Real-Time 3D Object Detection with Inference-Aligned Learning

Chenyu Zhao[1], Xianwei Zheng[1*], Zimin Xia[2], Linwei Yue[3], Nan Xue[4]

[1]The State Key Lab. LIESMARS, Wuhan University
[2]École polytechnique fédérale de Lausanne (EPFL)
[3]School of Geography and Information Engineering, China University of Geosciences
[4]Ant Group
{cyzhao, zhengxw}@whu.edu.cn, zimin.xia@epfl.ch, yuelw@cug.edu.cn, xuenan@ieee.org

Abstract

Real-time 3D object detection from point clouds is essential for dynamic scene understanding in applications such as augmented reality, robotics, and navigation. We introduce a novel Spatial-prioritized and Rank-aware 3D object detection (SR3D) framework for indoor point clouds, to bridge the gap between how detectors are trained and how they are evaluated. This gap stems from the lack of spatial reliability and ranking awareness during training, which conflicts with the ranking-based prediction selection used at inference. Such a training-inference gap hampers the model's ability to learn representations aligned with inference-time behavior. To address the limitation, SR3D consists of two components tailored to the spatial nature of point clouds during training: a novel spatial-prioritized optimal transport assignment that dynamically emphasizes well-located and spatially reliable samples, and a rank-aware adaptive self-distillation scheme that adaptively injects ranking perception via a self-distillation paradigm. Extensive experiments on ScanNet V2 and SUN RGB-D show that SR3D effectively bridges the training-inference gap and significantly outperforms prior methods in accuracy while maintaining real-time speed.

Code — https://github.com/zhaocy-ai/sr3d

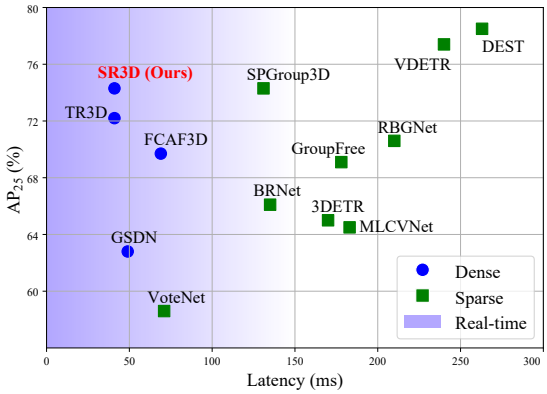


Figure 1: AP_{25} vs. latency on the ScanNet V2 validation set. Our proposed SR3D achieves accurate and fast detection from indoor point clouds. Latency is measured on a single RTX 4090 GPU. The metrics AP_{25} is mean Average Precision under the IoU threshold of 0.25.

Introduction

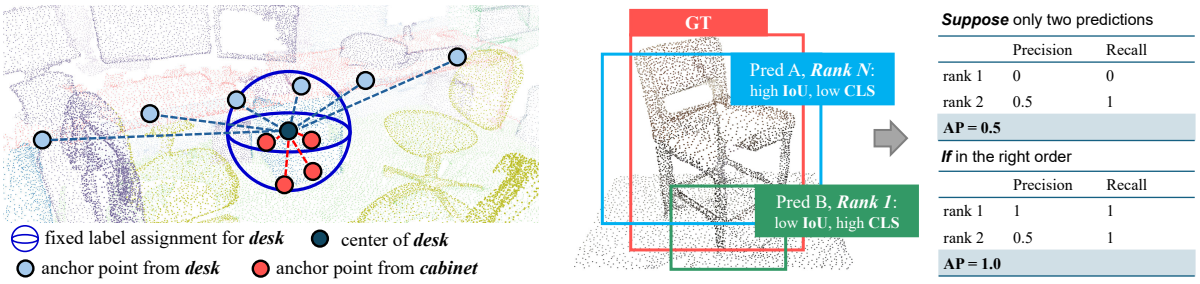
With the increasing availability of 3D sensing technologies, understanding 3D point clouds has become a crucial task in computer vision. We are interested in *3D object detection for point clouds of indoor scenes*, aiming to localize 3D bounding boxes and determine their semantic classes in real time. Robust and real-time 3D object detection is vital for holistic and dynamic scene understanding, enabling critical applications in augmented reality, embodied robotics, and navigation. In such scenarios, the perception system must process input point clouds and make decisions within tight time constraints to ensure responsiveness and safety.

Current 3D detectors are broadly divided into sparse and dense detection paradigms based on their proposal generation mechanisms. Sparse detectors (Qi et al. 2019; Misra, Girdhar, and Joulin 2021) rely on refining sparse proposals, which incur high memory costs and limit scalability, making them less suitable for real-time applications. As shown in Fig. 1, dense detectors (Gwak, Choy, and Savarese 2020; Rukhovich, Vorontsova, and Konushin 2022) are more efficient, as they utilize anchors densely to cover objects and predict bounding boxes with semantic labels in a single pass. Thus, we adopt dense detection frameworks as they are better aligned with the real-time requirements.

However, a fundamental problem in current dense 3D object detectors lies in their inability to align training supervision with inference behavior: *predictions are selected and optimized without considering either their spatial reliability or relative ranking* during training, as shown in Fig. 2. This misalignment ultimately hinders the detector's ability to learn robust and discriminative representations.

To address the issue, we propose a simple but effective framework SR3D tailored for real-time 3D object detection that explicitly integrates the awareness of spatial reliability and ranking into the supervision process. First, we observe that detectors often fail to focus on the most informative samples, especially when dealing with occluded or geometrically ambiguous objects. This is largely due to the use

*Corresponding author.



