widely adopt the pseudo

labeling strategy, leveraging the student-teacher paradigm (Xu et al. 2021; Liu et al. 2021; Liu, Ma, and Kira 2022; Li et al. 2022; Chen, DeBattista, and Han 2022), and focus on developing novel techniques to better utilize the teacher model’s supervision signal. (Liu et al. 2021) presents the Unbiased Teacher and a class-balance loss to eliminate pseudo-labeling bias. Unbiased Teacher V2 (Liu, Ma, and Kira 2022) further introduces the Listen2Student mechanism for anchor-free detectors to address the ineffectiveness of pseudo-labeling bounding box regression. PseCo (Li et al. 2022) concentrates on merging pseudo labeling and consistency training, proposing a noisy pseudo box learning method for robust label assignment and localization quality evaluation. VCL (Chen, DeBattista, and Han 2022) offers a solution that leverages confusing samples without label correction by assigning a virtual category to each sample.

Semi-supervised 3D object detection. Traditional semi-supervised learning techniques also benefit semi-supervised 3D object detection. However, due to disparities in data domain and model structure, state-of-the-art semi-supervised 2D object detection methods cannot be directly or trivially applied to the 3D domain. Current approaches (Zhao, Chua, and Lee 2020; Wang et al. 2021a,b; Park et al. 2022; Zhang et al. 2022b; Yin et al. 2022; Chen et al. 2023; Wu et al. 2022) mainly build upon frameworks introduced by earlier studies (Zhao, Chua, and Lee 2020; Wang et al. 2021a). One significant difference between semi-supervised detection in the 2D and 3D domains is the distinct nature of the backbones or techniques employed for indoor and outdoor scenes in 3D domain, due to the unique characteristics and complexities of indoor and outdoor environments. While semi-supervised 3D object detection in outdoor scenes (Wang et al. 2021a,b; Park et al. 2022; Zhang et al. 2022b; Yin et al. 2022) has shown promising results, the performance of its indoor counterpart remains sub-optimal. Several methods have been proposed to improve semi-supervised 3D object detection in indoor scenarios. Some methods, such as (Chen et al. 2023; Wang et al. 2021a), focus on improving pseudo labeling to enhance performance. 3DIoUMatch (Wang et al. 2021a) designs an IoU branch and applies high thresholds to select high-quality pseudo labels, but at the cost of reduced pseudo label recall.

Methodology

Problem Definition

In 3D semi-supervised detection, the goal is to leverage both labeled and unlabeled data for training detection models. The labeled data consists of N_L samples, represented as $\{X_L^i, Y_L^i\}_{i=1}^{N_L}$, where X_L^i represents i -th point cloud and Y_L^i represents the corresponding annotation. The annotation Y_L^i includes information such as the category label y_{cls} , the center location c_x, c_y, c_z , the size s_x, s_y, s_z , and the orientation θ of interested objects. In addition to the labeled data, we also have N_U unlabeled data, denoted as $\{X_U^i\}_{i=1}^{N_U}$. These unlabeled samples do not have corresponding annotations but can still be utilized to improve the performance of the detection models. Note that N_U and N_L are determined by

the label ratios set in the experiments, but typically follow the condition $N_U \gg N_L$.

Dual-Perspective Knowledge Enrichment

Our DPKE approach, as shown in Figure 2, follows the student-teacher paradigm (Tarvainen and Valpola 2017) commonly used in existing semi-supervised 3D detection methods (Wang et al. 2021a; Wu et al. 2022; Zhao, Chua, and Lee 2020). Both the student and teacher models share the same backbone architecture, specifically an improved version of VoteNet (Qi et al. 2019) from (Wang et al. 2021a). The student and teacher models are initialized with parameters exclusively trained on labeled data during the pre-training stage. In the training stage, the student model’s weights are updated using the loss function gradients, while the teacher model’s weights are updated using the exponential moving average of the student model’s weights.

Following previous works (Wang et al. 2021a; Wu et al. 2022; Zhao, Chua, and Lee 2020), our DPKE utilizes different augmentation methods for the student and teacher models. The teacher model undergoes weak augmentation of sub-sampling on the input point clouds. On the other hand, the student model is subjected to a strong augmentation scheme, including random flipping and scaling of the point clouds. However, unlike traditional scene level augmentation techniques, our DPKE introduces a class-probabilistic data augmentation method. This method adaptively augments instances into the scenes based on their class probabilities, leading to data-perspective knowledge enrichment.

