

DQS3D: Densely-matched Quantization-aware Semi-supervised 3D Detection

Huan-ang Gao^{1,2} Beiwen Tian^{1,2} Pengfei Li^{1,2} Hao Zhao¹ Guyue Zhou¹

¹Institute for AI Industry Research (AIR), THU

²Department of Computer Science and Technology, THU

{gha20, tbw18, li-pf22}@emails.tsinghua.edu.cn {zhaohao, zhouguyue}@air.tsinghua.edu.cn

Abstract

In this paper, we study the problem of semi-supervised 3D object detection, which is of great importance considering the high annotation cost for cluttered 3D indoor scenes. We resort to the robust and principled framework of self-teaching, which has triggered notable progress for semi-supervised learning recently. While this paradigm is natural for image-level or pixel-level prediction, adapting it to the detection problem is challenged by the issue of proposal matching. Prior methods are based upon two-stage pipelines, matching heuristically selected proposals generated in the first stage and resulting in spatially sparse training signals. In contrast, we propose the first semi-supervised 3D detection algorithm that works in the single-stage manner and allows spatially dense training signals. A fundamental issue of this new design is the quantization error caused by point-to-voxel discretization, which inevitably leads to misalignment between two transformed views in the voxel domain. To this end, we derive and implement closed-form rules that compensate this misalignment on-the-fly. Our results are significant, e.g., promoting ScanNet mAP@0.5 from 35.2% to 48.5% using 20% annotation. Codes and data are publicly available¹.

1. Introduction

3D object detection (and reconstruction/tracking) [3, 25, 28, 37, 42, 62] is a fundamental problem in 3D scene understanding [17, 26, 51, 57, 59, 63], but its progress still lags behind 2D detection due to a high annotation cost. As such, semi-supervised 3D object detection [48, 56, 60] has recently attracted much attention as it holds the promise to improve accuracy using enormous unlabeled data. These semi-supervised 3D detectors are trained with a widely recognized framework called mean teachers (MT) [44]. While semi-supervised image classification [52] and semantic segmentation [2] using MT boil down to pairing predictions at

Comparison between Dense Matching and Proposal Matching

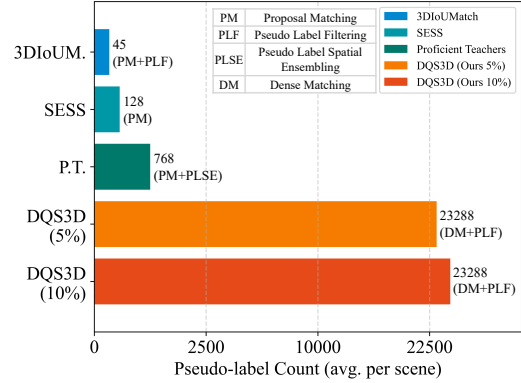


Figure 1: This figure demonstrates the average count of box pairs in representative proposal matching methods SESS [60], 3DIOUMatch [48] and Proficient Teachers [56]. Our dense matching formulation (DQS3D) allows significantly more box pairs and spatially dense training signals. The x-axis is distorted according to a squared root mapping.

the image or pixel level, how to pair predictions between two sets of 3D boxes remains an open question.

This open question is not yet well answered by prior methods [48, 56, 60], as demonstrated by the analysis in Fig. 1. Shown by the upper three bars, they only exploit a very limited number of box pairs for MT training and we attribute this limitation to the two-stage architecture (i.e., VoteNet [37]) they are built upon. VoteNet makes final box predictions using seed proposals extracted by the first stage and only a limited number of proposal pairs are aligned.

Being densely-matched. The emergence of fully convolutional 3D detection [39] inspires us to address the aforementioned issue using densely matched boxes, and it turns out fruitful. As shown by the lower two bars in Fig. 1, our method allows much more box pairs for MT training even after label filtering. This change leads to spatially dense training signals that translate to notable performance improvement (Table. 1). In one word, our method predicts one 3D box for each voxel, getting rid of the intermediate proposal generation stage. Thus pairing teacher and stu-

¹Code: <https://github.com/AIR-DISCOVER/DQS3D>

dent predictions in a voxel-wise manner becomes a natural choice and this directly leads to dense training signals.

Being quantization-aware. During the development of our densely matched paradigm, we identify a fundamental issue specific to 3D detection: point-to-voxel quantization. It is widely known that the power of MT is unleashed only with diverse data augmentation [2, 14, 52] and random transformation is a typical augmentation strategy [10, 18, 45] for 3D point cloud. Unfortunately, applying random transformation inevitably leads to a different point-to-voxel mapping due to the existence of quantization error and a mismatch between teacher and student predictions on each voxel. To this end, we derive a closed-form compensation rule and implement it on-the-fly, which leads to consistent performance gains in various settings.

Highlighting our two technical contributions mentioned above, we name our method **DQS3D**, which is short for densely-matched quantization-aware semi-supervised 3D detection. Our contributions are summarized as follows:

- We shed light on the superiority of dense matching over proposal matching in semi-supervised 3D object detection, which could not only harvest more pseudo labels but also improve the pseudo-label quality.
- We propose the first framework for densely-matched quantization-aware semi-supervised 3D object detection, where we point out the problem of quantization error and come up with an on-the-fly fix to it.
- We conduct extensive experiments on public datasets and achieve significant results. For example, DQS3D scores 48.5% mAP@0.5 on ScanNet using 20% data while the best published result is 35.2%.

2. Related Works

2.1. Self-Training for Semi-supervised Learning

Semi-supervised learning (SSL) is a powerful learning paradigm that improves performance by leveraging both labeled and unlabeled data, making it especially useful in situations where obtaining manually annotated labels is costly or difficult. Recent works strive to apply this paradigm to tasks including semantic segmentation [12, 20], object detection [1, 23, 34], text recognition [36], action recognition [35], facial expression recognition [24], video paragraph grounding [19], *etc.*

In particular, self-training using pseudo-labeling [22, 58] is a principled method that has been widely adopted for SSL [7, 11, 13, 15, 50, 53, 55, 64]. A typical architecture for online self-training is mean teachers (MT) [44], which successfully integrates the self-training method into end-to-end frameworks. MT involves two identical but independent networks during training, with one (referred to as the student network) updated by gradient descent and the other one

(referred to as the teacher network) updated by exponential moving average (EMA) of the student model’s parameters. Predictions of the teacher network on unlabeled data are regarded as online pseudo-labels for the student network, and self-teaching is implemented by enforcing predictions of the two networks to be consistent. The architecture of MT has been proven highly effective on various tasks [1, 23, 29, 30, 40, 43, 47, 54, 55].

2.2. Semi-supervised 3D Object Detection

Proposal Matching for Voting-based Detector. Specifically on the task of semi-supervised 3D object detection, numerous prior arts are also based on the MT architectures and take the voting-based VoteNet [37] as base detectors. SESS [60] introduced the nearest-center matching scheme (which we refer to as *proposal matching*) to generate pseudo-labels from all teacher proposals. 3DIoUMatch [48] proposed a filtering mechanism to impose multiple thresholds on teacher predictions for improving quality of pseudo labels. It further performs non-maximum suppression (NMS) on pseudo-labels to reduce redundancy. Proficient Teachers [56] implemented a spatial-ensembling module that generates detections from multiple augmented views of input point clouds, which are then combined to produce more pseudo-labels. Although these methods have shown promise on the task of semi-supervised 3D object detection, they rely heavily on proposal matching, which we argue to be ineffective as the harvested pseudo training signals are sparse in space.

Dense Prediction Detector. The dense prediction scheme for 2D object detection task has garnered a lot of interest in the research community [21, 27, 38, 46]. However, directly applying the backbones for 2D detection to 3D tasks [32] is not cost-efficient due to the sparse nature of point clouds in space, requiring non-trivially larger amount of computational resources than 2D counterparts.

Nevertheless, the advent of high-dimensional convolutional neural networks [4–6, 16, 61] has reduced both time and space complexity, making it possible to efficiently extract hierarchical features from 3D point clouds. Leveraging sparse 3D convolution, 3D object detection can scale to much larger scenes while remaining memory-efficient [31, 39, 49]. Motivated by this design, FCAF3D [39] uses a voxelized modification of ResNet as the backbone, which enables feature extraction and object prediction on a voxel basis. The voxelization, however, inevitably brings about the issue of quantization error in the point-to-voxel discretization when the input point cloud is randomly augmented. In this paper, we propose *dense matching for dense prediction detector*, identify the problem of quantization error and propose a solution to it on-the-fly.