(a) Spatial reliability absence (b) Ranking awareness absence

Figure 2: Illustration of core limitations in current dense 3D detectors. (a) Fixed heuristic label assignment misidentifies high-quality anchors for the *desk*, being misled by spatial clutter in the indoor scene. (b) Rank-agnostic supervision leads to incorrect ranking of *chair* predictions, degrading performance under Average Precision (AP) evaluation. We use 2D boxes for simplicity.

of fixed heuristic label assignment strategies, such as center priors and IoU thresholds, which ignore the actual spatial reliability of anchors during training. These strategies thereby overlook important and various geometric cues inherent in 3D objects and often mislead the detector to suboptimal solutions. Thus, we introduce the Spatial-Prioritized Optimal Transport Assignment (SPOTA), a formulation that casts label assignment as an optimal transport (OT) problem, where anchor-ground truth pairs are matched dynamically based on spatial-prioritized cost metrics. Unlike other dynamic assignment strategies (Zhang et al. 2019b; Ge et al. 2021a), SPOTA prioritizes geometric cues by introducing normalized vertex distances and a spatial-prioritized strategy. These designs shift focus from semantic scores to spatial alignment, which is particularly crucial in 3D scenes where geometry dominates object representation. As a result, SPOTA establishes more stable optimization and better captures the spatial dependencies unique to 3D indoor scenes.

Then we find that dense detectors treat all positive samples equally, regardless of their relative rankings of localization accuracy or semantic reliability. This ranking absence prevents the model from learning to highlight good-quality predictions during training, and causes inconsistency with the rank-sensitive evaluation metric Average Precision (AP). To resolve this challenge, we propose the Rank-aware Adaptive Self-Distillation (RAS) scheme that explicitly incorporates ranking information into the training process. RAS guides the classifier with localization-aware soft targets, while adaptively blended with classification loss according to confidence ranking. This scheme penalizes overconfident but poorly localized predictions, thereby promoting inference-aligned learning.

The main contributions of this paper are as follows:

- We introduce SR3D, a highly efficient detection framework for indoor 3D object detection to mitigate the training-inference gap. Extensive experiments on ScanNet V2 and SUN RGB-D demonstrate that SR3D enhances the performance of dense detectors, while preserving real-time inference speed.
- We present a novel Spatial-Prioritized OTA strategy, which incorporates more comprehensive geometric information and dynamically assigns labels by their qualities in a global view.
- We design a novel Rank-aware Adaptive Self-Distillation scheme to adaptively integrate ranking awareness into the training process by self-knowledge distillation and improve compatibility with evaluation metrics.

Related Work

We here review the works about 3D object detection, dynamic label assignment and self-knowledge distillation.

3D Object Detection. We focus on indoor 3D object detection and exclude outdoor methods, which differ significantly. Prior works are broadly categorized into sparse and dense detection paradigms. Sparse detection methods aim to generate a limited set of high-quality proposals for object localization by the deep hough voting mechanism (Qi et al. 2019; Xie et al. 2020; Gupta et al. 2022; Wang et al. 2022b,a; Zhu et al. 2024) or using query matching as in DETR-based methods (Misra, Girdhar, and Joulin 2021; Liu et al. 2021; Shen et al. 2024; Wang et al. 2025). In contrast, dense detection methods typically tile anchors across the spatial domain to enable dense prediction in a single shot. These works (Gwak, Choy, and Savarese 2020; Rukhovich, Vorontsova, and Konushin 2022) often inherit designs from 2D frameworks (Tian et al. 2019), and thus carry over limitations from 2D detection. However, these limitations become more pronounced under the modality variations and the geometric complexity of sparse and irregular point clouds.

Dynamic Label Assignment. Label assignment, which is fundamental to 2D and 3D object detection, significantly influences the optimization of a detector, especially for dense detectors. FreeAnchor (Zhang et al. 2019b) first identified the best samples based on the customized likelihood by classification scores and IoUs. Some other works (Ke et al. 2020; Li et al. 2020a; Zhu et al. 2020; Kim and Lee 2020) were proposed to select training samples by the joint criteria of the classification and the regression scores. Alternatively, OTA (Ge et al. 2021a) and simOTA (Ge et al. 2021b) handle the assignment process as an optimal transportation problem to minimize the transportation costs. DLLA (Liu et al. 2025)

utilized learnable feature embedding and similarity matching to find the best assignment. However, these methods fail to address the systematic training-inference gap stemming from the ranking awareness absence.

Self-Knowledge Distillation. Self-knowledge distillation enhances the effectiveness of training a student network by leveraging its own knowledge without an external teacher network (Furlanello et al. 2018). Some approaches (Zhang et al. 2019a; Lan, Zhu, and Gong 2018) introduce auxiliary networks to facilitate this process, while others (Xu and Liu 2019; Yun et al. 2020) adopt contrastive learning schemes to refine the internal representation learning. Later, self-knowledge distillation has been successfully applied to various tasks such as classification (Zhang, Bao, and Ma 2021), semantic segmentation (An et al. 2022), and object detection (Zhang et al. 2022). In contrast to conventional self-distillation frameworks, our approach focuses on embedding ranking awareness into the supervision process, thereby establishing a strong interaction between classification and regression branches without the need for additional modules.

Methodology

In 3D object detection on point clouds, we are given a point cloud $\mathcal{S}_i$ of an indoor scene with the coordinates $\{(x, y, z)\}$ and the colors $\{(r, g, b)\}$ to produce a set of bounding boxes $\{\mathbf{b}_k\}_i$ with semantic labels $\{l_k\}_i$ to cover all objects.

In the following sections, we first revisit the motivation behind our approach, and then detail the two core components: Spatial-Prioritized Optimal Transport Assignment (SPOTA) and Rank-aware Adaptive Self-Distillation (RAS).