Our DPKE combines the high-precision yet low-recall pseudo labels used in 3DIoUMatch with knowledge from less confident proposals to improve overall performance. To achieve this, our DPKE introduces a geometry-aware feature matching method. This method applies adaptive consistency regularization on the features of potential foreground proposals, based on the geometry similarity between points within the proposal bounding boxes. The computed geometry-aware feature matching loss is combined with the original loss used in 3DIoUMatch to optimize the student model’s parameters.

Class-Probabilistic Data Augmentation

Our class-probabilistic data augmentation method draws inspiration from semi-sampling (Wu et al. 2022) but does not use the instance segmentation annotations in the ScanNet dataset. Instead, we employ a process where we randomly crop bounding boxes from labeled scenes, creating a *proposal bank*. These bounding boxes are then inserted into the present scenes, similar to the gt-sampling (Yan, Mao, and Li 2018) technique used in fully-supervised 3D outdoor object detection. During the insertion process, we conduct collision detection to ensure that the inserted boxes do not intersect with existing objects in the scenes. However, since we do not have ground truth bounding boxes for the unlabeled scenes, we use the prediction results from the unlabeled scenes for collision detection. This allows us to leverage the information from the unlabeled data in a meaningful way during the data augmentation process.

Both semi-sampling and gt-sampling approaches typically apply data augmentation in a class-uniform manner. However, in the case of indoor scenes, where there are multiple categories with different learning statuses (Chen et al. 2023), it becomes necessary to assign different sampling probabilities to each category. By incorporating category-specific sampling probabilities, our class-probabilistic data augmentation method can better balance the augmentation process and address the specific learning needs of each category in the indoor scene dataset. Specifically, to assign different sampling probabilities to each category, we utilize the average logits for each category during training. These logits are then normalized into probabilistic weights, ranging from 0 to 1. We achieve this normalization by first applying min-max normalization to ensure the values are within the desired range and subsequently applying the sigmoid function to obtain the final probabilistic weights. We do not use the softmax function to obtain the probabilistic weights because its exponential operation would amplify differences between input values and produce a sharper/distorted output, which harms the diversity of the augmented scenes. Subsequently, we use the obtained probabilities to sample the proposals from the proposal bank and paste it into the existing scene at the first N epochs. The experimental results in Figure 4 show the superiority of our class-probabilistic data augmentation method over the class-uniform semi-sampling.

Additional Analysis. We analyze our class-probabilistic data augmentation method by concentrating on the seed features generated during the intermediate stage of VoteNet (Qi et al. 2019). A detailed description of seed features can be found in the Appendix A. Formally, $P(y|f_c, f_n)$ represents the predicted probabilities of various predictions y , such as the locations and classes of proposals, given the seed features within the target proposals (f_c) and the seed features outside the target proposals (f_n). Here we expect the model to learn the relations between f_c and y , which is meaningful. By leveraging the relations between (f_c, f_n) and y , we can derive the following equation:

$$P(y|f_c, f_n) = P(f_c, y) \cdot \frac{P(f_n|y, f_c)}{P(f_n, f_c)}. \quad (1)$$

$P(f_c, y)$ represents the joint probability of f_c and the prediction y , which the model aims to learn. However, considering that f_n is unrelated to f_c and y , the term $P(f_n|y, f_c)/P(f_n, f_c)$ captures the potential for significant fluctuations when the model struggles to effectively distinguish between f_c and f_n . These fluctuations can impact the training process as it now must accommodate the additional variability introduced by the f_n features. To address this issue, our sampling strategy that pastes more well-learned object proposals into existing scenes appears to be a feasible solution, which can simultaneously balance enhancing diversity and improving model learning stability.

Geometry-aware Feature Matching

3DIoUMatch (Wang et al. 2021a) adopts high thresholds to select pseudo labels with high precision. However, it

comes at the cost of decreasing the recall of pseudo labels. To address this, we propose a new geometry-aware feature matching method that focuses on enriching knowledge from proposals with lower confidence. Our method introduces geometry-aware weights for each proposal alignment to regulate the discrepancy in proposal features between the student and teacher models.

The presence of random down-sampling steps, such as farthest point sampling (FPS), in the model backbone leads to different sets of proposals for the student and teacher models when given the same input point cloud. Consequently, the alignment of proposals between the two models cannot be guaranteed. To address this issue, we follow (Zhao and Lee 2022) to save the indexes of selected points from the student model, which are reused to sample points in FPS when processing the same input with the teacher model. This ensures that the predicted results of the teacher model and student model are aligned, enabling the computation of a feature matching loss.