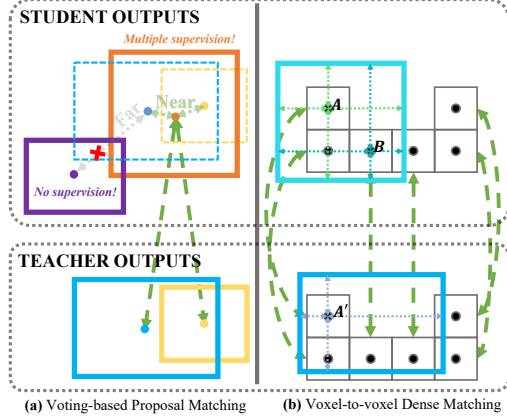


Figure 2: Illustration of two matching schemes. (a) Proposal matching: each teacher prediction is matched with the student prediction whose center is closest to that of the teacher prediction. (b) Dense matching: matching is established through spatially-aligned voxel anchors. Dashed boxes are only demonstrations for spatial locations of the teacher predictions.

3. Methodology

This section presents a detailed exposition of DQS3D. In Sec. 3.1, we formally define the task of semi-supervised 3D object detection. In Sec. 3.2, we introduce dense matching scheme and compare it with prior arts of proposal matching. In Sec. 3.3, we introduce the densely matched self-training framework and the loss design, combined to address the task of semi-supervised 3D object detection. In Sec. 3.4, we point out the problem of quantization error and derive a closed-form solution.

3.1. Preliminary

We formally define the task of 3D object detection as to predict all objects $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^K$ given an input point cloud $\mathbf{X} \in \mathbb{R}^{N \times 3}$, where K denotes the number of objects in the scene and each target object \mathbf{y}_i is represented by its bounding box parameters δ_i and corresponding semantic label q_i . Specifically, in terms of 3d object detection in the semi-supervised setting, only a small proportion of the training dataset (denoted by $\{\mathbf{X}^L\}$) is equipped with ground-truth object bounding-box labels (denoted by $\{\mathbf{Y}^L\}$), whereas the remainder (denoted by $\{\mathbf{X}^U\}$) has no labels.

3.2. Dense Matching

To address the task of semi-supervised 3D object detection, self-training methods (e.g., mean teachers [44, 48, 56, 60]) enforce consistency between the predictions of the student and teacher networks. Thus, it is crucial to establish a mapping between student and teacher predictions in aligned views, which we refer to as *matching*.

Prior arts with self-training adopt *proposal matching* [48, 60] to align the predicted objects (referred to as *proposals*) of the student and teacher networks, which is typically done through a nearest-center strategy. More specifically, each teacher proposal is aligned with the student proposal whose center is the nearest to that of the teacher proposal, as illustrated in Fig. 2(a). Note that, the dashed boxes are only demonstration for spatial locations of teacher outputs.

Despite the fact that teacher proposals are generally more accurate than student proposals, we argue that *proposal matching* is ineffective and may hinder knowledge propagation from the teacher to the student. The ineffectiveness is mainly attributed to the two adverse situations illustrated in Fig. 2(a), inevitably caused by the sparseness of the proposals in space: (1) **Adjacent teacher proposals are aligned to the same student proposal and cause confused supervision for the student.** (2) **Student proposals that are distant from any teacher proposal are aligned to none and receive no supervision from teacher proposals.**

To address the aforementioned issues, a sufficient condition is a bijection between the student and teacher predictions. Inspired by the facts that the objects are predicted on a voxel basis with the dense prediction base detector, and that the voxels anchors of the student and teacher views are inherently corresponded in space, we propose *dense matching* (illustrated in Fig. 2(b)) to establish the bijection, simply by pairing the predictions at corresponding voxel anchors.

We believe that the dense matching scheme has the following advantages: (1) Each predicted object is represented by multiple bounding box predictions whose regression scores varies at different voxel anchors (e.g., points A and B in Fig. 2(b)). This phenomenon imposes spatial regularization on the dense prediction model and improves the models’s awareness of local geometry, as the optimization process forces the predicted bounding box parameters of the same object but at different voxel anchors to be sampled from a smooth function in space. (2) The required bijection between the student and teacher predictions is naturally established with the correspondence of the voxel anchors, with which each student prediction receives supervision from only one teacher prediction. This eliminates the two aforementioned adverse situations, namely the “multiple supervision” and the “no supervision”. In light of the benefits of dense matching over proposal matching, we propose a self-training framework specifically tailored for the dense matching scheme in the upcoming section.

3.3. Densely Matched Self-Training Framework

Overall Architecture. Following prior arts [48, 56, 60], we adapt the robust self-training approach for our densely-matched semi-supervised learning framework, as depicted in Fig. 3. The framework includes two identical but independent networks (teacher and student models) as the

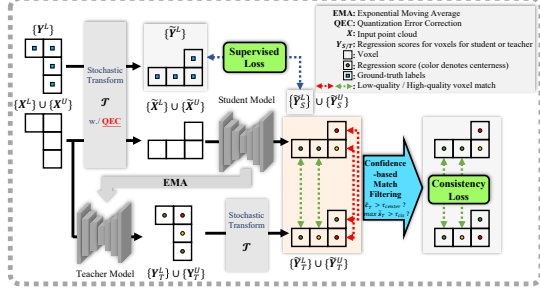


Figure 3: Illustration of our proposed densely-matched quantization-aware semi-supervised 3D object detection framework.

base detectors, which are implemented by FCAF3D [39]. During training, an input batch is composed of both labeled data with ground-truth object annotations and unlabeled data. The input batch is then augmented by asymmetric quantization-aware transformation $\mathbf{R}_{\theta, \Delta \mathbf{r}}$, which consists of a random rotation θ around the upright axis, a random translation $\Delta \mathbf{r}$ and the quantization error correction (detailed later in Sec. 3.4).

The augmented and unaugmented batches are then fed into the student and teacher models, producing voxel-level predictions $\tilde{\mathbf{Y}}_S = \{\tilde{\mathbf{Y}}_S^L, \tilde{\mathbf{Y}}_S^U\}$ and $\mathbf{Y}_T = \{\mathbf{Y}_T^L, \mathbf{Y}_T^U\}$, respectively. The teacher predictions are then aligned to the student predictions with the same transformation $\mathbf{R}_{\theta, \Delta \mathbf{r}}$, resulting in the transformed teacher predictions $\tilde{\mathbf{Y}}_T = \{\tilde{\mathbf{Y}}_T^L, \tilde{\mathbf{Y}}_T^U\}$. With the same transformation, the dense matching between the two sets of predictions is naturally established, which is further filtered to increase the quality of pseudo-labels. The student network is optimized by enforcing a consistency loss on the remaining match set and a supervised loss between the student predictions and the ground-truth labels. In the following sections, we describe three core components of the framework. The colors of the titles are the same as the corresponding regions in Fig. 3.

Aligning Teacher-Student Predictions. Suppose $\mathbf{A} = (\hat{x}, \hat{y}, \hat{z})$ represents the coordinate of the voxel. Following [39], the teacher prediction $\mathbf{y}^{\mathbf{A}} \in \mathbf{Y}_T$ is formulated by the bounding box parameters $\delta^{\mathbf{A}} = \{\delta_i^{\mathbf{A}}\}_{i=1}^8$, the centerness $c^{\mathbf{A}} \in [0, 1]$ and the semantic regression scores $\{p_i^{\mathbf{A}}\}_{i=1}^{N_{\text{cls}}}$, where N_{cls} denotes the number of semantic categories. The first six bounding box parameters $\delta_1, \delta_2, \dots, \delta_6$ represent the distance to the opposite surfaces of the bounding box in the width, length and height dimension, and δ_7, δ_8 utilize the topological equivalency of pair $(\frac{w}{l}, \theta)$ to a Mobius strip for the disambiguation of heading angles of symmetric objects, namely:

$$\begin{aligned} \delta_1^{\mathbf{A}} &= (x + \frac{w}{2}) - \hat{x}, \delta_2^{\mathbf{A}} = \hat{x} - (x - \frac{w}{2}), \delta_3^{\mathbf{A}} = (y + \frac{l}{2}) - \hat{y}, \\ \delta_4^{\mathbf{A}} &= \hat{y} - (y - \frac{l}{2}), \delta_5^{\mathbf{A}} = (z + \frac{h}{2}) - \hat{z}, \delta_6^{\mathbf{A}} = \hat{z} - (z - \frac{h}{2}), \\ \delta_7^{\mathbf{A}} &= \log \frac{w}{l} \sin(2\theta), \delta_8^{\mathbf{A}} = \log \frac{w}{l} \cos(2\theta) \end{aligned} \quad (1)$$