Motivation

A key challenge in dense 3D object detection is the training-inference gap, the misalignment between how detectors are trained and how they are evaluated. We identify two primary gaps. (1) *Missing spatial reliability*: During training, sample selection often relies on ad-hoc heuristic rules (*e.g.*, instance scales) or prior knowledge (*e.g.*, center prior). These fixed methods fail to reflect the actual spatial quality of anchors, especially in cluttered indoor scenes. (2) *Missing ranking awareness*: While evaluation metrics like AP are inherently rank-sensitive, standard training pipelines apply uniform, rank-agnostic supervision across all positive samples. Without ranking cues, the detectors struggle to align classification confidence with true localization accuracy, leading to suboptimal results in evaluation.

To overcome these gaps, we propose SR3D, a Spatial-prioritized and Rank-aware 3D detector. SR3D tackles missing spatial reliability through Spatial-Prioritized Optimal Transport Assignment (SPOTA) and addresses missing ranking awareness via Rank-aware Adaptive Self-Distillation (RAS) during training, effectively bridging the training-inference gap in an inference cost-free manner.

The overall pipeline of SR3D is illustrated in Fig. 3. The input point clouds are processed by a sparse convolutional backbone (Choy et al. 2019) with an FPN (Lin et al. 2017a), followed by two task-specific heads that generate dense predictions. During training, SR3D first assigns ground-truth labels to anchors using the proposed SPOTA, and supervises the positives with the RAS scheme. At inference time, Non-Maximum Suppression (NMS) is applied to remove redundant low-confidence detections.

Preliminaries

The Optimal Transport Assignment (OTA) (Ge et al. 2021a) is a representative and widely used label assignment method that formulates the assigning procedure as an optimal transport problem, a special form of Linear Programming (LP) in Optimization Theory. In such a perspective, it views each ground truth t_i as a supplier who holds s_i units of positive labels, and each proposal a_j as a demander who needs d_j units of label. Thus the goal of this problem is to find a transportation plan $\pi^* = \{\pi_{ij}\}$, according to which all goods from suppliers can be transported to demanders at a minimal transportation cost:

$$\begin{aligned} \min_{\pi} \quad & \sum_i \sum_j C_{ij}\pi_{ij}. \\ \text{s.t.} \quad & \sum_i \pi_{ij} = d_j, \sum_j \pi_{ij} = s_i, \\ & \sum_i s_i = \sum_j d_j, \pi_{ij} \geq 0. \end{aligned} \tag{1}$$

The transportation cost from ground truth t_i to proposal a_j is defined as the weighted summation of their classification and regression losses in OTA:

$$C = \lambda \cdot \mathcal{C}_{cls} + \mathcal{C}_{reg}, \tag{2}$$

where $\mathcal{C}_{cls}$ and $\mathcal{C}_{reg}$ stand for cross-entropy loss and IoU loss generally, and λ is the balancing coefficient.

Spatial-Prioritized OTA

While OTA provides a theoretical framework introducing the awareness of prediction reliability to alleviate the missing spatial reliability, its direct application to 3D detectors suffers from several challenges: (1) Compared to 2D object detection, 3D object detection relies more heavily on geometric information, as point cloud coordinates rather than color serve as the primary input modality. (2) It leads to a multi-objective conflict, where predictions with either high IoU but low classification score, or vice versa, are equally likely to be selected as positives.

To overcome the limitations mentioned above, we propose a Spatial-Prioritized Optimal Transport Assignment (SPOTA). Compared to OTA, our proposed SPOTA integrates more geometric cues to adapt to the characteristics of 3D object detection and mitigates the multi-objective conflict by prioritizing localization quality, leading to a more coherent and stable optimization path across tasks.

First, we introduce the normalized vertex distance, a more precise and shape-sensitive measurement of spatial reliability, as shown in Fig. 4. While IoU may assign similar scores to geometrically distinct predictions (Rezatofighi et al. 2019; Zheng et al. 2020, 2022), the proposed normalized vertex distance captures fine-grained differences (*e.g.*, scale and shape) in box vertex alignment, which is defined as:

$$\mathcal{R}_{VD} = \frac{d(\hat{\mathbf{v}}_1, \mathbf{v}_1) + d(\hat{\mathbf{v}}_2, \mathbf{v}_2)}{2\rho(\hat{\mathbf{b}}, \mathbf{b})}, \tag{3}$$

