Leveraging Imagery Data with Spatial Point Prior for Weakly Semi-supervised 3D Object Detection

Hongzhi Gao[1], Zheng Chen[1], Zehui Chen[1], Lin Chen[1], Jiaming Liu[2], Shanghang Zhang[2], Feng Zhao[1]*

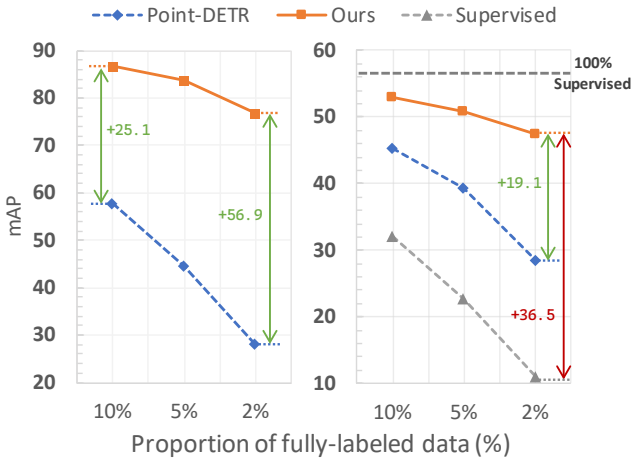
[1] University of Science and Technology of China
[2] Peking University
{hongzhigao, chenzheng, lovesnow, chlin}@mail.ustc.edu.cn,
liujiaming@bupt.cn, shanghang@pku.edu.cn, fzhao956@ustc.edu.cn

Abstract

Training high-accuracy 3D detectors necessitates massive labeled 3D annotations with 7 degree-of-freedom, which is laborious and time-consuming. Therefore, the form of point annotations is proposed to offer significant prospects for practical applications in 3D detection, which is not only more accessible and less expensive but also provides strong spatial information for object localization. In this paper, we empirically discover that it is non-trivial to merely adapt Point-DETR to its 3D form, encountering two main bottlenecks: 1) it fails to encode strong 3D prior into the model, and 2) it generates low-quality pseudo labels in distant regions due to the extreme sparsity of LiDAR points. To overcome these challenges, we introduce Point-DETR3D, a teacher-student framework for weakly semi-supervised 3D detection, designed to fully capitalize on point-wise supervision within a constrained instance-wise annotation budget. Different from Point-DETR which encodes 3D positional information solely through a point encoder, we propose an explicit positional query initialization strategy to enhance the positional prior. Considering the low quality of pseudo labels at distant regions produced by the teacher model, we enhance the detector's perception by incorporating dense imagery data through a novel Cross-Modal Deformable RoI Fusion (D-RoI). Moreover, an innovative point-guided self-supervised learning technique is proposed to allow for fully exploiting point priors, even in student models. Extensive experiments on representative nuScenes dataset demonstrate our Point-DETR3D obtains significant improvements compared to previous works. Notably, with only 5% of labeled data, Point-DETR3D achieves over 90% performance of its fully supervised counterpart.

Introduction

3D object detection is one of the fundamental tasks in autonomous driving perception. In recent years, contemporary 3D object detectors (Lang et al. 2019; Wang et al. 2022; Chen et al. 2022b; Wang and Solomon 2021; Chen et al. 2022c,a) have made significant strides. Current advanced detectors typically necessitate training on myriad scenes with precise 3D annotations, delineating the 3D objects' locations, dimensions, and orientations using 7 degree-of-freedom (DoF). However, the manual annotation process for 3D labels is time-consuming and expensive (Su, Deng, and Fei-Fei 2012), emphasizing the need to diminish dependence on massive 3D box annotations.



(a) Teacher Model (b) Student Model

Figure 1: Comparison of mAP on teacher and student models. (a) Our teacher model outperforms a large margin over the Point-DETR baseline. (b) Our student model which only uses 10% fully-labeled data achieves comparable performance with the 100% supervised paradigm on the Center-Point baseline.

To mitigate manual annotation efforts, previous studies have proposed weakly-supervised (Meng et al. 2020; Peng et al. 2022; Xu et al. 2022) and semi-supervised (Zhao, Chua, and Lee 2020; Yin et al. 2022; Park et al. 2022) methods for 3D object detection. These strategies either leverage incomplete annotations or utilize a small subset of meticulously annotated scenes combined with a more extensive collection of unlabeled scenes for training. Nevertheless, they yield limited improvements in detection performance. Weakly Semi-Supervised Learning (WSSL) (Li, Arnab, and Torr 2018; Chen et al. 2021; Zhang et al. 2022) bridges the gap between these approaches, offering a balance between labeling costs and model efficacy by annotating large-scale scenes with only one point per instance and a few scenes with 3D box annotations.