After aligning the proposals, we obtain the proposal features z^s and z^t for the student and teacher models, respectively. These features are obtained by concatenating the features from all three convolutional layers in the proposal module (Qi et al. 2019). The feature matching loss is computed as:

$$L_\delta(z^s, z^t) = \begin{cases} \frac{1}{2}(z^s - z^t)^2, & \text{if } |z^s - z^t| \leq \delta, \\ \delta|z^s - z^t| - \frac{1}{2}\delta^2, & \text{otherwise.} \end{cases} \quad (2)$$

To focus on informative proposals, we apply an objectness score threshold τ_{obj} that is lower than the one used for selecting high-precision pseudo labels. This allows us to exclude numerous background proposals that may not contribute significantly to knowledge enrichment. Additionally, in complex scenes where objects can overlap, even slight changes in the locations and sizes of proposals can result in noticeable variations in geometry properties. We believe that proposal pairs with similar geometry properties should undergo stronger regularization than those with dissimilar geometry. To incorporate the similarity of geometry properties, we utilize the Chamfer distance (Fan, Su, and Guibas 2017) to calculate the similarity between point sets within the two proposals. Let $P^t \in \mathbb{R}^{n_t \times 3}$ and $P^s \in \mathbb{R}^{n_s \times 3}$ represent the point sets in the predicted proposals of the teacher model and student model, respectively, with n_t and n_s denoting the number of points in each set. Based on the Chamfer distance, we can calculate the geometric-aware weight for a proposal pair as follows:

$$\mathcal{W}_{\text{geometry}} = \exp\left(-\sum_{i=1}^{n_t} \min_{P_j^s \in P^s} \|P_i^t - P_j^s\|_2^2 - \sum_{i=1}^{n_s} \min_{P_j^t \in P^t} \|P_i^s - P_j^t\|_2^2\right). \quad (3)$$

However, directly calculating $\mathcal{W}_{\text{geometry}}$ may consume too many computational resources because we have no limitations on the number of points in P^t and P^s . We address this

| Dataset | Model | 5% (mAP) | | 10% (mAP) | | 20% (mAP) | | 100% (mAP) | |
|-----------|--------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-------------|-------------|
| | | @0.25 | @0.5 | @0.25 | @0.5 | @0.25 | @0.5 | @0.25 | @0.5 |
| ScanNet | VoteNet | 27.9 | 10.8 | 36.9 | 18.2 | 46.9 | 27.5 | 57.8 | 36.0 |
| | SESS | – | – | 39.7 | 18.6 | 47.9 | 26.9 | 62.1 | 38.8 |
| | 3DIoUMatch* | 39.4±1.2 | 22.2±1.1 | 47.5±1.8 | 28.6±1.4 | 52.7±1.6 | 35.3±0.7 | 62.8 | 42.2 |
| | Confid-3DIoUMatch* | 39.9±0.6 | 24.0±0.4 | 47.9±2.7 | 29.3±1.5 | 53.2±0.4 | 35.8±0.3 | 61.7 | 41.7 |
| | Semi-Sampling* | 41.9±2.3 | 24.9±1.5 | 48.3±0.9 | 30.2±0.8 | 55.4±1.1 | 37.6±2.0 | 64.5 | 47.5 |
| | Ours | 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 |
| | improvements | 2.1 | 2.1 | 3.6 | 3.9 | 2.2 | 3.8 | 0.8 | 1.2 |
| SUN RGB-D | VoteNet | 29.9 | 10.5 | 38.9 | 17.2 | 45.7 | 22.5 | 58.0 | 33.4 |
| | SESS | – | – | 42.9 | 14.4 | 47.9 | 20.6 | 61.1 | 37.3 |
| | 3DIoUMatch* | 38.7±1.2 | 21.3±1.7 | 46.0±0.5 | 28.9±0.6 | 50.4±0.9 | 30.7±0.5 | 61.8 | 41.9 |
| | Confid-3DIoUMatch* | 39.0±1.7 | 21.8±0.8 | 45.9±0.8 | 28.9±0.3 | 50.4±1.5 | 30.6±1.1 | 60.1 | 40.4 |
| | Semi-Sampling* | 39.8±1.1 | 22.6±0.6 | 48.6±0.7 | 31.1±0.5 | 51.4±1.2 | 33.2±0.4 | 63.2 | 45.7 |
| | Ours | 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 |
| | improvements | 1.7 | 2.4 | 1.3 | 1.4 | 1.9 | 1.8 | 0.7 | 1.2 |