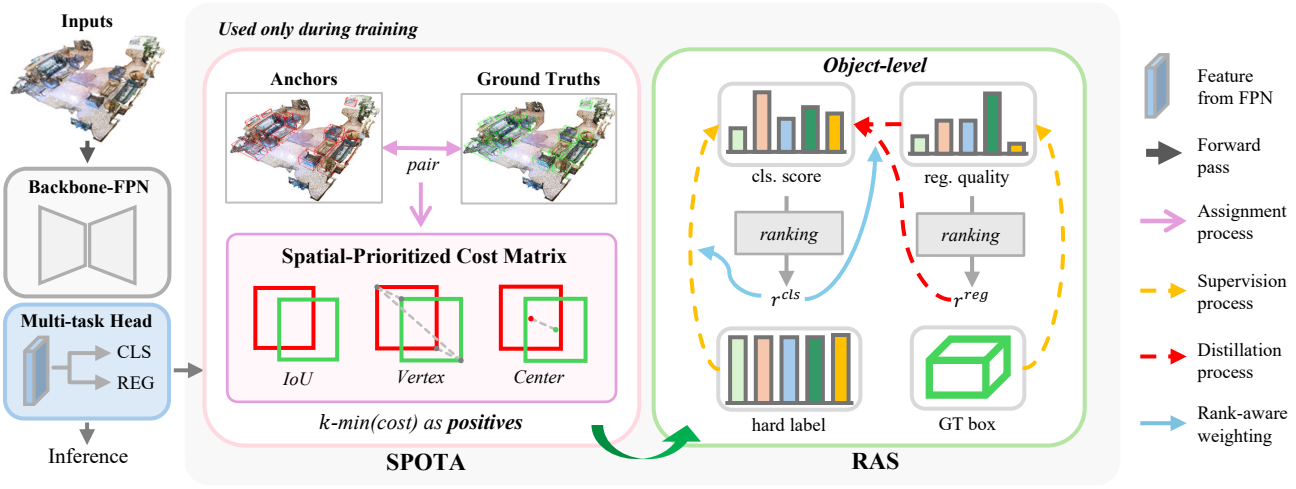


Figure 3: The overall framework of our Spatial-prioritized and Rank-aware network for indoor 3D object detection (SR3D). The spatial-prioritized OTA (SPOTA) and rank-aware adaptive self-distillation (RAS) scheme are employed only during training. SPOTA dynamically assigns positive labels to those truly informative and high-reliability anchors by leveraging geometry hints from prediction–ground truth pairs, such as the IoU and normalized vertex distances. RAS introduces ranking perception into training via a self-distillation mechanism and adaptively reweights the supervision based on relative ranking signals.

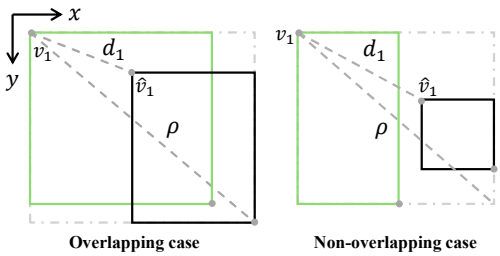


Figure 4: A simplified illustration of normalized vertex distances $\mathcal{R}_{VD}$. Dashed box indicates the smallest enclosing box. We use 2D boxes for simplicity.

where $d(\cdot)$ is the Euclidean distance, $\hat{\mathbf{v}}_1$, $\hat{\mathbf{v}}_2$ denote the vertices of the predicted box, $\mathbf{v}_1$, $\mathbf{v}_2$ are the corresponding ground truth vertices, $\rho(\hat{\mathbf{b}}, \mathbf{b})$ is the diagonal length of the smallest enclosing box covering the two boxes. This allows SPOTA to better distinguish predictions with similar IoUs but varying spatial structures, thereby avoiding ambiguous assignments and enabling more discriminative label assignment and consistent optimization, especially in cluttered indoor environments.

Then, to mitigate the multi-objective conflict during assignment, we design a spatial-prioritized strategy for the assigning process, which is driven solely by geometric cues that better reflect the spatial nature of 3D object detection. While classification remains essential for the recognition of 3D indoor scenes, its feature preferences and output behaviors fundamentally diverge from regression objectives, often leading to conflicting optimization signals. Instead of merely reducing the classification cost weight by λ, we choose to completely remove it from the assignment cost. This decision stems from the observation that, in 3D detection of point clouds, semantic cues are inherently encoded in geometric structures, *e.g.*, object shapes, edges, and layouts (Fan et al. 2024; Mei et al. 2024). Retaining an explicit classification term would thus introduce redundancy and potentially bias the assignment toward semantic patterns rather than robust geometric alignment.

Additionally, to force detectors to focus on potential positive areas and then help stabilize the training process, especially in the early stage, we impose a center prior term:

$$\gamma_c = 1 - \exp\left(-\mu d^2(\mathbf{c}, \mathbf{c}^{gt})\right), \tag{4}$$

where $\mathbf{c}$ and $\mathbf{c}^{gt}$ are the centers of the anchor and the corresponding ground truth box, respectively. Consequently, the final cost matrix is formulated as:

$$C = \gamma_c \cdot (\mathcal{C}_{reg} + \mathcal{R}_{VD}), \tag{5}$$

where $\mathcal{C}_{reg}$ is the regression loss, such as IoU Loss (Yu et al. 2016). Then, following Ge et al. (2021b), we select the top k predictions with the least cost as positive samples for each ground truth, while the rest are negative samples.

Rank-aware Adaptive Self-Distillation

To resolve the issue of missing ranking awareness, we design a unified Rank-aware Adaptive Self-Distillation (RAS) paradigm, as illustrated in Fig. 3, which injects localization and ranking cues from the model itself during training to align supervision with inference behavior. It consists of a self-distillation loss and an adaptive weighting strategy, both guided by the relative ranking of predictions.

We first propose a self-distillation loss to guide the classification branch with localization-aware soft targets derived from the model's own regression branch. Specifically, for each ground truth, we compute the localization accuracy q

(*i.e.*, IoU) and the corresponding soft rank r^{reg} for its positives, where higher r^{reg} indicates better localization. The rank-aware self-distillation loss is then defined as:

$$\mathbf{RDL}(\sigma) = (1 - r^{reg})^{\beta} q \log(\sigma) + q(1-q)\log(1-\sigma), \quad (6)$$

where σ is the classification confidence, and β controls the strength of rank-based modulation. This formulation penalizes poorly localized samples more heavily, effectively suppressing unreliable positives and encouraging the model to assign higher confidence to well-localized predictions.

Then, we need to balance the contribution of the standard classification loss and the distillation loss to stabilize training. A fixed coefficient offers a simple way to balance the two components, but it lacks flexibility across diverse training stages and sample conditions, and fails to exploit the relative ranking crucial for optimizing AP. To address this, we propose a rank-aware adaptive weighting mechanism that dynamically adjusts the contribution of each loss component based on the model's own confidence ranking. Concretely, we blend the Focal Loss (Lin et al. 2017b) and rank-aware self-distillation loss as the final classification loss:

$$\mathcal{L}_{cls} = \sum_{i \in \mathcal{P}} ((1 - r_i^{cls})\mathbf{FL}_i + r_i^{cls}\mathbf{RDL}_i) + \sum_{j \in \mathcal{N}} \mathbf{FL}_j, \quad (7)$$

where r_{cls} is the soft relative ranking of classification scores, with higher values indicating higher scores. $\mathcal{P}$ and $\mathcal{N}$ denote the sets of positive and negative anchors, respectively.