However, weakly semi-supervised 3D object detection us-

*Corresponding Author
Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

ing points remains an under-explored area. Modern WSSL frameworks primarily focus on 2D object detection (Chen et al. 2021; Zhang et al. 2022) and 2D segmentation (Kim et al. 2023; Papandreou et al. 2015; Li, Arnab, and Torr 2018; Lee et al. 2019). Despite the impressive performance in these tasks, we empirically find that applying them to 3D object detection is non-trivial, mainly due to overlooking the inherent characteristics of LiDAR points. In this paper, we start the point by adapting the competitive WSSL 2D-based Point-DETR (Chen et al. 2021) to a 3D-based WSSL detector, and identify two core impediments for the final detection performance. On the one hand, even though Point-DETR explicitly encodes the coordinates and category information through the proposed Point Encoder, it still struggles with limited positional priors, making instance identification challenging. On the other hand, it generates extremely low-quality pseudo labels in distant regions, likely due to the extreme sparsity of LiDAR points (*i.e.*, with only a few LiDAR points, accurate shape, and position annotation become unfeasible, even for skilled human experts).

To address these challenges, we introduce the so-called Point-DETR3D, a weakly semi-supervised 3D object detection framework designed to fully capitalize on point-level annotations within a constrained instance-wise label budget. Firstly, we propose an explicit positional query initialization strategy, rather than merely encoding 3D point coordinates through a learnable point encoder. This approach seamlessly integrates absolute 3D positions with object queries, initializing them directly from point annotations, and significantly accelerates model convergence (*i.e.*, our model can generate high-quality 3D bounding boxes with only one transformer decoder). Secondly, considering the characteristics of LiDAR points, we leverage the dense imagery data to assist the pseudo-label generation in distant regions. Specifically, we propose a novel cross-modal feature interaction module, namely Deformable RoI Fusion Module, to dynamically aggregate 2D information for accurate boundary localization. Furthermore, we introduce a general point-guided self-supervised learning paradigm for the student model, which makes full use of point annotations in a parameter-free manner, substantially mitigating pseudo-label noise and enhancing representation robustness.

In summary, the contributions are three-fold:

- We identify the main bottlenecks of the current WSSL framework in 3D detection, which mainly attributes to the insufficient usage of 3D point priors and the sparsity characteristic of LiDAR inputs.
- Based on the above observations, we propose a new WSSL framework for 3D object detection, namely Point-DETR3D, which consists of a point-centric teacher model and a self-motivated student model.
- Extensive experiments on the representative nuScenes benchmark demonstrate the effectiveness of our proposed detector. Notably, with only 5% of labeled data, Point-DETR3D achieves over 90% performance of its fully supervised counterpart.

Related Work

3D Object Detection. Existing 3D object detectors can be broadly categorized as uni-modal and cross-modal approaches. Uni-modal methods utilize point clouds (Zhou and Tuzel 2018; Lang et al. 2019) or images (Wang et al. 2022; Chen et al. 2022b) only as inputs. VoxelNet (Zhou and Tuzel 2018) and PointPillar (Lang et al. 2019) processes the point features at voxel or BEV space, respectively. CenterPoint (Yin, Zhou, and Krahenbuhl 2021) revisits the label assignment in 3D detection and proposes a general center-based assignment strategy. Object-DGCNN (Wang and Solomon 2021) and SST (Fan et al. 2021) prove that transformer-based detectors can also yield competitive results. Recently, there has been increasing attention on multi-modal data to further boost the 3D detection performance. TransFusion (Bai et al. 2022a) incorporates imagery information through a transformer-style fusion module. BEVFusion (Liang et al. 2022) directly concatenates the different modalities by unifying them into the same BEV space. DeepInteraction (Yang et al. 2022) introduces a modality interaction strategy that leverages representational and predictive interaction to prevent the loss of unique perception strengths. SparseFusion (Xie et al. 2023) proposes instance-level sparse feature fusion and cross-modality information transfer to take advantage of the strengths of each modality while mitigating their weaknesses. Despite being effective, how to utilize imagery information with point prior is still under-explored.

Semi-Supervised 3D Object Detection. Although supervised 3D object detection methods yield promising results, the heavy reliance on massive accurate annotations makes it hard to be deployed in real-world scenarios. Semi-supervised 3D detection presents a potential solution to alleviate the labeling cost issue, which only requires a small fraction of labeled data. Based on the Mean Teacher (Tarvainen and Valpola 2017) paradigm, SESS (Zhao, Chua, and Lee 2020) applies a consistency loss between teacher predictions and student predictions. DetMatch (Park et al. 2022) introduces a flexible framework for joint semi-supervised learning on 2D and 3D modalities, generating more robust pseudo-labels. The quality of the pseudo-labels is further enhanced by 3DIoUMatch (Wang et al. 2021), which employs a dedicated filter approach. ProficientTeacher (Yin et al. 2022) goes a step further by proposing an STE module to generate sufficient seed boxes, and then the CBV module is used to select high-quality predictions.