Table 1: Comparison with previous methods on ScanNet and SUN RGB-D datasets. Mark * indicates that the result is reproduced by ourselves under the same data splits with 3DIoUMatch’s.

| ID | Class-Prob. Data Aug. | | Geo.-aware Feat. Matching | | ScanNet 10% | | SUN-RGBD 5% | |
|-----|-----------------------|-------------|---------------------------|--------------|--------------|--------------|--------------|--------------|
| | uniform | class-prob. | feature matching | geo. weights | mAP@0.25 | mAP@0.5 | mAP@0.25 | mAP@0.5 |
| (a) | | | | | 47.53 | 28.63 | 38.72 | 21.31 |
| (b) | ✓ | | | | 48.32 | 30.15 | 39.79 | 22.60 |
| (c) | ✓ | ✓ | | | 50.63 | 31.72 | 40.35 | 23.75 |
| (d) | ✓ | | ✓ | ✓ | 50.82 | 32.60 | 40.79 | 23.90 |
| (e) | ✓ | ✓ | ✓ | ✓ | 51.88 | 34.06 | 41.51 | 24.98 |
| (f) | | | | | 47.53 | 28.63 | 38.72 | 21.31 |
| (g) | | | ✓ | | 48.59 | 29.34 | 39.19 | 21.86 |
| (h) | | | ✓ | ✓ | 49.68 | 31.16 | 39.44 | 22.29 |
| (i) | ✓ | ✓ | ✓ | | 50.89 | 32.37 | 40.79 | 24.27 |
| (j) | ✓ | ✓ | ✓ | ✓ | 51.88 | 34.06 | 41.51 | 24.98 |

Table 2: Ablation Study on ScanNet and SUN-RGBD datasets. Comprehensive ablation experiments are conducted on the proposed DPKE, validating the independent effectiveness of each individual module.

challenge by developing an efficient computation method, which is detailed in the Appendix B. Finally, the geometry-aware feature matching loss can be computed as:

$$\mathcal{L}_f = \mathbb{1}(o^t \geq \tau_{\text{obj}}) \cdot \mathcal{L}_\delta(z^s, z^t) \cdot \mathcal{W}_{\text{geometry}}, \quad (4)$$

where o^t is the objectness score predicted by the student model, $\mathbb{1}(x)$ takes a value of 1 only when the condition x is true; otherwise, it is 0.

Experiments

Evaluation Setup

Datasets and Training Details. We validate our proposed method on the ScanNet (Dai et al. 2017) and SUN RGB-D (Song, Lichtenberg, and Xiao 2015) datasets. Our code is built based on (Wang et al. 2021a), adhering to their optimization and training parameters. We conduct experiments with label ratios of 0.05, 0.1, 0.2, and 1.0, as presented in Table 1. For both the ScanNet and SUN RGB-D datasets, we follow the standard preprocessing (Zhao, Chua, and Lee 2020; Wang et al. 2021a) and evaluation protocols to ensure consistency across different experiments. These protocols involve reporting the mean average precision (mAP)

over three data splits with 3D Intersection over Union (IoU) thresholds of 0.25 and 0.5. Further details on implementation and datasets can be found in the Appendix B.

Methods. We compare our method with five other methods, as displayed in Table 1. Both 3DIoUMatch (Wang et al. 2021a) and Confid-3DIoUMatch (Chen et al. 2023) have open-source code. We do experiments using their code on the same splits and report their performance. Semi-sampling (Wu et al. 2022) does not have a codebase available online. We contact their authors and reproduce their results, removing the extra segmentation annotations on ScanNet.