This mechanism aims to identify and rectify predictions where the model exhibits high classification confidence that is inconsistent with poor localization accuracy. By assigning stronger distillation supervision to these potentially overconfident yet poorly localized predictions, the model is encouraged to recalibrate its internal consistency between classification confidence and localization accuracy.

Loss Function

The overall loss function $\mathcal{L}_{det}$ is formulated as:

$$\mathcal{L}_{det} = \mathcal{L}_{cls} + \mathcal{L}_{reg}, \quad (8)$$

where $\mathcal{L}_{cls}$ is defined in Eq. 7, and $\mathcal{L}_{reg}$ is DIoU Loss.

Experiments

Datasets and Evaluation Metrics

We use two challenging 3D indoor scene datasets, ScanNet V2 (Dai et al. 2017) and SUN RGB-D (Song, Lichtenberg, and Xiao 2015), with the data splits from (Qi et al. 2019).

ScanNet V2 is a richly annotated dataset that provides a comprehensive collection of 1,513 reconstructed 3D indoor scans, with per-point annotated 3D indoor scenes and bounding boxes for the 18 object categories. The dataset is divided into 1,201 training samples and 312 for validation.

SUN RGB-D is a widely recognized dataset designed for 3D scene understanding in indoor environments. This dataset is divided into approximately 5,285 training and 5,050 validation point clouds. We follow (Qi et al. 2019) to evaluate our approach on the 10 classes of objects.

For both datasets, We follow the evaluation protocol used in 3D object detection from (Qi et al. 2019; Liu et al. 2021).

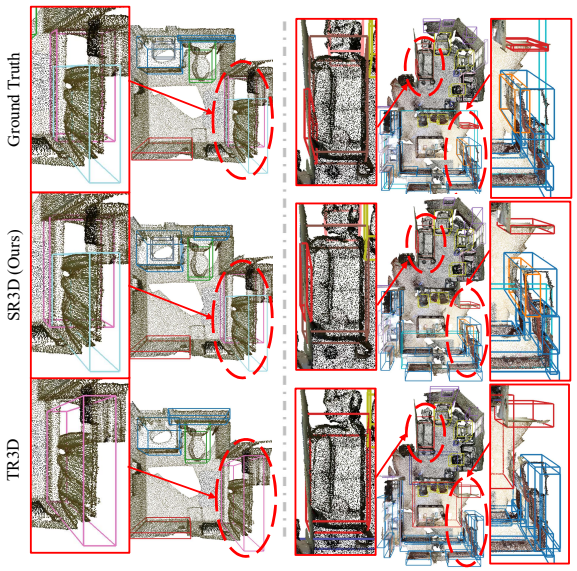


Figure 5: Qualitative results on validation set of ScanNet V2. We only visualize the most confident predictions. As compared to TR3D, our SR3D enables robust detection of more challenging objects in cluttered scenes. Different classes are indicated by bounding boxes in different colors.

Specifically, we train the model five times and test each trained model five times, yielding a total of 25 evaluation runs. We report both the best and average values of mean Average Precision over the 25 runs, under different IoU thresholds of 0.25 (AP_{25}) and 0.5 (AP_{50}).

Compared with State-of-the-arts

We evaluate our SR3D with the recent state-of-the-art dense 3D detectors on ScanNet V2 (Dai et al. 2017) and SUN RGB-D (Song, Lichtenberg, and Xiao 2015) benchmarks. As indicated in Tab. 1, SR3D outperforms the previous state-of-the-art dense detectors in all metrics, whether measured by the highest performance or the average results over multiple trials, and virtually has no effect on latency. In terms of AP_{25}, our method achieves 1.1 and 1.0 improvements over the previous state-of-the-art methods on ScanNet V2 and SUN RGB-D, respectively. While our method improves the highest AP_{50} score by only 0.3 and 0.5 on ScanNet V2 and SUN RGB-D, it achieves substantial average gains of 1.1 and 0.9, suggesting that SR3D delivers more stable and reliable performance. SR3D and DLLA show comparable accuracy. However, DLLA suffers from higher computational overhead due to its auxiliary branch and learnable parameters. Our systemic inference-aligned framework is superior by achieving this performance efficiently through spatial and rank awareness, without adding any learnable components.

The visualization of 3D Object detection with predicted bounding boxes on ScanNet V2 is shown in Fig. 5. To highlight the improvements, we compare only the most confident predictions from TR3D and our SR3D, where SR3D clearly produces more accurate predictions than TR3D.