Weakly Semi-Supervised Object Detection. Previous works (Bilen and Vedaldi 2016; Tang et al. 2017; Yang, Li, and Dou 2019) commonly utilize image-level annotations as weak supervision. However, the absence of location information greatly hinders the model's performance. Recent works propose to leverage point annotations as weak supervision signals. WS3D (Meng et al. 2020) devises a two-stage framework that achieves competitive performance by utilizing only a small set of weakly annotated scenes, along with a few precisely labeled object instances. Point-Teaching (Ge et al. 2023) applies a Hungarian-based point-matching method to obtain pseudo labels, which are then utilized in conjunction with multiple instance learning (MIL)

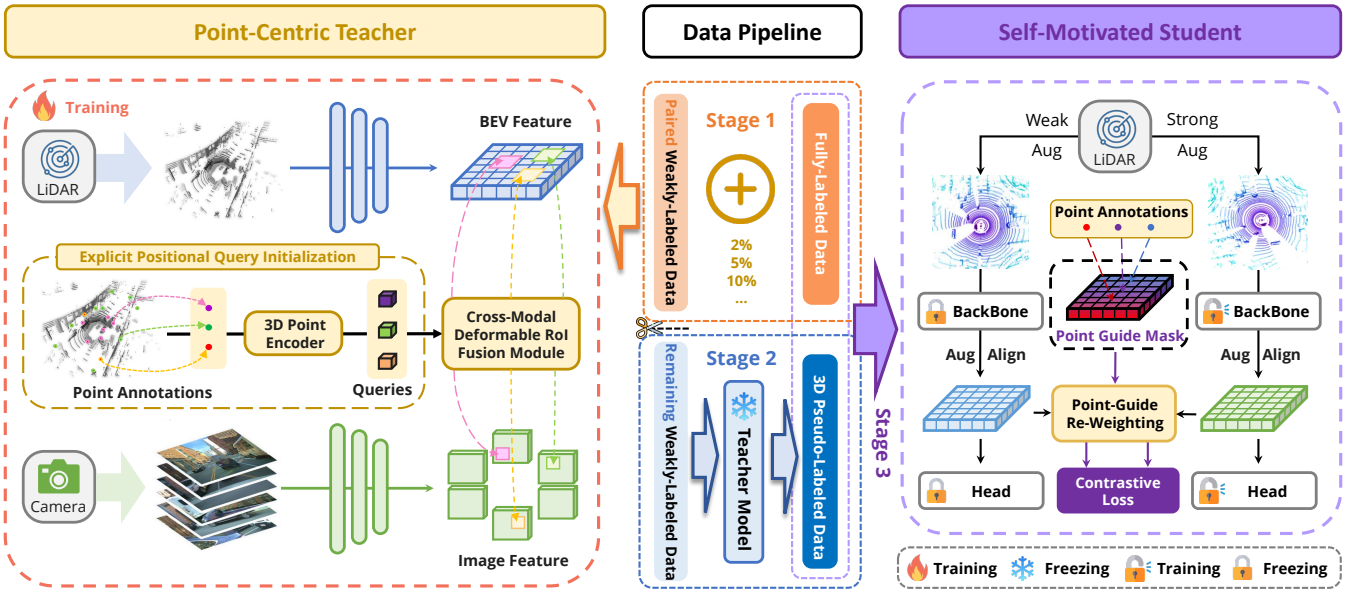


Figure 2: The overall framework of Point-DETR3D. The stage 1 is the training stage of the teacher model which utilizes limited fully-labeled data and paired weakly-labeled data as the training set. The stage 2 represents the pseudo-label generation stage of the teacher model, which transforms the remaining weakly-labeled data into 3D pseudo boxes. The stage 3 is the self-supervised training process of the student model, which is trained with the full set of both labeled and pseudo annotations.

approaches to supervise the object detector. To leverage point annotations, Point-DETR (Chen et al. 2021) introduces a transformer-based point-to-box regressor that transforms point annotations into pseudo boxes so that a student object detector can be trained. Group R-CNN (Zhang et al. 2022) designs a RPN-based point-to-box projector, which enhances RPN recall by establishing an Instance Group and further improves accuracy through the adoption of a new assignment strategy. However, most of works are applicable only to 2D target detection. In the realm of 3D detection, there remains a need to further investigate weakly semi-supervised approaches.