Assuming the transformation $\mathbf{R}_{\theta, \Delta \mathbf{r}}$ maps $\mathbf{y}^{\mathbf{A}}$ to $\tilde{\mathbf{y}}^{\mathbf{A}'} := \tilde{\mathbf{y}}^{\mathbf{A} \mathbf{R}_{\theta, \Delta \mathbf{r}}}$. Since the rotation around the upright-axis and the spatial translation have no effect on the semantics or the relative location towards anchor voxel of the predicted bounding box, we have $\tilde{c}^{\mathbf{A}'} = c^{\mathbf{A}}, \tilde{s}^{\mathbf{A}} = s^{\mathbf{A}'}$. The relationship between $\tilde{\delta}^{\mathbf{A}'}$ and $\delta^{\mathbf{A}}$ can be derived from Eq. 1, which goes:

$$\begin{aligned} \tilde{\delta}_1^{\mathbf{A}'} &= \frac{\cos \theta + 1}{2} \delta_1^{\mathbf{A}} + \frac{-\cos \theta + 1}{2} \delta_2^{\mathbf{A}} + \frac{-\sin \theta}{2} \delta_3^{\mathbf{A}} + \frac{\sin \theta}{2} \delta_4^{\mathbf{A}}, \\ \tilde{\delta}_2^{\mathbf{A}'} &= \frac{-\cos \theta + 1}{2} \delta_1^{\mathbf{A}} + \frac{\cos \theta + 1}{2} \delta_2^{\mathbf{A}} + \frac{\sin \theta}{2} \delta_3^{\mathbf{A}} + \frac{-\sin \theta}{2} \delta_4^{\mathbf{A}}, \\ \tilde{\delta}_3^{\mathbf{A}'} &= \frac{\sin \theta}{2} \delta_1^{\mathbf{A}} + \frac{-\sin \theta}{2} \delta_2^{\mathbf{A}} + \frac{\cos \theta + 1}{2} \delta_3^{\mathbf{A}} + \frac{-\cos \theta + 1}{2} \delta_4^{\mathbf{A}}, \\ \tilde{\delta}_4^{\mathbf{A}'} &= \frac{-\sin \theta}{2} \delta_1^{\mathbf{A}} + \frac{\sin \theta}{2} \delta_2^{\mathbf{A}} + \frac{-\cos \theta + 1}{2} \delta_3^{\mathbf{A}} + \frac{\cos \theta + 1}{2} \delta_4^{\mathbf{A}}, \\ \tilde{\delta}_5^{\mathbf{A}'} &= \delta_5^{\mathbf{A}}, \tilde{\delta}_6^{\mathbf{A}'} = \delta_6^{\mathbf{A}}, \tilde{\delta}_7^{\mathbf{A}'} = \delta_7^{\mathbf{A}} \cos(2\theta), \tilde{\delta}_8^{\mathbf{A}'} = \delta_8^{\mathbf{A}} \cos(2\theta). \end{aligned} \quad (2)$$

The detailed derivation is in the supplementary material.

Matching Filtering. After establishing the matching between the two sets of predictions, a filtering strategy based on confidence is applied to the matching to reduce low-quality supervision. Specifically, with the predicted centerness score and semantic distribution in teacher outputs denoted by \tilde{c}_T and \tilde{s}_T , only matching that satisfies $\tilde{c}_T > \tau_{\text{center}}$ and $\max(\text{softmax}(\tilde{s}_T)) > \tau_{\text{cls}}$ is retained. τ_{center} and τ_{cls} are hyperparameters. Note that, even after the filtering, the matching in our method is still dense.

The primary distinction between the proposed *dense matching* method and prior arts with *proposal matching* is the processing order of the matching and filtering. In proposal matching methods, teacher proposals are first filtered using confidence scores for higher quality and then matched, resulting in even sparser teacher proposals. In the dense matching framework, on the contrary, the matching is established first and then filtered, preserving the spatial alignment of the predictions.

Optimization. The student model is optimized by gradient descent with the supervised loss $\mathcal{L}_{\text{supervised}}$ and the consistency loss $\mathcal{L}_{\text{consistency}}$.

The supervised loss $\mathcal{L}_{\text{supervised}}$ is enforced between the student predictions of the labeled input point clouds $\{\tilde{\mathbf{Y}}_S^L\}$ and the corresponding labels after the augmentation $\{\tilde{\mathbf{Y}}_T^L\}$. Following [39], we adopt 3DIoU loss on the predicted bounding boxes, a binary cross entropy loss on the predicted centerness and a cross entropy loss on the predicted semantic distribution.

The consistency losses $\mathcal{L}_{\text{consistency}}$ are enforced on the filtered matching between student and teacher predictions. For box parameters, we adopt Huber loss, which is less sensitive to outliers in pseudo labels:

$$\mathcal{L}_{\text{box}}^{\mathbf{A}} = \begin{cases} \frac{1}{2}(\Delta \delta_i^{\mathbf{A}})^2, & \text{for } |\Delta \delta_i^{\mathbf{A}}| < \tau_{\text{box}}, \\ \tau_{\text{box}} \cdot (|\Delta \delta_i^{\mathbf{A}}| - \frac{1}{2}\tau_{\text{box}}), & \text{otherwise.} \end{cases} \quad (3)$$

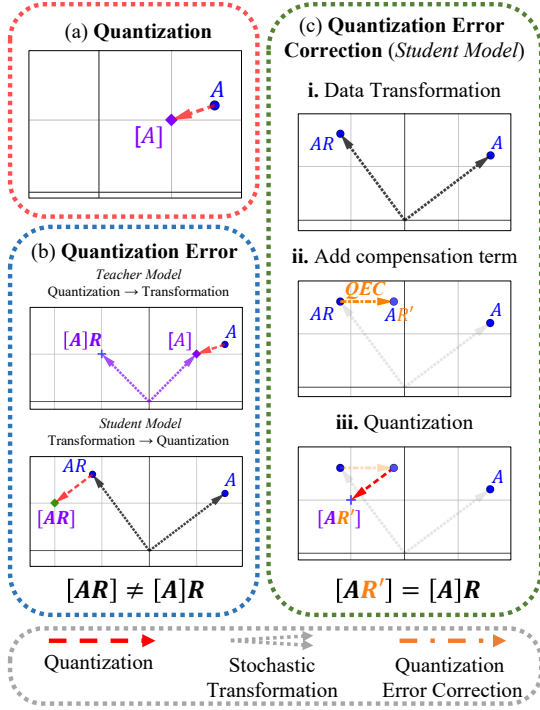


Figure 4: Demonstration of quantization error and correction. R denotes stochastic augmentation and $\lfloor \cdot \rfloor$ denotes quantization. For the purpose of illustration, voxels and transformations are depicted in 2D space. (a) Concept of quantization. (b) Concept of quantization error. Without loss of generality, the random transformation is represented by a 90° counter-clockwise rotation. (c) The process of quantization error correction (QEC). QEC is applied on the student branch between random transformation and quantization to eliminate the quantization error.

where $\Delta\delta^A = \delta^A - \tilde{\delta}^A$ and τ_{box} is a hyperparameter. For centerness we adopt L2 loss $\mathcal{L}_{\text{center}}^A = \|c^A - \tilde{c}^A\|_2^2$. For predicted semantic distribution, we adopt KL divergence $\mathcal{L}_{\text{semantic}}^A = \sum_{c=1}^{N_{\text{cls}}} s_c^A \log(\frac{s_c^A}{\tilde{s}_c^A})$. The final consistency loss is then formulated as:

$$\mathcal{L}_{\text{consistency}} = \lambda_{\text{box}} \mathcal{L}_{\text{box}} + \lambda_{\text{center}} \mathcal{L}_{\text{center}} + \lambda_{\text{semantic}} \mathcal{L}_{\text{semantic}}. \quad (4)$$

where λ_{box} , λ_{center} and $\lambda_{\text{semantic}}$ are loss weights.

As for the teacher model, the gradients are detached and the model parameters are updated using exponential moving average (EMA) of those of the student model:

$$\theta_t^{n+1} = \alpha \theta_t^n + (1 - \alpha) \theta_s^n \quad (5)$$

where θ_t^n and θ_s^n denote the parameters of the teacher and student networks at the n -th step, and α is the average factor. The quality of the guidance provided by the teacher model is gradually improved with the knowledge from the student model.

3.4. Quantization Error Correction

In this section, we shed light on the problem of quantization error and propose a quantization error correction (QEC) module with closed-form solutions to address the problem.

Following implementation in MinkowskiEngine [4], we define the quantization (or voxelization) operator $\lfloor \cdot \rfloor$ on a vector as $\lfloor \mathbf{A} \rfloor = (\lfloor x_{\mathbf{A}} \rfloor, \lfloor y_{\mathbf{A}} \rfloor, \lfloor z_{\mathbf{A}} \rfloor)$, where the notation $\lfloor \cdot \rfloor$ denotes the floor operator. The process of quantization is illustrated in Fig. 4(a). Since the stochastic transformation does not commute with voxelization (depicted in Fig. 4(b)), the spatial location of an input point corresponds to two different ones after being processed by the student and teacher networks ($\lfloor \mathbf{A} \mathbf{R} \rfloor$ and $\lfloor \mathbf{A} \rfloor \mathbf{R}$ in Fig. 4(b)), the difference of which is defined as the *quantization error*.