Methods	ScanNet V2			SUN RGB-D		
	AP_{25} ↑	AP_{50} ↑	Latency ↓	AP_{25} ↑	AP_{50} ↑	Latency ↓
Sparse detectors:						
VoteNet (Qi et al. 2019)	58.6	33.5	71ms	57.7	-	41ms
MLCVNet (Xie et al. 2020)	64.5	41.4	183ms	59.8	-	-
3DETR (Misra et al. 2021)	65.0	47.0	170ms	59.1	32.7	-
BRNet (Gupta et al. 2022)	66.1	50.9	135ms	61.1	43.7	-
GroupFree (Liu et al. 2021)	69.1 (68.6)	52.8 (51.8)	178ms	63.0 (62.6)	45.2 (44.4)	-
RBGNet (Wang et al. 2022b)	70.6 (69.6)	55.2 (54.7)	210ms	64.1 (63.6)	47.2 (46.3)	-
CAGroup3D (Wang et al. 2022a)	75.1 (74.5)	61.3 (60.3)	472ms	66.8 (66.4)	50.2 (49.5)	-
SPGroup3D (Zhu et al. 2024)	74.3 (73.5)	59.6 (58.3)	131ms	65.4 (64.8)	47.1 (46.4)	-
V-DETR (Shen et al. 2024)	77.4 (76.8)	65.0 (64.5)	240ms	67.5 (66.8)	50.4 (49.7)	-
DEST (Wang et al. 2025)	**78.5** (78.3)	**66.6** (66.2)	263ms	**68.4** (67.4)	**51.8** (50.9)	-
Dense detectors:						
GSDN (Gwak et al. 2020)	62.8	34.8	49ms	-	-	-
FCAF3D (Rukhovich et al. 2022)	71.5 (70.7)	57.3 (56.0)	64ms	64.2 (63.8)	48.9 (48.2)	56ms
↳ + DLLA (Liu et al. 2025)	71.4 (71.0)	60.0 (59.0)	97ms	63.8 (63.4)	48.3 (47.4)	-
TR3D (Rukhovich et al. 2023)	72.9 (72.0)	59.3 (57.4)	**42ms**	67.1 (66.3)	50.4 (49.6)	**36ms**
↳ + DLLA (Liu et al. 2025)	73.8 (72.8)	**60.2** (58.9)	-	67.3 (67.0)	50.6 (50.5)	-
SR3D (Ours)	**74.0** (73.2)	59.7 (58.5)	**42ms**	**68.1** (67.2)	**50.9** (50.5)	**36ms**

Table 1: Results of ours and recent indoor 3D object detection methods on the validation set of ScanNet V2 and SUN RGB-D datasets. The main comparison is based on the best results of multiple experiments between different methods, and the average value of 25 trials is given in brackets. For fair comparision, we focus on dense 3D detectors and measure the latency.

Analysis of Inference Alignment

To evaluate whether SR3D effectively achieves inference-aligned learning, we conduct a detailed analysis of the consistency through the following three aspects.

First, we introduce the Average Inconsistency Coefficient (AIC), defined as $AIC = |\mathcal{P}|^{-1} \sum_{i \in \mathcal{P}} |p_i - q_i|$, which measures the L_1 disparity between the classification score (p_i) and the localization quality (q_i). A smaller AIC indicates better consistency. We visualize the AIC curves during training in Fig. 6(a), and observe that SR3D consistently exhibits significantly lower AIC than TR3D, demonstrating its improved inference alignment.

Then, we examine whether this improved alignment translates into better inference behavior. As shown in Fig. 6(b), we visualize the top-30 high-confidence predictions of each class after NMS, plotting their classification confidence against the corresponding IoU with ground truth.

The distribution of predictions from SR3D aligns more closely with the ideal confidence–accuracy diagonal, indicating better inference consistency than TR3D. Furthermore, we compute the absolute errors between the classification scores and IoUs of these predictions, called Prediction Consistency Error (PCE), as shown in Fig. 6(c). Clearly, SR3D achieves lower prediction errors, further validating its ability to maintain consistent and reliable confidence outputs aligned with true localization accuracy.

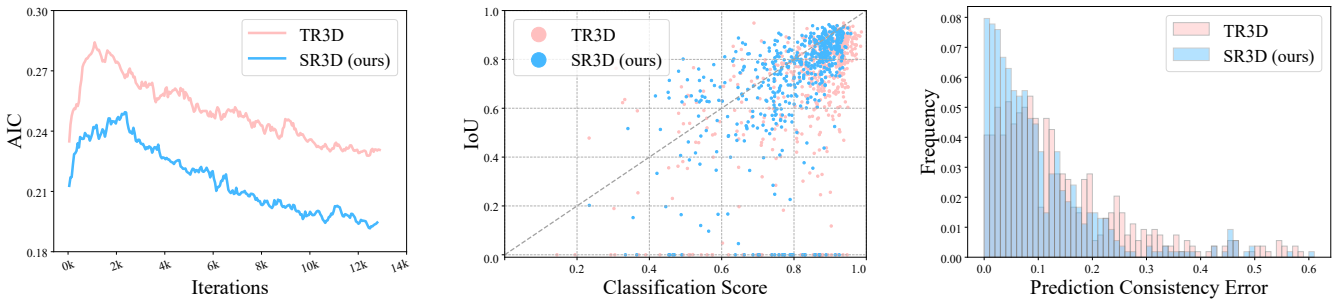
These experiments collectively confirm that SR3D successfully bridges the training-inference gap in dense 3D object detection, validating the effectiveness of our inference-aligned learning designs.

Ablation Study

We conduct detailed ablation studies on the validation set of ScanNet V2, reporting ***the average performance*** to robustly analyze the contribution of each proposed component.

Effect of different components of SR3D. We systematically verify the effectiveness of each component in our proposed methods on the basic fully sparse convolutional dense detector used in (Rukhovich et al. 2023), as shown in Tab. 2. We first introduce each component individually, and the average results of all components surpass the baseline, demonstrating independent efficacy of SR3D. Finally, the complete model achieves 73.2 AP_{25}, establishing a 2.4 absolute gain over the baseline while maintaining real-time inference speed (42ms). The detailed ablation study shows that our proposed modules consistently improve performance across various combinations, demonstrating the robustness and effectiveness of SR3D.