Point-DETR3D

In this section, we first review the preliminaries of weakly semi-supervised 3D object detection with point annotations (WSS3D-P). After that, we introduce Point-DETR3D, an effective teacher-student-based framework for WSS3D-P. It consists of two main components: the Point-Centric Teacher and the Self-Motivated Student. The overall framework is shown in Figure 2.

Preliminaries

3D Point Annotations. Each 3D point annotation includes an object point and its category label. We form each object as (x, y, z, c), where (x, y, z) and c represent point location in 3D space and object category, respectively. Note that the object point annotations are randomly sampled within the 3D box following a normal distribution prior. Such an approach enables us to dramatically alleviate the massive burden of precisely annotating objects in 3D tasks.

WSS3D-P Pipeline. WSS3D-P is a challenging task due to point annotations providing limited information, making it difficult to precisely predict 3D bounding boxes in a scene. It typically utilizes a small portion of well-labeled scenes and a large number of weakly point-level annotations as training data. Previous works adapt the teacher-student training paradigm as the default training pipeline, which leads to significant progress in semi-supervised learning (Wang et al. 2021; Yin et al. 2022). The steps are summarized as follows:

(1) Train a point-to-box teacher model with a small portion of paired point- and fully-annotated data.

(2) Generate pseudo labels for weakly-labeled scenes using the trained teacher model.

(3) Train a student model with the combination of fully-labeled scenes and the remaining pseudo-labeled scenes.

Point-Centric Teacher

The Point-Centric Teacher is designed to generate precise 3D pseudo boxes by fully capitalizing on the potential of weak point annotations. This is achieved through two core components: the Explicit Positional Query Initialization strategy and the Cross-Modal Deformable RoI Fusion module, which together seamlessly harness strong positional priors and dense visual features for precise localization.

Explicit Positional Query Initialization. While Point-DETR (Chen et al. 2021) utilizes point annotations by encoding both coordinates and category information through a learnable projector, such an approach yields minimal im-

provements when extended to the 3D domain. This limitation may be attributed to the significant increase in candidate space when lifting from 2D to 3D. To address this issue, we opt for an explicit positional query initialization strategy rather than merely embedding prior information implicitly via a 3D point encoder. Specifically, we enable an explicit bind between the absolute 3D position and each object query. Thanks to the generality of Object-DGCNN (Wang and Solomon 2021), this bind can be easily achieved through the *reference point*. To this end, we can generate n object queries, and each reference point of them is initialized as the n GT point annotations. Since there is a direct correlation between each object query and GT, we replace the originally complex bipartite matching with a straightforward one-to-one match based on such initialization, which significantly reduces the training instability observed in (Li et al. 2022).

Point-Centric Deformable RoI Fusion. Despite being informed with strong 3D points prior, we empirically observe that our teacher model produces subpar results in distant regions. In fact, even a skilled human expert would struggle to determine the exact boundaries of target instances with only a few points. To make up for this information deficit, we introduce dense imagery data as a valuable reference. Previous works (Yang et al. 2022; Chen et al. 2022c; Xie et al. 2023) have demonstrated remarkable performance gains by combining 2D and 3D data in 3D object detection. Yet, most works mainly focus on general cross-modal fusion at a voxel/point level, while effective instance-level fusion with 3D priors has rarely been explored. In this work, we present a novel Point-Centric Deformable RoI Cross-Modal Fusion operation (see Figure 3), which seamlessly aggregates instance-level features guided by the 3D prior through RoI-wise point sampling and dynamically decides the imagery regions for reference, especially in distant regions.

Formally, given a 3D instance point annotation $p_i^{3d} = (x_i, y_i, z_i)$, we first project it onto different image views using the projection matrix of the respective cameras. The rotation and translation matrices from the LiDAR coordinate to the camera coordinate of the j^{th} view are represented by $R_j \in \mathbb{R}^{3\times3}$ and $t_j \in \mathbb{R}^3$, respectively, and the intrinsic parameter of the j^{th} camera is denoted as $K_j \in \mathbb{R}^{3\times3}$. The point projection from LiDAR to the camera can be written as:

$$p_i^{img}(u_i, v_i) = K_j(R_j p_i^{3d} + t_j), \tag{1}$$

where $p_i^{img} = (u_i, v_i)$ is image coordinates of point annotations.