Main Results

Table 1 presents the comparison results against baselines under different ratios of labeled data. The previous state-of-the-art is semi-sampling (Wu et al. 2022), which verifies that the diversity of input data is the most severe problem for semi-supervised 3D detection. Compared with previous methods, our method achieves 2%-4% increase on ScanNet. Even on the much more difficult dataset SUN RGB-D (Song, Lichtenberg, and Xiao 2015), we still achieve around 2% increase under all settings on SUN RGB-D. We hypothesize that the

| strategy | ScanNet 10% | | SUN-RGBD 5 % | |
|------------|--------------|--------------|--------------|--------------|
| | mAP @0.25 | mAP @0.5 | mAP @0.25 | mAP @0.5 |
| Uniform | 48.32 | 30.15 | 39.79 | 22.60 |
| LLS | 48.07 | 29.32 | 39.42 | 22.39 |
| HLS (ours) | 50.63 | 31.72 | 40.35 | 23.75 |

| weighting | ScanNet 10% | | SUN-RGBD 5 % | |
|------------|--------------|--------------|--------------|--------------|
| | mAP @0.25 | mAP @0.5 | mAP @0.25 | mAP @0.5 |
| Constant | 48.59 | 29.34 | 39.19 | 21.86 |
| HCD | 48.83 | 30.19 | 39.31 | 21.99 |
| LCD (ours) | 49.68 | 31.16 | 39.44 | 22.29 |

(a) Different sampling strategies for Class-Probabilistic Data Augmentation. Here LLS is short for Low-Logit Sampling and HLS is short for High-Logit Sampling.

(b) Different weighting strategies for geometry-aware feature matching. LCD means giving samples with Low Chamfer Distance higher weights. HCD is short for High Chamfer Distance, thus reverse.

Table 3: Further validation of the effectiveness of DPKE.

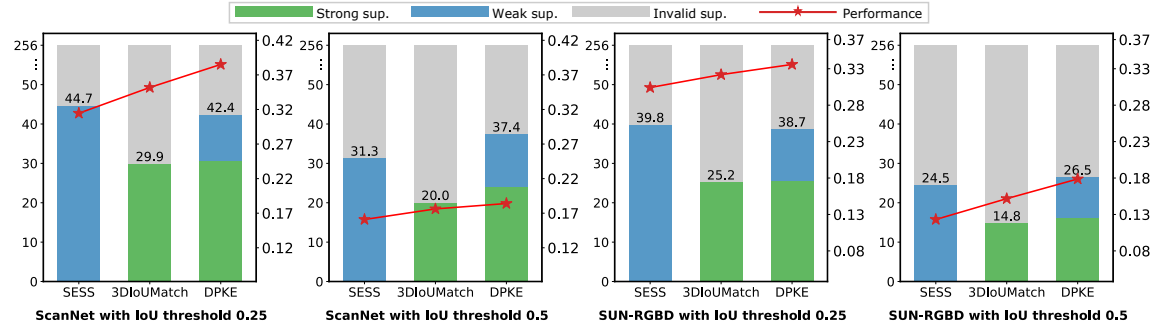


Figure 3: The distribution of samples providing different supervision from the teacher model predictions. The invalid supervision represents the samples failing to match ground truth. The rest part could be divided into samples providing strong (e.g. 3DIoUMatch) and weak supervision (e.g. SESS or our Geometry-aware Feature Matching).

weaker improvement observed on the SUN RGB-D dataset, in comparison to the ScanNet dataset, is attributable to the inferior quality of point clouds and ground truth proposals. Our class-probabilistic data augmentation is impacted on SUN RGB-D, as it depends on ground truth information to generate the proposal bank. Additionally, the geometry-aware feature matching loss is also influenced by the considerable incompleteness present in low-quality point clouds.

Ablation Study

Ablation of class-probabilistic data augmentation. To validate the effectiveness of our proposed class-probabilistic data augmentation, we test a degraded version of our proposed algorithm, which uses uniform sampling. Then compare it to the results obtained when applying our class-probabilistic augmentation strategy. Results are displayed in Table 2 from (a) to (e). (a) shows the performance of the 3DIoUMatch. The results with or without geometry-aware feature matching loss are shown in (b) and (c), or (d) and (e). We also compare our class-probabilistic sampling strategy with different sampling strategies in Table 3 (a). The performance indicates that forcing the model to learn from examples with low logits even harms the mAP under both ScanNet and SUN RGB-D. That is because the categories with lower logits are usually the ones that are difficult to recognize, such as “picture”. Utilizing such categories to augment the current scene does not effectively serve the purpose of increasing diversity.