Effect of SPOTA. In SPOTA, we introduce several tailored designs to overcome the limitations of the standard OTA scheme. As shown in Tab. 3, we conduct ablation studies to demonstrate the reasonableness of these designs. First, reintroducing the classification loss into the cost matrix (w/ $\mathcal{C}_{cls}$) leads to a noticeable performance drop of 0.7 AP_{25} and 1.6 AP_{50}. This strongly supports our spatial-prioritized strategy, demonstrating that fine-grained geometric cues are sufficient to drive effective label assignment. Then, removing the normalized Vertex Distances regularizer (w/o $\mathcal{R}_{VD}$) causes 0.5 AP_{25} and 0.7 AP_{50} drop, confirming the importance of fine-grained geometric cues. These results highlight the importance of spatial cues in cluttered 3D scenes, where



(a) (b) (c)

Figure 6: The analysis of inference-aligned learning in SR3D. (a) The training Average Inconsistency Coefficient (AIC) shows lower inconsistency of SR3D. (b) Confidence vs. IoU scatter plots show that the distribution of SR3D's outputs more closely with the ideal diagonal, *i.e.*, Confidence = IoU, indicating better inference consistency. (c) Prediction Consistency Error (PCE) reveals smaller gaps between classification scores and localization accuracy in SR3D.

SPOTA	RAS	AP_{25}	AP_{50}	latency
		70.8	55.6	42ms
✓		72.3	57.4	42ms
	✓	72.5	57.7	42ms
✓	✓	**73.2**	**58.5**	42ms

Table 2: Ablation study of SR3D on each component.

Setting	AP_{25}	AP_{50}
SPOTA (Ours)	**73.2**	**58.5**
w/ $\mathcal{C}_{cls}$	72.5	56.9
w/o $\mathcal{R}_{VD}$	72.7	57.8

Table 3: Effect of the designs in SPOTA.

OTA-based method	AP_{25}	AP_{50}
simOTA (Ge et al. 2021b)	72.3	56.5
AlignOTA (Xu et al. 2022)	72.2	56.9
SPOTA (Ours)	**73.2**	**58.5**

Table 4: Superiority of SPOTA.

Quality-aware loss	AP_{25}	AP_{50}
QFL (Li et al. 2020b)	71.9	57.7
VFL (Zhang et al. 2021)	71.7	58.3
RAS (Ours)	**73.2**	**58.5**

Table 5: Superiority of RAS.

spatial structure is more informative than visual appearance.

To further demonstrate the effect of SPOTA, we compare it with other recent OT-based label assignment methods, simOTA (Ge et al. 2021b) and AlignOTA (Xu et al. 2022). As shown in Tab. 4, Our method achieves more reliable detection results (73.2 for AP_{25} and 58.5 for AP_{50}) compared to these methods. This demonstrates that our spatial-prioritized strategy eliminates redundant influences from classification and achieves improved detection performance.

Effect of RAS. Recent works such as Quality Focal Loss (QFL) (Li et al. 2020b) and Varifocal Loss (VFL) (Zhang et al. 2021) aim to embed localization quality into classification targets by directly supervising the classification branch using localization quality (*i.e.*, IoU) as labels. While our Rank-aware Adaptive Self-Distillation (RAS) differs in formulation from these losses, their objectives remain similar. As shown in Tab. 5, our RAS demonstrates superior compatibility, surpassing QFL by 1.3 for AP_{25} and 0.8 for AP_{50}, and VFL by 1.5 for AP_{25} and 0.2 for AP_{50}. These results highlight a fundamental limitation of quality-aware losses in the 3D domain: Low 3D IoU scores create optimization conflicts when localization quality is used to supervise classification, thereby harming training stability and performance. In contrast, our RAS distills signals of localization quality and relative ranking while preserving the original objective, providing more stable and discriminative supervision, even under the low-IoU conditions common in 3D detection.

Conclusion

This paper presents SR3D, a novel and efficient framework for real-time 3D object detection of indoor scenes. SR3D targets the fundamental training-inference gap in dense detectors, which primarily arises from missing spatial reliability and missing ranking awareness. We introduce two novel components to bridge this gap: the Spatial-Prioritized Optimal Transport Assignment (SPOTA) and the Rank-aware Adaptive Self-Distillation (RAS) scheme. The effectiveness of SR3D is validated on ScanNet V2 and SUN RGB-D, setting new benchmarks while maintaining real-time efficiency. Extensive analysis and ablation studies further demonstrate that SR3D effectively resolves the inconsistency issue, ensuring the inference-aligned learning.

Acknowledgments

This research is supported by the National Key Research and Development Program of China 2024YFC3811000, the NSFC-projects 42471447, and the Fundamental Research Funds for the Central Universities of China 2042022dx0001.

References

An, S.; Liao, Q.; Lu, Z.; and Xue, J.-H. 2022. Efficient semantic segmentation via self-attention and self-distillation. *IEEE Transactions on Intelligent Transportation Systems*, 23(9): 15256–15266.

Choy, C.; Gwak, J.; and Savarese, S. 2019. 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3075–3084.

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, G.; Qi, Z.; Shi, W.; and Ma, K. 2024. Point-GCC: Universal Self-supervised 3D Scene Pre-training via Geometry-Color Contrast. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, 4709–4718. New York, NY, USA: Association for Computing Machinery. ISBN 9798400706868.

Furlanello, T.; Lipton, Z. C.; Tschannen, M.; Itti, L.; and Anandkumar, A. 2018. Born-Again Neural Networks. In *Proceedings of the 35th International Conference on Machine Learning,ICML 2018*, volume 80 of *Proceedings of Machine Learning Research*, 1602–1611. PMLR.

Ge, Z.; Liu, S.; Li, Z.; Yoshie, O.; and Sun, J. 2021a. Ota: Optimal transport assignment for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 303–312.

Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021b. YOLOX: Exceeding YOLO Series in 2021. arXiv:2107.08430.