To obtain the RoI reference points p_{ik}^{RoI}, we generate a square grid centered around p_i^{img}. Here, k refers to the index of the RoI reference points in the RoI region ($k \in 1, ..., K^2, e.g., K = 4$), while K serves as the square grid size. Unlike in previous works(Yang et al. 2022; Chen et al. 2022c), we do not simply sample RoI features using reference points. Instead, we introduce a learnable offset Δ_{ik} to each RoI reference point, which dynamically determines the most informative neighboring features for aggregation and is not limited to fixed RoI regions. The process of deformable RoI cross-attention (DRoICA) can be formulated as:

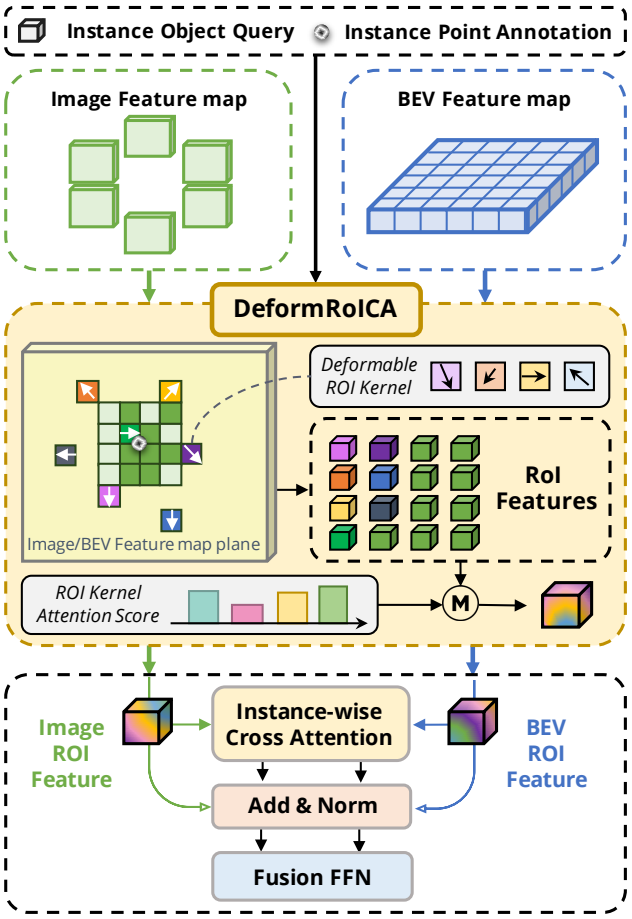


Figure 3: Illustration of the Point-Centric Deformable RoI Cross-Modal Fusion Module. First, it projects the instance point annotations onto the image and Bird's Eye View (BEV) plane, generating RoI-wise grid points. These grid points function as reference points and can be deformed through learnable coordinate offsets. Subsequently, each object query interacts with the corresponding RoI features to assimilate both LiDAR and imagery information. Finally, an instance-wise cross-attention module is applied to aggregate instance features from both modalities.

$$\begin{aligned}&\text{DRoICA}(Q_i, F) = \\ &\sum_{k=1}^{K} \text{MSDeformAttn}(Q_i, (p_{mlik}^{RoI} + \Delta_{mlik}), \{F_l\}_{l=1}^{L}),\end{aligned} \tag{2}$$

where m indexes the attention head and l indexes the input feature level. Q_i denotes the instance query and F is the corresponding multi-level image features with the RoI center point p_i^{img}. For the BEV feature map, we follow the same process to generate RoI grids for each instance query in the 2D plane. We then apply deformable RoI cross-attention to obtain the instance feature F_i^I and F_i^L from camera and LiDAR features. To maintain the spatial localization information from the LiDAR RoI feature and the textural informa-

tion from the image RoI feature, the corresponding features from the two modalities interact at the instance level through a cross-attention module proposed in (Vaswani et al. 2017).

Self-Motivated Student

In the original Point-DETR, the student model is solely trained on a combination of fully labeled data and pseudo-labeled data generated by the teacher model. However, this approach overlooks the potential benefits that the student model could derive from point annotations during training. In order to maximize the utilization of prior information of points and enhance the final performance of the model, we propose a simple and parameter-free self-supervised approach called Point-Centric Feature-Invariant Learning. It particularly focuses on mitigating the influence of label noises and strengthening the robustness of the model's representations.

Drawing inspiration from (Chen and He 2021; Chen et al. 2020), we employ the standard contrastive learning training paradigm as our foundational framework (see Figure 2). For a given input x, it is fed into two randomly augmented pipelines corresponding to weak and strong augmentations, producing two distinct inputs x_1 and x_2, respectively. Both of these views are processed by a feature extractor to derive BEV features. Subsequently, we reverse the BEV features given the augmented pipelines to generate the corresponding paired features for consistency regularization. Previous CL works(Liang et al. 2021; Chen and He 2021) directly force the student to mimic the feature map with equal supervision:

$$L_{feat} = \frac{1}{H \times W} \sum_{i}^{H} \sum_{j}^{W} ||F_{ij}^{2D} - F_{ij}^{2D}||_2, \tag{3}$$