The quantization error is detrimental to dense pseudo-label self-training scheme, as it violates the exact dense matching we pursue and causes inaccurate training signals and performance decrease. We propose an online solution that finds a compensation term $\mathbf{r}^T(\mathbf{A}, \mathbf{R}_{\theta, \Delta \mathbf{r}})$ for the given location \mathbf{A} and transformation $\mathbf{R}_{\theta, \Delta \mathbf{r}} \in \mathcal{T}$, namely find \mathbf{r}^T that satisfies:

$$\lfloor \mathbf{A} \mathbf{R}_{\theta, \Delta \mathbf{r}} + \mathbf{r}^T \rfloor \stackrel{\text{same voxel}}{=} \lfloor \mathbf{A} \rfloor \mathbf{R}_{\theta, \Delta \mathbf{r}} \quad (6)$$

We rewrite Eq. 6 by applying voxelization to both sides to replace the *same voxel* equality with the arithmetic equality. Since quantization operation holds the property of idempotence, we have:

$$\lfloor \mathbf{A} \mathbf{R}_{\theta, \Delta \mathbf{r}} + \mathbf{r}^T \rfloor = \lfloor \lfloor \mathbf{A} \rfloor \mathbf{R}_{\theta, \Delta \mathbf{r}} \rfloor \quad (7)$$

By refactoring $\mathbf{A} \mathbf{R}_{\theta, \Delta \mathbf{r}}$ into $\mathbf{A} \mathbf{R}_{\theta} + \Delta \mathbf{r}$ and using Lemma.1 and Lemma.2 from the supplementary material, we have:

$$\{\lfloor \mathbf{A} \rfloor \mathbf{R}_{\theta} + \{\Delta \mathbf{r}\} + \mathbf{r}^T\} = \mathbf{0} \quad (8)$$

when $\theta \in \{\frac{k\pi}{2}\}_{k=0}^3$. The operator $\{\cdot\} : \mathbf{x} \mapsto \mathbf{x} - \lfloor \mathbf{x} \rfloor$ is defined as the remainder after quantization. We solve Eq. 26 by interpreting the equation as a requirement for the terms on the left-hand side to lie within the voxel represented by the original point. Therefore, assuming $\gamma \in [0, S_v]^3$ (S_v denotes the voxel size), we derive the compensation term for $\mathbf{r}^T(\mathbf{A}, \mathbf{R}_{\theta, \Delta \mathbf{r}})$ as:

$$\mathbf{r}^T(\gamma) = \gamma - \{\mathbf{A}\} \mathbf{R}_{\theta} - \{\Delta \mathbf{r}\} \quad (9)$$

To alleviate the negative impacts caused by the perturbations to the point cloud structures, we select γ_0 such that:

$$\gamma_0 = \operatorname{argmin}_{\gamma \in [0, S_v]^3} \|\gamma - \{\mathbf{A}\} \mathbf{R}_{\theta} - \{\Delta \mathbf{r}\}\|_2 \quad (10)$$

Finding γ_0 is a typical mathematical optimization problem, and we provide a closed-form solution to Eq. 28 in the supplementary material.

4. Experiments

4.1. Datasets

Following prior arts [48] [60] aiming at semi-supervised 3D object detection, we evaluate our framework on ScanNet v2 [9] and SUN RGB-D [41].

ScanNet v2 [9] is a widely used 3D indoor scene dataset which contains 1512 scans of indoor scenes reconstructed from 2.5 million high-quality RGB-D images. The annotations include per-point instance labels which enable the derivation of axis-aligned object bounding boxes for training and evaluation of 3D object detection methods. The challenge with this dataset in the semi-supervised setting is the limited amount of labeled data. For instance, the 5% labeled setting corresponds to only a few dozen labeled scenes, making it difficult to learn a good detector from labeled data alone.

SUN RGB-D [41] is a widely used benchmark dataset with 10335 indoor scene scans for evaluating scene understanding algorithms, particularly in the context of 3D object detection. Apart from the RGB-D data, the dataset also provides ground-truth 3D bounding box annotations, which enables the evaluation of the task of 3D object detection. The main challenge of this dataset is that the scenes are not axis-aligned. This rotational variability makes it difficult to predict object bounding boxes accurately in the semi-supervised setting, as the model is challenged to recognize objects with any possible orientation after training on limited labeled data.

4.2. Implementation Details

Hyperparameters. We use the same set of hyperparameters for both datasets. As suggested in [39], the voxel size is set to $S_v = 0.01\text{m}$. The confidence thresholds are set to $\tau_{\text{center}} = 0.40$ and $\tau_{\text{cls}} = 0.20$. The threshold for Huber loss is set to $\tau_{\text{box}} = 0.30$. The weights for consistency losses are set to $\lambda_{\text{box}} = 1.00$, $\lambda_{\text{center}} = 0.25$, and $\lambda_{\text{semantic}} = 0.50$. The same warmup strategy as in [60] is adopted for the consistency losses. The average factor α of the exponential moving average is set to 0.999. As for the stochastic transformation strategies, rotation θ around the upright-axis is randomly chosen from $\{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$ and random translation Δr is sampled uniformly from $[-0.5\text{m}, 0.5\text{m}]^3$.

Details during training and evaluation. We use MMDetection3D [8] toolbox to implement our proposed framework. For semi-supervised detection on both ScanNet and SUN RGB-D, our method runs for 12000 training steps which empirically leads to good convergence. During training, we adopt the AdamW optimizer [33] with an initial learning rate of 10^{-3} and a weight decay factor of 10^{-4} , and a scheduler decaying the learning rate by 90% at 67% and 90% of the training process. In the semi-supervised setting, a training batch contains 8 labeled samples and 8 un-

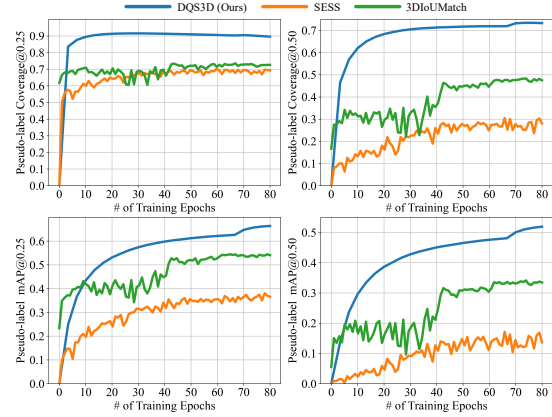


Figure 5: Transductive analysis on different matching schemes in semi-supervised 3D object detection. Experiments are conducted on ScanNet with 10% training data equipped with labels. Coverage@{0.25, 0.50} and mAP@{0.25, 0.50} on the unlabeled set are reported. The proposed dense matching scheme achieves significantly higher performance than prior arts.

labeled samples. During evaluation, to ensure fair comparison with former semi-supervised 3D object detection methods [48, 60], we perform only one forward pass without test-time augmentation used by [39]. Meanwhile, we keep other evaluation settings including IoU thresholds for NMS the same as [39]. We report the mAP@0.25 and mAP@0.50 metrics.

4.3. Comparison of Matching Schemes

In this section, we first provide the comparison between our proposed *dense matching* method and the *proposal matching* methods. We validate the superiority of our proposed methods by demonstrating the quantity and quality of the pseudo labels generated by dense matching strategy.

Are more pseudo-labels harvested? We trained our proposed method as well as former arts on ScanNet [9] dataset with 10% data equipped with labels and collect the pseudo labels harvested by these methods. The average amounts of pseudo-labels harvested from one scene are illustrated in Fig. 1. As shown in the figure, our methods with dense matching strategy harvest a significantly larger amount of pseudo labels compared with SESS [60], 3DIoUMatch [48] and Proficient Teachers [56], which is a contributing factor to better performance for semi-supervised 3D object detection. As shown by later experiments in Sec. 4.4, this translates to the improvement of the detection performance.