Ablation of geometry-aware feature matching. Our proposed geometry-aware feature matching loss consists of the feature matching loss and geometry weights for each proposal pair. We present ablation studies for each component under various conditions in Table 2, ranging from (f) to (j). Regardless of whether class-probabilistic data augmentation is applied, each module in our geometry-aware feature matching loss can contribute to improvements independently. Our geometry-aware constraints encourage the model to focus more on proposal pairs with lower Chamfer Distance, denoted as LCD. To validate our method’s efficacy, we compare it to an alternative weighting strategy that assigns higher weight to proposal pairs with higher Chamfer Distance. The “constant” means the weights for all proposal pairs equal to one. As demonstrated in Table 3 (b), our method outperforms HCD and constant weights under all settings.

Pseudo-label convergence. Our DPKE can effectively utilize more supervision signals. As illustrated in Figure 3, we collect statistics on the teacher model’s output during the training process, focusing on supervision signals from samples that can be matched with the ground truth, namely those corresponding to the blue and green portions. The other samples are represented by the gray portion and cannot provide valid supervision to the student model. The red line illustrates the performance of SESS, 3DIoUMatch, and DPKE, respectively. Under various settings, our method is able to

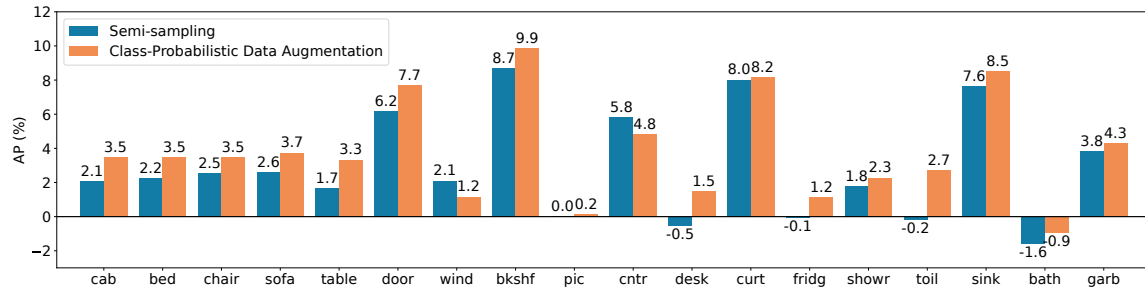


Figure 4: Per class average precision (AP) comparison by applying different data augmentation strategies on 3DIoUMatch. The experiment is conducted on the ScanNet with the label ratio as 5%.

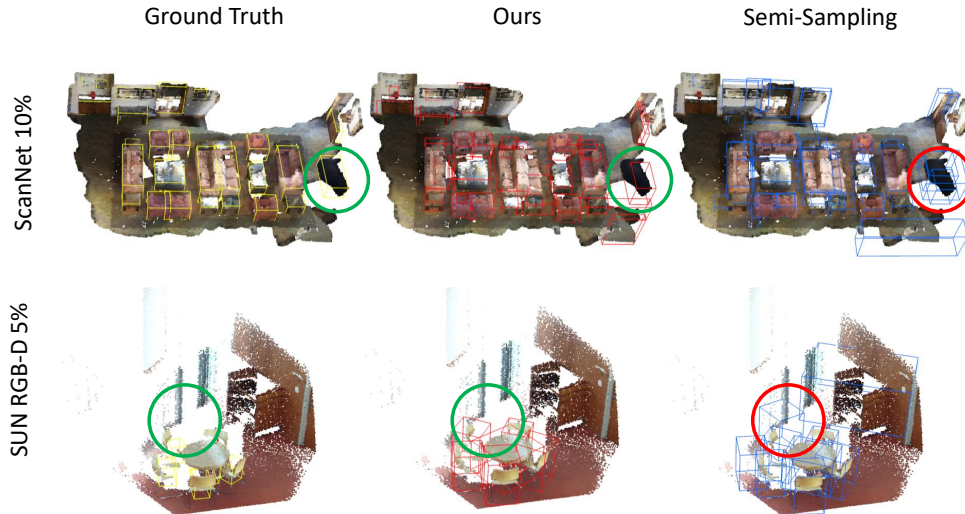


Figure 5: Visualization results on ScanNet and SUN-RGBD. The models are trained on ScanNet 5% and SUN RGB-D 10%, respectively. The areas within the green circles contain the ground truth and correct predictions, while those in the red circles contain wrong predictions.

utilize more valid samples predicted by the teacher model compared to 3DIoUMatch, and is more efficient than SESS.

Class-wise Performance Comparison with Different Augmentations. Figure 4 compares the mAP improvement for each class. Our method outperforms semi-sampling in all categories except "counter" and "window". This might be due to the fact that these two classes face the challenges of limited samples or high learning difficulty.