where H, W denotes the width and height of the image feature map, and $|| \cdot ||_2$ is the L_2 norm. However, considering the characteristics of LiDAR points, there is a large portion of empty space in the BEV plane. Conducting supervision on these *NULL* representations makes no contribution to the model learning or even deteriorates the final performance. Therefore, we leverage the point annotations as foreground regional prompts to guide the learning positions. Concretely, we generate a Gaussian distribution for each point annotation (x_i, y_i) in the BEV space following (Yin, Zhou, and Krahenbuhl 2021):

$$m_{i,x,y} = \exp(-\frac{(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2}{2\sigma_i^2}), \tag{4}$$

where σ_i is a constant (default value of 2). Since the feature maps are class-agnostic, we combine all $m_{i,x,y}$ into one single mask M. For overlapping regions among different $m_{i,x,y}$ at the same position, we simply take the maximum value of them. After generating the mask M, we use it to guide the student in mimicking the feature for dense feature contrastive learning:

$$\begin{aligned} L_{feat} = & \frac{1}{H \times W \times \sum max(M_{i,j})} \\ & \sum_{i}^{H} \sum_{j}^{W} max(M_{ij}) ||F_{ij}^{3D} - F_{ij}^{3D}||_2 \end{aligned} \tag{5}$$

Such a point-guided re-weighting strategy enables the model to concentrate on the foreground regions from the teacher while avoiding the useless imitation of empty 3D features in the excessive background regions.

Experiments

Dataset and Evaluation Metrics

Dataset. We conduct the experiments on the nuScenes dataset(Caesar et al. 2020), a well-established dataset in the domain of 3D object detection. It consists of 700 training scenes, 150 validation scenes, and 150 testing scenes. For each scene, it provides point clouds collected from a 32-beam LiDAR system and images captured by 6 surrounding cameras with a resolution of 1600 × 900. For the 3D object detection task, 1.4M objects are annotated with 3D bounding boxes and classified into 10 categories.

Evaluation Metrics. We report *Static Properties nuScenes Detection Score* (SPNDS) and mean Average Precision (mAP) as our evaluation metrics. Specifically, SPNDS is derived from standard NDS by removing the velocity-related measurements including mean Average Velocity Error (mAVE) and mean Average Attribute Error (mAAE), since we only focus on the static attributes of 3D bounding boxes. Formally, SPNDS is a weighted sum of mean Average Precision (mAP) and the aforementioned True Positive (TP), which is defined as:

$$SPNDS = \frac{1}{8} \left[5 \times mAP + \sum_{\text{mTP} \in TP} (1 - \min(1, \text{mTP})) \right] \tag{6}$$

Implementation Details

Our codebase is built on MMDetection3D toolkit. All models are trained on an 8 NVIDIA A100 GPU machine. Following prior work(Chen et al. 2021), we sample 2%, 5%, and 10% of training scenes as the fully labeled set, and the rest scenes are set as weakly labeled. We employ a sequential sampling approach to avoid data leakage. To maintain consistency with Point-DETR, we select a transformer-based 3D detector Object-DGCNN(Wang and Solomon 2021) as our default teacher model. Note that any 3D detector can be set as the student model only if it provides BEV features. Without loss of generality, we choose CenterPoint (Yin, Zhou, and Krahenbuhl 2021) with both pillar-based and voxel-based settings.

Main Results

In this section, we report the experimental results on teacher and student models given different ratios of fully labeled sets. For a fair comparison, we carefully modify the Point-DETR to adapt it to the 3D domain, which serves as a competitive baseline.

Teacher Models Performance. Table 1 compares the results of our teacher model with Point-DETR under varying ratios of fully labeled data. Our proposed Point-DETR3D outperforms Point-DETR by more than 20.0 SPNDS and 30.0 mAP across all data configurations. Such significant

Backbone	Ratio	Paradigm	SPNDS ↑	mAP ↑
Pillar-based	2%	Point-DETR	23.87	23.55
		Ours	**66.03**	**80.47**
	5%	Point-DETR	35.28	39.89
		Ours	**69.18**	**84.09**
	10%	Point-DETR	51.33	61.41
		Ours	**73.19**	**86.50**
Voxel-based	2%	Point-DETR	27.49	28.19
		Ours	**63.98**	**76.82**
	5%	Point-DETR	39.28	44.56
		Ours	**70.28**	**83.63**
	10%	Point-DETR	49.35	57.58
		Ours	**76.46**	**87.46**

Table 1: Comparison of different teacher models on the validation set with varying ratios of labeled data. Note that point annotations are used in Point-DETR and ours(Point-DETR3D) during training and validation. ResNet-50 is the default image backbone in Point-DETR3D.

improvements highlight the efficacy of our explicit positional query initialization coupled with the incorporation of imagery data. Interestingly, the pillar-based teacher yields similar performance to the voxel-based one, suggesting that our models are nearing the upper bound of detection accuracy when only using point annotations.