Are the pseudo-labels of good quality? In this section, we investigate the quality of pseudo-labels generated by different matching schemes. We borrow the concept of transductive analysis [60] where we regard the model performance on the unlabeled set as the indicating measure of

	Model	5%		10%		20%		100%	
		mAP @0.25	mAP @0.50	mAP @0.25	mAP @0.50	mAP @0.25	mAP @0.50	mAP @0.25	mAP @0.50
ScanNet [9]	VoteNet [37]	27.9	10.8	36.9	18.2	46.9	27.5	57.8	36.0
	FCAF3D [39]	43.8	29.3	51.1	35.7	58.2	42.1	69.5	55.1
	SESS [60]	32.0	14.4	39.5	19.8	49.6	29.0	61.3	39.0
	3DIoUMatch [48]	40.0	22.5	47.2	28.3	52.8	35.2	62.9	42.1
	DQS3D (Ours)	49.2	35.0	57.1	41.8	64.3	48.5	71.9	56.3
	Improv.	+9.2 ↑	+12.5 ↑	+9.9 ↑	+13.5 ↑	+11.5 ↑	+13.3 ↑	+2.4 ↑	+1.2 ↑
SUN-RGBD [41]	VoteNet [37]	29.9	10.5	38.9	17.2	45.7	22.5	58.0	33.4
	FCAF3D [39]	49.5	31.7	50.7	33.4	54.3	36.5	63.6	47.5
	SESS [60]	34.2	13.1	42.1	20.9	47.1	24.5	60.5	38.1
	3DIoUMatch [48]	39.0	21.1	45.5	28.8	49.7	30.9	61.5	41.3
	DQS3D (Ours)	53.2	35.6	55.7	38.2	58.0	42.3	64.1	48.2
	Improv.	+14.2 ↑	+14.5 ↑	+10.2 ↑	+9.4 ↑	+8.3 ↑	+11.4 ↑	+0.5 ↑	+0.7 ↑

Table 1: Experiment results on the task of 3D object detection in various semi-supervised settings (5%, 10%, 20% labels available) and the fully-supervised setting on ScanNet [9] and SUN-RGBD [41] datasets. The proposed DQS3D is compared with **semi-supervised** 3D object detection frameworks SESS [60] and 3DIoUMatch [48], with the margins over 3DIoUMatch [48] marked in blue. Proficient Teachers [56] is currently not comparable as their experiments were only conducted in outdoor scenes and their codes are not currently available, which hinders us to reproduce their experiments on indoor benchmarks. DQS3D is also compared with **fully-supervised** 3D object detectors VoteNet [37] and FCAF3D [39], with the margins over FCAF3D [39] marked in magenta.

the pseudo-label quality. The results are obtained by training our methods and former arts on ScanNet [9] dataset with 10% of data equipped with labels. In Fig. 5, we depict the Coverage@{0.25, 0.50} and mAP@{0.25, 0.50} on the unlabeled set during the training stage, where Coverage indicates the recall rate of objects in the scene. It can be observed that dense matching provides significantly more informative and accurate pseudo-labels, compared with the proposal-based counterparts. We attribute the improved transductive results to the way in which the dense prediction scheme resolves the "multiple supervision" and "no supervision" issues demonstrated in Fig. 2(a) and provides spatially dense training signals for the student network. For more evidence, see Fig. 6.

4.4. Comparisons with prior SOTAs

In this section, we conduct extensive experiments and report the performance of DQS3D and the prior arts in both semi-supervised and fully-supervised settings on the ScanNet and SUN RGB-D datasets. In the semi-supervised setting, the proportion of the labeled set varies among 5%, 10%, and 20%. The consistency losses are imposed on both the labeled and unlabeled sets. In the fully-supervised setting, the entire dataset is regarded as both the labeled and unlabeled sets to examine if the proposed framework can further learn from the additional supervision of pseudo-labels. The experiment results are presented in Tab. 1.

Our method outperforms prior proposal matching meth-

ods by large margins and sets new state-of-the-art results on the semi-supervised 3D object detection benchmark for both ScanNet and SUN RGB-D datasets. It is noteworthy that the improvements of mAP@0.50 are generally larger than those of mAP@0.25. We attribute this to the denseness of the pseudo labels, which provides more spatially fine-grained supervision for the student model. In this way, the predicted object bounding boxes overlap with the target objects to a larger extent, which helps achieve more distinct margins with higher IoU thresholds.

Surprisingly, in the fully-supervised setting, our method also pushes the boundaries of 3D object detection, as shown in Tab. 1. We attribute these improvements to the way in which the framework of self-training serves as regularization and helps improve the stability of the training procedure, as the pseudo-labels generated by the teacher networks are not affected by the bias in mini-batches.

4.5. Ablation Studies

Quantization Error Correction. To demonstrate effectiveness of our proposed quantization error correction (QEC) module (detailed in Sec. 3.4), we conduct experiments on ScanNet (20% of training set equipped with labels), in which we train the proposed framework with various voxel sizes and ablate on the compensation term. In addition to the mAP@{0.25, 0.50} on the *validation* set, we also report the weighted IoU of the predicted and ground-truth bounding boxes on the *labeled training* set. The ob-

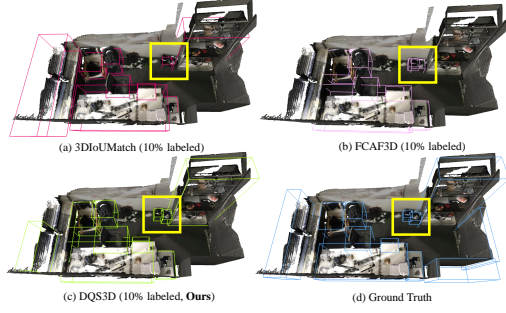


Figure 6: Qualitative results for the 10% labeled setting. Notice the two trash cans against the back wall. As for (a) *3DioUMatch* [48], which employs proposal matching, a *multiple supervision* problem arises, because when adjacent teacher proposals align with the same student proposal, the student receives noisy supervision. However, (c) *our dense matching* successfully resolves this issue.

ject’s weight is determined by the predicted centerness.

The results are presented in Tab. 2. Notably, with all voxel sizes, experiments trained with the compensation term achieve higher performances (up to +2.49% IoU) than those trained without the term. The non-trivial improvements demonstrate the effectiveness of the QEC module in addressing the inherent issue of quantization error and consequently improving the detection accuracy.

Furthermore, we conduct a statistical analysis for intuitively understanding the QEC module. In experiments training on ScanNet dataset, we collect the compensation terms of 80 million points and plot the distribution of the L2 norms and the directions of the terms in Fig. 7. As depicted in Fig. 7(a), the L2 norms of more than 80% of the compensation terms lie in the range of $[0.03S_v, S_v]$ (S_v denotes the voxel size), demonstrating that the quantization error is a non-trivial phenomenon. As depicted in Fig. 7(b), the majority of the compensation terms are aligned with axes. This is because the QEC terms have the smallest possible magnitude by design to preserve the point cloud structure.

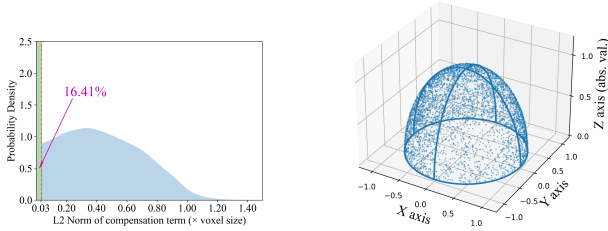


Figure 7: Visualized statistics of the Quantization Error Correction terms. QEC terms are collected from 80M points from ScanNet scenes under training-stage transformations. (a) L2 norm distribution of QEC terms. (b) Directions of QEC terms, note that **the solid blue lines are actually formed by a large amount of crowded data points**.

Consistency Losses. Ablation experiments on the con-

Voxel Size (m)	QEC	IoU (%)	mAP@0.25	mAP@0.50
0.01		75.19	63.7	47.1
	✓	76.82 (+1.63)	64.3 (+0.6)	48.5 (+1.4)
0.02		71.06	59.0	42.2
	✓	72.63 (+1.57)	60.2 (+1.2)	42.8 (+0.6)
0.03		65.63	51.9	35.2
	✓	68.12 (+2.49)	52.7 (+0.8)	35.9 (+0.7)

Table 2: Ablations on the QEC term. Experiments are conducted on ScanNet with 20% training data with labels.

T	Box	Centerness	Class	mAP@0.25	mAP@0.50
✓	✓	✓	✓	64.3 (+4.3)	48.5 (+3.9)
✓		✓	✓	62.3 (-2.0) (+2.3)	45.3 (-3.2) (+0.7)
✓	✓		✓	63.5 (-0.8) (+3.5)	46.4 (-2.1) (+1.8)
✓	✓	✓		63.4 (-0.9) (+3.4)	47.3 (-1.2) (+2.7)
✓	✓			63.0 (-1.3) (+3.0)	46.3 (-2.2) (+1.7)
✓		✓		61.6 (-2.7) (+1.6)	44.9 (-3.6) (+0.3)
✓			✓	62.1 (-2.2) (+2.1)	45.3 (-3.2) (+0.7)
✓				60.0 (-4.3)	44.6 (-3.9)
				58.2 (-6.1) (-1.8)	42.1 (-6.4) (-2.5)

Table 3: Ablations on the consistency losses. Experiments are conducted on ScanNet with 20% training data equipped with labels. T denotes random transformations. **Red margins** are comparisons with DQS3D with all consistency losses. **Blue margins** are comparisons with DQS3D with no consistency losses.

sistency losses are conducted on the ScanNet dataset with 20% training data equipped with labels. The results are reported in Tab. 3. According to the results, the absence of the box consistency loss has the largest influence with the largest performance drops of **-2.0%** (mAP@0.25) and **-3.2%** (mAP@0.50), while the absences of other two consistency losses also bring about drops in performance. These results indicate that each component of the proposed consistency losses is necessary.