Visualization

Figure 5 visualizes the qualitative comparison between our proposed method and the existing SOTA semi-sampling (Wu et al. 2022) on a testing scene of ScanNet and SUN RGB-D, respectively. We observe that semi-sampling exhibits a tendency to predict redundant proposals, while our proposed DPKE aligns with ground truth well. This may be attributed to the geometry-aware feature matching, which perceps the geometry properties of objects.

Conclusion

In this study, we present DPKE, an approach that addresses the challenges in 3D semi-supervised object detection from both data and feature perspectives. We develop class-probabilistic data augmentation, which effectively handles the limited diversity in 3D datasets at the data level and minimizes potential inter-class influences resulting from varying learning progress. Furthermore, we introduce the geometry-aware feature matching loss to overcome the low recall of pseudo labels encountered in previous semi-supervised 3D detection techniques. Our geometry-aware feature matching resolves the low recall issue by utilizing feature-level similarity between student and teacher models, thereby enhancing the quantity of supervision. Comprehensive experiments on two benchmark datasets, ScanNet and SUN-RGBD, demonstrate that our proposed DPKE method surpasses existing state-of-the-art techniques under various label ratio conditions, highlighting its capacity to reduce the labor-intensive process of data annotation in cluttered 3D indoor environments.

Acknowledgments

The authors would like to thank the reviewers for their comments that help improve the manuscript. This research work is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-RP-2021-022), the Hyundai research grant (04OIS000257C130), and the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021).

References

- Chen, B.; Li, P.; Chen, X.; Wang, B.; Zhang, L.; and Hua, X.-S. 2022. Dense Learning Based Semi-Supervised Object Detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4805–4814. New Orleans, LA, USA: IEEE. ISBN 978-1-66546-946-3.
- Chen, C.; Debattista, K.; and Han, J. 2022. Semi-supervised object detection via virtual category learning. *arXiv preprint arXiv:2207.03433*.
- Chen, Z.; Jing, L.; Yang, L.; Li, Y.; and Li, B. 2023. Class-Level Confidence Based 3D Semi-Supervised Learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 633–642.
- Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5828–5839.
- Fan, H.; Su, H.; and Guibas, L. J. 2017. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 605–613.
- Jeong, J.; Lee, S.; Kim, J.; and Kwak, N. 2019. Consistency-Based Semi-supervised Learning for Object Detection. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Lee, J.; Walsh, S.; Harakeh, A.; and Waslander, S. L. 2018. Leveraging pre-trained 3d object detection models for fast ground truth generation. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2504–2510. IEEE.
- Li, G.; Li, X.; Wang, Y.; Wu, Y.; Liang, D.; and Zhang, S. 2022. Pseco: Pseudo labeling and consistency training for semi-supervised object detection. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, 457–472. Springer.
- Liu, L.; Zhang, B.; Zhang, J.; Zhang, W.; Gan, Z.; Tian, G.; Zhu, W.; Wang, Y.; and Wang, C. 2023. MixTeacher: Mining Promising Labels with Mixed Scale Teacher for Semi-Supervised Object Detection. *arxiv:2303.09061*.
- Liu, Y.-C.; Ma, C.-Y.; He, Z.; Kuo, C.-W.; Chen, K.; Zhang, P.; Wu, B.; Kira, Z.; and Vajda, P. 2021. Unbiased teacher for semi-supervised object detection. *arXiv preprint arXiv:2102.09480*.
- Liu, Y.-C.; Ma, C.-Y.; and Kira, Z. 2022. Unbiased Teacher v2: Semi-supervised Object Detection for Anchor-free and Anchor-based Detectors. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9809–9818. New Orleans, LA, USA: IEEE. ISBN 978-1-66546-946-3.
- Papadopoulos, D. P.; Uijlings, J. R.; Keller, F.; and Ferrari, V. 2016. We don’t need no bounding-boxes: Training object class detectors using only human verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 854–863.
- Park, J.; Xu, C.; Zhou, Y.; Tomizuka, M.; and Zhan, W. 2022. Detmatch: Two teachers are better than one for joint 2d and 3d semi-supervised object detection. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, 370–389. Springer.
- Qi, C. R.; Litany, O.; He, K.; and Guibas, L. J. 2019. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, 9277–9286.
- Rukhovich, D.; Vorontsova, A.; and Konushin, A. 2023. Tr3d: Towards real-time indoor 3d object detection. *arXiv preprint arXiv:2302.02858*.
- Sohn, K.; Zhang, Z.; Li, C.-L.; Zhang, H.; Lee, C.-Y.; and Pfister, T. 2020. A Simple Semi-Supervised Learning Framework for Object Detection. *arxiv:2005.04757*.
- Song, S.; Lichtenberg, S. P.; and Xiao, J. 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 567–576.
- Tang, Y.; Chen, W.; Luo, Y.; and Zhang, Y. 2021. Humble Teachers Teach Better Students for Semi-Supervised Object Detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3131–3140. Nashville, TN, USA: IEEE. ISBN 978-1-66544-509-2.
- Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.
- Wang, H.; Cong, Y.; Litany, O.; Gao, Y.; and Guibas, L. J. 2021a. 3dioumatch: Leveraging iou prediction for semi-supervised 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14615–14624.
- Wang, H.; Dong, S.; Shi, S.; Li, A.; Li, J.; Li, Z.; Wang, L.; et al. 2022. Cagroup3d: Class-aware grouping for 3d object detection on point clouds. *Advances in Neural Information Processing Systems*, 35: 29975–29988.
- Wang, J.; Gang, H.; Ancha, S.; Chen, Y.-T.; and Held, D. 2021b. Semi-supervised 3D object detection via temporal graph neural networks. In *2021 International Conference on 3D Vision (3DV)*, 413–422. IEEE.
- Wu, X.; Zhao, Y.; Peng, L.; Chen, H.; Huang, X.; Lin, B.; Liu, H.; Cai, D.; and Ouyang, W. 2022. Boosting Semi-Supervised 3D Object Detection with Semi-Sampling. *arXiv preprint arXiv:2211.07084*.