Student Models Performance. We evaluate the benefits of different training paradigms by comparing the student model CenterPoint (Yin, Zhou, and Krahenbuhl 2021), which is trained on the pseudo labels generated from the corresponding teacher (pillar/voxel-based). The results are shown in Table 2 and Table 3. Both Point-DETR and Point-DETR3D markedly outperform the vanilla baseline, demonstrating the merit of weakly semi-supervised learning with point annotations. Moreover, Point-DETR3D leads the original Point-DETR by 5.5 $\sim$ 19.9 SPNDS under various labeled data conditions, substantiating the improved pseudo-label quality and the advantage of the point-centric feature-invariant learning module. It's worth noting that with only 5% fully labeled samples, our student model achieves almost 90% performance of a fully supervised 3D detector, which demonstrates that our Point-DETR3D is able to dramatically reduce the labeling efforts on 3D annotations.

Ablation Experiments

In this section, we conduct extensive ablation studies to study the effectiveness of each component of Point-DETR3D. If not specified, the labeled ratio and the backbone are set to 10% and Voxel-based by default.

Explicit Positional Query Initialization. In Table 4, we compare our explicit positional query initialization strategy with the implicit point encoder proposed in Point-DETR. Simply encoding the x, y coordinates and category information brings a reasonable improvement of 37 mAP over the baseline detector. Then, we replace such implicit initialization with our explicit positional query initialization, and the detection accuracy is enhanced by more than 10.0 mAP compared to the original Point-DETR-style encoding approach. We also discover that linking each object query to a unique GT annotation during training alleviates the model's learning challenges, enabling competitive results with just one decoder compared to six. We conjecture that our one-to-one label assignment strategy stabilizes the training process compared to original Hungarian matching, echoing the observation in (Li et al. 2022).

3D Detector	Ratio	Paradigm	SPNDS	mAP
CenterPoint -Pillar	100%	-	56.28	48.90
	2%	-	16.97	5.52
		Point-DETR	24.40	21.02
		Ours	**40.09**	**39.49**
	5%	-	23.85	13.34
		Point-DETR	35.14	30.75
		Ours	**45.81**	**41.56**
	10%	-	31.29	20.44
		Point-DETR	45.07	39.00
		Ours	**48.96**	**43.86**
CenterPoint -Voxel	100%	-	61.80	56.37
	2%	-	20.82	11.01
		Point-DETR	29.14	28.40
		Ours	**48.21**	**47.49**
	5%	-	33.21	22.82
		Point-DETR	42.55	39.26
		Ours	**53.99**	**50.80**
	10%	-	42.18	32.07
		Point-DETR	50.58	45.30
		Ours	**57.81**	**53.03**

Table 2: Comparison of different student models which are trained with the combined of labeled data and pseudo labels generated by difference teacher models. "-" means the model is trained on the full labeled data only.

Paradigm	2%	5%	10%
-	21.71/10.29	36.90/25.61	46.60/35.96
Point-DETR	30.55/30.55	43.88/42.81	54.56/51.26
Ours†	**51.04/54.53**	**56.39/57.15**	**60.93/58.99**

Table 3: Additional experiments of SPNDS/mAP results on FocalFormer3D. † denotes training without self-supervised paradigm in the student model.

Point-Centric Deformable RoI Cross-Modal Fusion. In this part, we study the effectiveness of our proposed DeformRoICM Fusion module, compared with other point/voxel-based multi-modal fusion strategies, including TransFusion, AutoAlignV2, and MoCa, and report the results in Table 5. We can find that simply adopting attention-

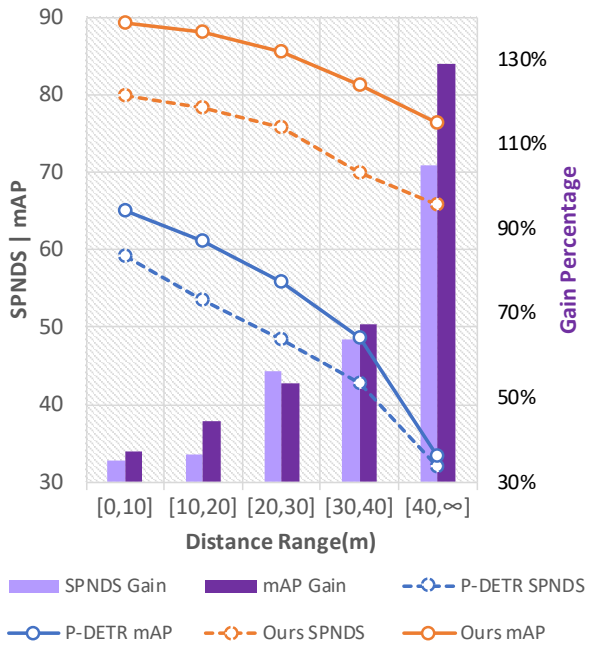


Figure 4: Comparison of Point-DETR3D (Ours) and Point-DETR (P-DETR) in terms of SPNDS, mAP, and Gains in different detection ranges.