5. Conclusion

This paper presents a densely-matched quantization-aware framework, DQS3D, for semi-supervised 3D object detection. By leveraging dense matching instead of proposal matching, and by addressing the issue of quantization error, DQS3D achieves significant improvements over former arts on two widely-used benchmarks, ScanNet v2 and SUN RGB-D, in the semi-supervised setting.

Furthermore, the paper provides evidence that the use of dense predictions leads to more meaningful pseudo-labels and promotes self-training. We hope the insights and techniques introduced in this work would inspire future research in the field of semi-supervised learning.²

²This work is sponsored by DiDi GAIA research program.

References

- [1] Binbin Chen, Weijie Chen, Shicai Yang, Yunyi Xuan, Jie Song, Di Xie, Shiliang Pu, Mingli Song, and Yueting Zhuang. Label matching semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14381–14390, June 2022. 2
- [2] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021. 1, 2
- [3] Xiaoxue Chen, Hao Zhao, Guyue Zhou, and Ya-Qin Zhang. Pq-transformer: Jointly parsing 3d objects and layouts from point clouds. *IEEE Robotics and Automation Letters*, 7(2):2519–2526, 2022. 1
- [4] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019. 2, 5
- [5] Christopher Choy, Junha Lee, Rene Ranftl, Jaesik Park, and Vladlen Koltun. High-dimensional convolutional networks for geometric pattern recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2
- [6] Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8958–8966, 2019. 2
- [7] Ruihang Chu, Xiaoqing Ye, Zhengzhe Liu, Xiao Tan, Xiaojuan Qi, Chi-Wing Fu, and Jiaya Jia. Twist: Two-way inter-label self-training for semi-supervised 3d instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1100–1109, June 2022. 2
- [8] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. <https://github.com/open-mmlab/mmdetection3d>, 2020. 6
- [9] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 6, 7
- [10] Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulencard, Andrea Tagliasacchi, and Leonidas J Guibas. Vector neurons: A general framework for so (3)-equivariant networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12200–12209, 2021. 2
- [11] Jiali Duan, Yen-Liang Lin, Son Tran, Larry S. Davis, and C.-C. Jay Kuo. Slade: A self-training framework for distance metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9644–9653, June 2021. 2
- [12] Jiashuo Fan, Bin Gao, Huan Jin, and Lihui Jiang. Ucc: Uncertainty guided cross-head co-training for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9947–9956, June 2022. 2
- [13] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14009–14018, June 2021. 2
- [14] Huan-ang Gao, Beiwen Tian, Pengfei Li, Xiaoxue Chen, Hao Zhao, Guyue Zhou, Yurong Chen, and Hongbin Zha. From semi-supervised to omni-supervised room layout estimation using point clouds. *arXiv preprint arXiv:2301.13865*, 2023. 2
- [15] Kirill Gavrilyuk, Mihir Jain, Ilia Karmanov, and Cees G. M. Snoek. Motion-augmented self-training for video recognition at smaller scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10429–10438, October 2021. 2
- [16] JunYoung Gwak, Christopher B Choy, and Silvio Savarese. Generative sparse detection networks for 3d single-shot object detection. In *European conference on computer vision*, 2020. 2
- [17] Siyuan Huang, Siyuan Qi, Yixin Zhu, Yinxue Xiao, Yuanlu Xu, and Song-Chun Zhu. Holistic 3d scene parsing and reconstruction from a single rgb image. In *Proceedings of the European conference on computer vision (ECCV)*, pages 187–203, 2018. 1
- [18] Siyuan Huang, Yichen Xie, Song-Chun Zhu, and Yixin Zhu. Spatio-temporal self-supervised representation learning for 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6535–6545, 2021. 2
- [19] Xun Jiang, Xing Xu, Jingran Zhang, Fumin Shen, Zuo Cao, and Heng Tao Shen. Semi-supervised video paragraph grounding with contrastive encoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2466–2475, June 2022. 2
- [20] Donghyeon Kwon and Suha Kwak. Semi-supervised semantic segmentation with error localization network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9957–9967, June 2022. 2
- [21] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018. 2
- [22] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013. 2
- [23] Aoxue Li, Peng Yuan, and Zhenguo Li. Semi-supervised object detection via multi-instance alignment with global class prototypes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9809–9818, June 2022. 2
- [24] Hangyu Li, Nannan Wang, Xi Yang, Xiaoyu Wang, and Xinbo Gao. Towards semi-supervised deep facial expression recognition with an adaptive confidence margin. In *Proceedings of the IEEE/CVF Conference on Computer Vision and*

- Pattern Recognition (CVPR)*, pages 4166–4175, June 2022. 2
- [25] Zizhang Li, Xiaoyang Lyu, Yuanyuan Ding, Mengmeng Wang, Yiyi Liao, and Yong Liu. Rico: Regularizing the unobservable for indoor compositional reconstruction. *arXiv preprint arXiv:2303.08605*, 2023. 1
 - [26] Zhengqin Li, Jia Shi, Sai Bi, Rui Zhu, Kalyan Sunkavalli, Miloš Hašan, Zexiang Xu, Ravi Ramamoorthi, and Manmohan Chandraker. Physically-based editing of indoor scene lighting from a single image. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, pages 555–572. Springer, 2022. 1
 - [27] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016. 2
 - [28] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022. 1
 - [29] Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4258–4267, June 2022. 2
 - [30] Yen-Cheng Liu, Chih-Yao Ma, and Zsolt Kira. Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9819–9828, June 2022. 2
 - [31] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel cnn for efficient 3d deep learning. *Advances in Neural Information Processing Systems*, 32, 2019. 2
 - [32] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2949–2958, 2021. 2
 - [33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
 - [34] Peng Mi, Jiangang Lin, Yiyi Zhou, Yunhang Shen, Gen Luo, Xiaoshuai Sun, Liujuan Cao, Rongrong Fu, Qiang Xu, and Rongrong Ji. Active teacher for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14482–14491, June 2022. 2
 - [35] Chihiro Noguchi and Toshihiro Tanizawa. Ego-vehicle action recognition based on semi-supervised contrastive learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5988–5998, January 2023. 2
 - [36] Gaurav Patel, Jan P. Allebach, and Qiang Qiu. Seq-ups: Sequential uncertainty-aware pseudo-label selection for semi-supervised text recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6180–6190, January 2023. 2
 - [37] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 1, 2, 7
 - [38] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2
 - [39] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Fcaf3d: fully convolutional anchor-free 3d object detection. In *European Conference on Computer Vision*, pages 477–493. Springer, 2022. 1, 2, 4, 6, 7
 - [40] Wonchul Son, Jaemin Na, Junyong Choi, and Wonjun Hwang. Densely guided knowledge distillation using multiple teacher assistants. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9395–9404, October 2021. 2
 - [41] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 6, 7
 - [42] Shuran Song and Jianxiong Xiao. Sliding shapes for 3d object detection in depth images. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VI 13*, pages 634–651. Springer, 2014. 1
 - [43] Yihe Tang, Weifeng Chen, Yijun Luo, and Yuting Zhang. Humble teachers teach better students for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3132–3141, June 2021. 2
 - [44] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 1, 2, 3
 - [45] Beiwen Tian, Liyi Luo, Hao Zhao, and Guyue Zhou. Vibus: Data-efficient 3d scene parsing with viewpoint bottleneck and uncertainty-spectrum modeling. *ISPRS Journal of Photogrammetry and Remote Sensing*, 194:302–318, 2022. 2
 - [46] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 2
 - [47] Hardik Uppal, Alireza Sepas-Moghaddam, Michael Greenspan, and Ali Etemad. Teacher-student adversarial depth hallucination to improve face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3671–3680, October 2021. 2
 - [48] He Wang, Yezhen Cong, Or Litany, Yue Gao, and Leonidas 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*, pages 14615–14624, 2021. 1, 2, 3, 6, 7, 8