Gupta, V.; Liao, W.-k.; Choudhary, A.; and Agrawal, A. 2022. Brnet: Branched residual network for fast and accurate predictive modeling of materials properties. In *Proceedings of the 2022 SIAM international conference on data mining (SDM)*, 343–351. SIAM.

Gwak, J.; Choy, C.; and Savarese, S. 2020. Generative sparse detection networks for 3d single-shot object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, 297–313. Springer.

Ke, W.; Zhang, T.; Huang, Z.; Ye, Q.; Liu, J.; and Huang, D. 2020. Multiple anchor learning for visual object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10206–10215.

Kim, K.; and Lee, H. S. 2020. Probabilistic anchor assignment with iou prediction for object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, 355–371. Springer.

Lan, X.; Zhu, X.; and Gong, S. 2018. Knowledge distillation by on-the-fly native ensemble. *Advances in neural information processing systems*, 31.

Li, H.; Wu, Z.; Zhu, C.; Xiong, C.; Socher, R.; and Davis, L. S. 2020a. Learning from noisy anchors for one-stage object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10588–10597.

Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; and Yang, J. 2020b. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in Neural Information Processing Systems*, 33: 21002–21012.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017a. Feature Pyramid Networks for Object Detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 936–944.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017b. Focal Loss for Dense Object Detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2999–3007.

Liu, X.; Zhao, L.; Fan, B.; Lu, J.; and Liu, H. 2025. Dynamic Learnable Label Assignment for Indoor 3D Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(10): 10134–10147.

Liu, Z.; Zhang, Z.; Cao, Y.; Hu, H.; and Tong, X. 2021. Group-free 3d object detection via transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2949–2958.

Mei, G.; Riz, L.; Wang, Y.; and Poiesi, F. 2024. Geometrically-Driven Aggregation for Zero-Shot 3D Point Cloud Understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 27896–27905.

Misra, I.; Girdhar, R.; and Joulin, A. 2021. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2906–2917.

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 International Conference on Computer Vision*.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 658–666.

Rukhovich, D.; Vorontsova, A.; and Konushin, A. 2022. Fcaf3d: Fully convolutional anchor-free 3d object detection. In *European Conference on Computer Vision*, 477–493. Springer.

Shen, Y.; Geng, Z.; Yuan, Y.; Lin, Y.; Liu, Z.; Wang, C.; Hu, H.; Zheng, N.; and Guo, B. 2024. V-DETR: DETR with Vertex Relative Position Encoding for 3D Object Detection. In *The Twelfth International Conference on Learning Representations*.

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.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9627–9636.

Wang, C.; Yang, W.; Liu, X.; and Zhang, T. 2025. State Space Model Meets Transformer: A New Paradigm for 3D Object Detection. In *The Thirteenth International Conference on Learning Representations*.

Wang, H.; Ding, L.; Dong, S.; Shi, S.; Li, A.; Li, J.; Li, Z.; and Wang, L. 2022a. Cagroup3d: Class-aware grouping for 3d object detection on point clouds. *Advances in Neural Information Processing Systems*, 35: 29975–29988.

Wang, H.; Shi, S.; Yang, Z.; Fang, R.; Qian, Q.; Li, H.; Schiele, B.; and Wang, L. 2022b. Rbgnet: Ray-based grouping for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1110–1119.

Xie, Q.; Lai, Y.-K.; Wu, J.; Wang, Z.; Zhang, Y.; Xu, K.; and Wang, J. 2020. Mlcvnet: Multi-level context votenet for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10447–10456.

Xu, T.-B.; and Liu, C.-L. 2019. Data-distortion guided self-distillation for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 5565–5572.

Xu, X.; Jiang, Y.; Chen, W.; Huang, Y.; Zhang, Y.; and Sun, X. 2022. DAMO-YOLO: A Report on Real-Time Object Detection Design. *arXiv preprint arXiv:2211.15444v2*.

Yu, J.; Jiang, Y.; Wang, Z.; Cao, Z.; and Huang, T. 2016. UnitBox: An Advanced Object Detection Network. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, 516–520. New York, NY, USA: Association for Computing Machinery. ISBN 9781450336031.

Yun, S.; Park, J.; Lee, K.; and Shin, J. 2020. Regularizing class-wise predictions via self-knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13876–13885.

Zhang, H.; Wang, Y.; Dayoub, F.; and Sunderhauf, N. 2021. Varifocalnet: An iou-aware dense object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8514–8523.

Zhang, L.; Bao, C.; and Ma, K. 2021. Self-distillation: Towards efficient and compact neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8): 4388–4403.

Zhang, L.; Song, J.; Gao, A.; Chen, J.; Bao, C.; and Ma, K. 2019a. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3713–3722.

Zhang, P.; Kang, Z.; Yang, T.; Zhang, X.; Zheng, N.; and Sun, J. 2022. Lgd: label-guided self-distillation for object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 3309–3317.

Zhang, X.; Wan, F.; Liu, C.; Ji, R.; and Ye, Q. 2019b. Freeanchor: Learning to match anchors for visual object detection. *Advances in neural information processing systems*, 32.

Zheng, T.; Zhao, S.; Liu, Y.; Liu, Z.; and Cai, D. 2022. SCALoss: Side and Corner Aligned Loss for Bounding Box Regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; and Ren, D. 2020. Distance-IoU loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 12993–13000.

Zhu, B.; Wang, J.; Jiang, Z.; Zong, F.; Liu, S.; Li, Z.; and Sun, J. 2020. Autoassign: Differentiable label assignment for dense object detection. *arXiv preprint arXiv:2007.03496*.

Zhu, Y.; Hui, L.; Shen, Y.; and Xie, J. 2024. Spgroup3d: Superpoint grouping network for indoor 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7811–7819.