- Xu, M.; Zhang, Z.; Hu, H.; Wang, J.; Wang, L.; Wei, F.; Bai, X.; and Liu, Z. 2021. End-to-End Semi-Supervised Object Detection with Soft Teacher. *arxiv:2106.09018*.
- Xu, Q.; Zhong, Y.; and Neumann, U. 2022. Behind the curtain: Learning occluded shapes for 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2893–2901.
- Yan, Y.; Mao, Y.; and Li, B. 2018. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10): 3337.
- Yang, Y.-Q.; Guo, Y.-X.; Xiong, J.-Y.; Liu, Y.; Pan, H.; Wang, P.-S.; Tong, X.; and Guo, B. 2023. Swin3D: A Pre-trained Transformer Backbone for 3D Indoor Scene Understanding. *arXiv preprint arXiv:2304.06906*.
- Yin, J.; Fang, J.; Zhou, D.; Zhang, L.; Xu, C.-Z.; Shen, J.; and Wang, W. 2022. Semi-supervised 3D object detection with proficient teachers. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, 727–743. Springer.
- Zhang, F.; Pan, T.; and Wang, B. 2021. Semi-Supervised Object Detection with Adaptive Class-Rebalancing Self-Training. *arxiv:2107.05031*.
- Zhang, Y.; Zhang, Q.; Zhu, Z.; Hou, J.; and Yuan, Y. 2022a. GLENet: Boosting 3D Object Detectors with Generative Label Uncertainty Estimation. *arXiv preprint arXiv:2207.02466*.
- Zhang, Z.; Ji, Y.; Cui, W.; Wang, Y.; Li, H.; Zhao, X.; Li, D.; Tang, S.; Yang, M.; Tan, W.; et al. 2022b. ATF-3D: Semi-supervised 3D object detection with adaptive thresholds filtering based on confidence and distance. *IEEE Robotics and Automation Letters*, 7(4): 10573–10580.
- Zhao, N.; Chua, T.-S.; and Lee, G. H. 2020. Sess: Self-ensembling semi-supervised 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11079–11087.
- Zhao, N.; and Lee, G. H. 2022. Static-dynamic co-teaching for class-incremental 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3436–3445.
- Zheng, W.; Tang, W.; Jiang, L.; and Fu, C.-W. 2021. SE-SSD: Self-ensembling single-stage object detector from point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14494–14503.
- Zhou, H.; Ge, Z.; Liu, S.; Mao, W.; Li, Z.; Yu, H.; and Sun, J. 2022. Dense Teacher: Dense Pseudo-Labels for Semi-supervised Object Detection. *arxiv:2207.02541*.