- [49] Haiyang Wang, Lihe Ding, Shaocong Dong, Shaoshuai Shi, Aoxue Li, Jianan Li, Zhenguo Li, and Liwei Wang. CA-Group3d: Class-aware grouping for 3d object detection on point clouds. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 2
- [50] Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10857–10866, June 2021. 2
- [51] Xin Wu, Hao Zhao, Shunkai Li, Yingdian Cao, and Hongbin Zha. Sc-wls: Towards interpretable feed-forward camera re-localization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pages 585–601. Springer, 2022. 1
- [52] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020. 1, 2
- [53] Bo Xiong, Haoqi Fan, Kristen Grauman, and Christoph Feichtenhofer. Multiview pseudo-labeling for semi-supervised learning from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7209–7219, October 2021. 2
- [54] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3060–3069, October 2021. 2
- [55] Qize Yang, Xihan Wei, Biao Wang, Xian-Sheng Hua, and Lei 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 (CVPR)*, pages 5941–5950, June 2021. 2
- [56] Junbo Yin, Jin Fang, Dingfu Zhou, Liangjun Zhang, Cheng-Zhong Xu, Jianbing Shen, and Wenguan Wang. Semi-supervised 3d object detection with proficient teachers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, pages 727–743. Springer, 2022. 1, 2, 3, 6, 7
- [57] Hao Zhao, Ming Lu, Anbang Yao, Yiwen Guo, Yurong Chen, and Li Zhang. Physics inspired optimization on semantic transfer features: An alternative method for room layout estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 10–18, 2017. 1
- [58] Hao Zhao, Ming Lu, Anbang Yao, Yiwen Guo, Yurong Chen, and Li Zhang. Pointly-supervised scene parsing with uncertainty mixture. *Computer Vision and Image Understanding*, 200:103040, 2020. 2
- [59] Hao Zhao, Rene Ranftl, Yurong Chen, and Hongbin Zha. Transferable end-to-end room layout estimation via implicit encoding. *arXiv preprint arXiv:2112.11340*, 2021. 1
- [60] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Sess: Self-ensembling semi-supervised 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11079–11087, 2020. 1, 2, 3, 6, 7
- [61] Tianchen Zhao, Niansong Zhang, Xuefei Ning, He Wang, Li Yi, and Yu Wang. Codedvtr: Codebook-based sparse voxel transformer with geometric guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1435–1444, June 2022. 2
- [62] Leisheng Zhong, Yu Zhang, Hao Zhao, An Chang, Wenhao Xiang, Shunli Zhang, and Li Zhang. Seeing through the occluders: Robust monocular 6-dof object pose tracking via model-guided video object segmentation. *IEEE Robotics and Automation Letters*, 5(4):5159–5166, 2020. 1
- [63] Chuhan Zou, Ruiqi Guo, Zhizhong Li, and Derek Hoiem. Complete 3d scene parsing from an rgb-d image. *International Journal of Computer Vision*, 127:143–162, 2019. 1
- [64] Longkun Zou, Hui Tang, Ke Chen, and Kui Jia. Geometry-aware self-training for unsupervised domain adaptation on object point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6403–6412, October 2021. 2

A. Proof of Equations

Lemma 1 If all elements in \mathbf{A} are integers, then the following equation holds:

$$[\mathbf{A} + \mathbf{B}] = \mathbf{A} + [\mathbf{B}] \quad (11)$$

Proof: By definition. \square

Lemma 2 If all elements in \mathbf{A} are integers and $\theta \in \{\frac{k\pi}{2}\}_{k=0}^3$, then all elements in \mathbf{AR}_θ are integers.

Proof: By considering the rotation matrix $\mathbf{R}_{\frac{k\pi}{2}}$ when $k = 0, 1, 2, 3$. Note that $x' = x \cos \theta + y \sin \theta$, $y' = -x \sin \theta + y \cos \theta$ and $z' = z$. When $k = 0, 1, 2, 3$, $\sin \theta$ and $\cos \theta$ produces integer values. According to the property of integer fields, x', y' and z' are also integers, which means all elements in \mathbf{AR}_θ are integers. \square

Proof of Equation 2 Here we prove:

$$\begin{aligned} \tilde{\delta}_1^{\mathbf{A}'} &= \frac{\cos \theta + 1}{2} \delta_1^{\mathbf{A}} + \frac{-\cos \theta + 1}{2} \delta_2^{\mathbf{A}} + \frac{-\sin \theta}{2} \delta_3^{\mathbf{A}} + \frac{\sin \theta}{2} \delta_4^{\mathbf{A}}, \\ \tilde{\delta}_2^{\mathbf{A}'} &= \frac{-\cos \theta + 1}{2} \delta_1^{\mathbf{A}} + \frac{\cos \theta + 1}{2} \delta_2^{\mathbf{A}} + \frac{\sin \theta}{2} \delta_3^{\mathbf{A}} + \frac{-\sin \theta}{2} \delta_4^{\mathbf{A}}, \\ \tilde{\delta}_3^{\mathbf{A}'} &= \frac{\sin \theta}{2} \delta_1^{\mathbf{A}} + \frac{-\sin \theta}{2} \delta_2^{\mathbf{A}} + \frac{\cos \theta + 1}{2} \delta_3^{\mathbf{A}} + \frac{-\cos \theta + 1}{2} \delta_4^{\mathbf{A}}, \\ \tilde{\delta}_4^{\mathbf{A}'} &= \frac{-\sin \theta}{2} \delta_1^{\mathbf{A}} + \frac{\sin \theta}{2} \delta_2^{\mathbf{A}} + \frac{-\cos \theta + 1}{2} \delta_3^{\mathbf{A}} + \frac{\cos \theta + 1}{2} \delta_4^{\mathbf{A}}, \\ \tilde{\delta}_5^{\mathbf{A}'} &= \delta_5^{\mathbf{A}}, \tilde{\delta}_6^{\mathbf{A}'} = \delta_6^{\mathbf{A}}, \tilde{\delta}_7^{\mathbf{A}'} = \delta_7^{\mathbf{A}} \cos(2\theta), \tilde{\delta}_8^{\mathbf{A}'} = \delta_8^{\mathbf{A}} \cos(2\theta). \end{aligned} \quad (12)$$

Proof: Assume the bounding box \mathbf{y} is centered at $\mathbf{c} \in \mathbb{R}^{3 \times 1}$ with dimension $\mathbf{d} \in \mathbb{R}^{3 \times 1}$ and yaw $\phi \in \mathbb{R}$. Since the spatial translation does not affect the relative position of voxels and bounding boxes, here we can only consider the effect of random rotation around the upright-axis θ . Since we have:

$$\begin{bmatrix} \tilde{\delta}_1^{\mathbf{A}'} \\ \tilde{\delta}_2^{\mathbf{A}'} \\ \tilde{\delta}_3^{\mathbf{A}'} \\ \tilde{\delta}_4^{\mathbf{A}'} \end{bmatrix} = \begin{bmatrix} \tilde{x} - \hat{x} \\ \tilde{y} - \hat{y} \end{bmatrix} + \begin{bmatrix} \frac{1}{2}w \\ \frac{1}{2}h \end{bmatrix} = \begin{bmatrix} x - \hat{x} \\ \hat{x} - x \\ y - \hat{y} \\ \hat{y} - y \end{bmatrix} + \begin{bmatrix} \frac{1}{2}w \\ \frac{1}{2}h \\ \frac{1}{2}h \\ \frac{1}{2}w \end{bmatrix} \quad (13)$$

By noting that,

$$\begin{aligned} \begin{bmatrix} x - \hat{x} \\ y - \hat{y} \end{bmatrix} &= \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} x - \hat{x} \\ y - \hat{y} \end{bmatrix} \\ &= \frac{1}{2} \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} \delta_1 - \delta_2 \\ \delta_3 - \delta_4 \end{bmatrix} \\ &= \frac{1}{2} \begin{bmatrix} \cos \theta & -\cos \theta & -\sin \theta & \sin \theta \\ \sin \theta & -\sin \theta & \cos \theta & -\cos \theta \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \end{bmatrix} \end{aligned} \quad (14)$$

And that,

$$\begin{bmatrix} w \\ h \end{bmatrix} = \begin{bmatrix} \delta_1 + \delta_2 \\ \delta_3 + \delta_4 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \end{bmatrix} \quad (15)$$

Then we have,

$$\begin{aligned} \tilde{\delta}_1^{\mathbf{A}'} &= \frac{\cos \theta + 1}{2} \delta_1^{\mathbf{A}} + \frac{-\cos \theta + 1}{2} \delta_2^{\mathbf{A}} + \frac{-\sin \theta}{2} \delta_3^{\mathbf{A}} + \frac{\sin \theta}{2} \delta_4^{\mathbf{A}}, \\ \tilde{\delta}_2^{\mathbf{A}'} &= \frac{-\cos \theta + 1}{2} \delta_1^{\mathbf{A}} + \frac{\cos \theta + 1}{2} \delta_2^{\mathbf{A}} + \frac{\sin \theta}{2} \delta_3^{\mathbf{A}} + \frac{-\sin \theta}{2} \delta_4^{\mathbf{A}}, \\ \tilde{\delta}_3^{\mathbf{A}'} &= \frac{\sin \theta}{2} \delta_1^{\mathbf{A}} + \frac{-\sin \theta}{2} \delta_2^{\mathbf{A}} + \frac{\cos \theta + 1}{2} \delta_3^{\mathbf{A}} + \frac{-\cos \theta + 1}{2} \delta_4^{\mathbf{A}}, \\ \tilde{\delta}_4^{\mathbf{A}'} &= \frac{-\sin \theta}{2} \delta_1^{\mathbf{A}} + \frac{\sin \theta}{2} \delta_2^{\mathbf{A}} + \frac{-\cos \theta + 1}{2} \delta_3^{\mathbf{A}} + \frac{\cos \theta + 1}{2} \delta_4^{\mathbf{A}}. \end{aligned} \quad (16)$$

The rotation around the upright-axis does not affect z -coordinates, so it is trivial that,

$$\tilde{\delta}_5^{\mathbf{A}'} = \delta_5^{\mathbf{A}}, \tilde{\delta}_6^{\mathbf{A}'} = \delta_6^{\mathbf{A}}. \quad (17)$$

The rotation does transform the yaw angle from ϕ to $\phi - \theta$, hence we have:

$$\begin{aligned} \tilde{\delta}_7^{\mathbf{A}'} &= \log\left(\frac{w}{l}\right) \sin(2\phi - 2\theta) \\ &= \log\left(\frac{w}{l}\right) (\sin(2\phi) \cos(2\theta) - \sin(2\theta) \cos(2\phi)) \\ \tilde{\delta}_8^{\mathbf{A}'} &= \log\left(\frac{w}{l}\right) \cos(2\phi - 2\theta) \\ &= \log\left(\frac{w}{l}\right) (\cos(2\phi) \cos(2\theta) + \sin(2\theta) \sin(2\phi)) \end{aligned} \quad (18)$$

By noting that when $\theta \in \{\frac{k\pi}{2}\}_{k=0}^3$, $\sin(2\theta) \equiv 0$. That produces,

$$\begin{aligned} \tilde{\delta}_7^{\mathbf{A}'} &= \log\left(\frac{w}{l}\right) \sin(2\phi) \cos(2\theta) = \delta_7^{\mathbf{A}} \cos(2\theta) \\ \tilde{\delta}_8^{\mathbf{A}'} &= \log\left(\frac{w}{l}\right) \cos(2\phi) \cos(2\theta) = \delta_8^{\mathbf{A}} \cos(2\theta) \end{aligned} \quad (19)$$

Eq. 16, Eq. 17 and Eq. 19 can be combined to form Eq. 12. \square

Proof of Equation 8 Here we prove:

$$[\{\mathbf{A}\}\mathbf{R}_\theta + \{\Delta\mathbf{r}\} + \vec{\mathbf{r}}] = \mathbf{0} \quad (20)$$

We start from Eq. 7 from the main paper:

$$[\mathbf{AR}_{\theta, \Delta\mathbf{r}} + \vec{\mathbf{r}}] = [[\mathbf{A}]\mathbf{R}_{\theta, \Delta\mathbf{r}}] \quad (21)$$

By defactoring $\mathbf{AR}_{\theta, \Delta\mathbf{r}}$ into $\mathbf{AR}_\theta + \Delta\mathbf{r}$, we have:

$$[\mathbf{AR}_\theta + \Delta\mathbf{r} + \vec{\mathbf{r}}] = [[\mathbf{A}]\mathbf{R}_\theta + \Delta\mathbf{r}] \quad (22)$$

Noting all elements in $[\mathbf{A}]$ are integers, hence by assuming $\theta \in \{\frac{k\pi}{2}\}_{k=0}^3$ and applying Lemma. 2, all elements in $[\mathbf{A}]\mathbf{R}_\theta$ are also integers. Then by Lemma. 1, we have:

$$[\mathbf{AR}_\theta + \Delta\mathbf{r} + \vec{\mathbf{r}}] = [\mathbf{A}]\mathbf{R}_\theta + [\Delta\mathbf{r}] \quad (23)$$

Leveraging the property that $\mathbf{X} = [\mathbf{X}] + \{\mathbf{X}\}$, we have:

$$[(\mathbf{A}] + \{\mathbf{A}\})\mathbf{R}_\theta + [\Delta\mathbf{r}] + \{\Delta\mathbf{r}\} + \vec{\mathbf{r}}'] = [\mathbf{A}]\mathbf{R}_\theta + [\Delta\mathbf{r}] \quad (24)$$

A simple deformation of this equation yields:

$$[(\mathbf{A}]\mathbf{R}_\theta + [\Delta\mathbf{r}] + \{\mathbf{A}\}\mathbf{R}_\theta + \{\Delta\mathbf{r}\} + \vec{\mathbf{r}}'] = [\mathbf{A}]\mathbf{R}_\theta + [\Delta\mathbf{r}] \quad (25)$$

By Lemma. 1, we move the term $[\mathbf{A}]\mathbf{R}_\theta + [\Delta\mathbf{r}]$ out of the left-hand side, and that yields:

$$[\{\mathbf{A}\}\mathbf{R}_\theta + \{\Delta\mathbf{r}\} + \vec{\mathbf{r}}'] = \mathbf{0} \quad (26)$$

That is the exact form as Eq. 8 in the original paper. \square

Solution to Equation 10 Here we find the solution γ_0 of:

$$\gamma_0 = \operatorname{argmin}_{\gamma \in [0, S_v]^3} \|\gamma - \{\mathbf{A}\}\mathbf{R}_\theta - \{\Delta\mathbf{r}\}\|_2 \quad (27)$$

We start by considering cases for unary functions. We find the solution ϕ_0 of:

$$\phi_0 = \operatorname{argmin}_{\phi \in [a, b]} \|\phi - M\|_2 \quad (28)$$

The solution is straight-forward. It denotes the closest value in $[a, b]$ to a fixed value M . We represent the solution to this problem as:

$$\operatorname{clamp}(M, a, b) = \begin{cases} a, & M < a, \\ M, & a \leq M < b, \\ b, & M \geq b. \end{cases} \quad (29)$$

Since in the target function of this problem, the three axes are uncorrelated, we can break this problem to a problem set of three problems each equivalent to Eq. 10. We can extend the clamping function to a vector version, namely for any $0 \leq i < \operatorname{len}(\mathbf{M})$:

$$\operatorname{clamp}(\mathbf{M}, \mathbf{a}, \mathbf{b})_i = \operatorname{clamp}(\mathbf{M}_i, \mathbf{a}_i, \mathbf{b}_i) \quad (30)$$

Then the closed-form solution of γ_0 can be formulated as:

$$\gamma_0 = \operatorname{clamp}(\mathbf{M} = \{\mathbf{A}\}\mathbf{R}_\theta + \{\Delta\mathbf{r}\}, (0, 0, 0), (S_v, S_v, S_v)) \quad (31)$$

That is the solution to the original problem. \square

B. Hyperparameter Study

τ_{center} and τ_{cls} . We conducted a hyperparameter study (Fig. 8) on τ_{center} and τ_{cls} . These two hyperparameters are utilized to filter the initially matched set and provide matching pairs that offer less noisy supervision. Finding the optimal values involves a trade-off, as setting the values too low introduces noisy supervision, while setting them too high reduces the number of matched pairs.

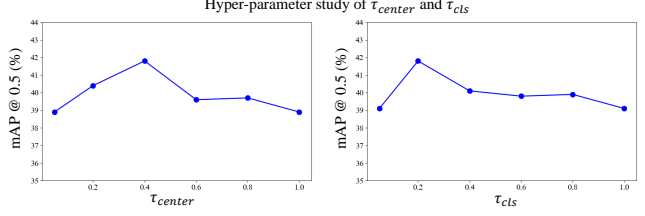


Figure 8: Hyper-parameter Study on τ_{center} and τ_{cls} .

Backbone (Semi-supervised Setting)	mAP@0.25	mAP@0.50
FCAF3D (baseline)	58.2	42.1
FCAF3D (+ Sparse Proposal Matching)	62.0	44.2
FCAF3D (+ Dense Matching, <i>ours</i> DQS3D)	64.3	48.5
TR3D (baseline)	62.5	46.8
TR3D (+ Dense Matching)	65.4	49.9

Table 4: Comparison of *Dense Matching* and *Proposal Matching* Strategies with *Different Backbones* on ScanNet Dataset (20% Labeled). *Proposal matching* involves filtering teacher proposals and matching them with the nearest-center student predictions, while dense matching establishes matching based on spatially-aligned voxel anchors and then applies filtering. In dense matching, the proposed **Quantization Error Correction** module is enabled.

Different Backbones. We conducted experiments (Tab. 4) that show the superiority of dense matching over proposal matching. We argue that the success is originate from addressing issues like *no supervision* and *multiple supervision* problems, which we also qualitatively illustrate in Fig. 6. Note that dense matching is applicable only to recent SOTA voxel-based detectors, not common two-stage proposal-based detectors based on Transformer or heatmaps. Hence we used TR3D (Rukhovich et al.), with the hyperparameters reported in our manuscript without further tuning. Remarkably, we observed an improvement of +3.1% on mAP@0.50.

C. Further Discussion

Computational Complexity Analysis. We utilized the NVIDIA GeForce RTX 2080Ti. Training employed 4 GPUs (2 labeled and 2 unlabeled scenes per GPU card, occupying approximately 7.5GB per GPU) and took around 7 hours to converge. In terms of inference speed, our system achieves 10.3 scenes per second on a single 2080Ti.

Limitation Analysis. The trade-off between memory and voxel size hampers our 3D detectors' performance in outdoor scenes, which is a common limitation in the family of sparse convolutional detectors.