Query Initialization	SPNDS	mAP
-	31.65	19.85
Implicit(Point-DETR)	49.35	57.58
Explicit(Ours)	**58.36(+9.01)**	**68.67(+11.09)**

Table 4: The effectiveness of explicit positional query initialization for the teacher model.

based approaches (TransFusion) yields limited improvement, which is mainly due to the weak positional correspondence during fusion. AutoAlignV2 and MoCa are two point-based fusion strategies, which keep more positional information and achieve 70 SPNDS. Furthermore, our proposed DeformRoICM module maximizes the advantage of point priors and yields an impressive 6.2 SPNDS and 7.5 mAP growth by enhancing the perception ability around objects. This is achieved by explicitly restricting the deformable sampling positions to be centered around RoI regions, rather than solely relying on the deformable offsets without any constraints. To further investigate the improvements in distant regions, we evaluate the detection performance with different distance intervals and report results in Figure 4. As the distance range increases, the detection performance of Point-DETR declines dramatically, which confirms the suppose that LiDAR inputs suffer from extreme sparsity of point clouds. Thanks to the dense visual information brought by imagery features, our approach suffers little performance drop as the distance goes up. Notably, at a distance of more than 40m, Point-DETR3D outperforms

Fusion Strategy	Venue	SPNDS	mAP
-	-	58.36	68.67
TransFusion	CVPR2022b	68.54	84.03
AutoAlignV2	ECCV2022a	70.26	79.88
MoCa	ICLRW2023	69.68	78.24
Ours	-	**76.46**	**87.46**

Table 5: The effectiveness of incorporating imagery data(Imagery) and Deformable RoI Fusion Module(D-RoI) for the teacher model.

SSL	Point-guided	SPNDS	mAP
		56.72	52.04
✓		56.70(-0.02)	51.85(-0.19)
✓	✓	**57.81(+1.09)**	**53.03(+0.99)**

Table 6: Comparison on different feature-based self-supervised learning(SSL) approaches for the student model.

Point-DETR, with an increase of 105% in SPNDS and 129% in mAP. These results demonstrate that our Point-Centric Teacher can significantly improve the quality of pseudo labels at far distances.

Point-Guided Self-Supervised Learning. We investigate the importance of our point-guided self-supervised paradigm in Table 6. Directly adopting full feature space for contrastive learning can even deteriorate the performance, with a slight drop of 0.2 mAP. We infer the reason that most regions in the BEV feature map are meaningless and noisy. Simply treating all regions equally would amplify the influence of the imprecise pseudo-labels on the student model. Then, we add the proposed point-guided supervision strategy to dynamically select the informative regions for contrastive learning, which enhances the performance by 1.0 mAP. Such an improvement demonstrates the effectiveness of the utilization of point priors at the student training stage, which is often omitted in previous WSSOD works.

Conclusion

In this work, we introduce Point-DETR3D, a point-centric teacher-student framework for weakly semi-supervised 3D object detection. Our approach capitalizes on spatial point priors, leveraging deformable RoI cross-attention networks to dynamically extract and aggregate RoI-wise features from different modalities. Additionally, we propose a general point-guided self-supervised learning paradigm for the student model, called Point-Centric Feature-Invariant Learning. This innovative approach can effectively mitigate the impact of label noise, thus enhancing the robustness of the model's representations. The experiments conducted on the challenging nuScenes benchmark demonstrate that Point-DETR3D outperforms supervised models by over 25 SPNDS and 35 mAP with only 2% labeled data, suggesting its potential applicability to various end-to-end LiDAR 3D detectors. We believe that Point-DETR3D is able to serve as a solid baseline and will inspire further exploration in the weakly semi-supervised 3D object detection area.

Acknowledgments

This work was supported by the JKW Research Funds under Grant 20-163-14-LZ-001-004-01, and the Anhui Provincial Natural Science Foundation under Grant 2108085UD12. We acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.

References

Bai, X.; Hu, Z.; Zhu, X.; Huang, Q.; Chen, Y.; Fu, H.; and Tai, C. L. 2022a. TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection with Transformers.

Bai, X.; Hu, Z.; Zhu, X.; Huang, Q.; Chen, Y.; Fu, H.; and Tai, C.-L. 2022b. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1090–1099.

Bilen, H.; and Vedaldi, A. 2016. Weakly supervised deep detection networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2846–2854.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

Chen, L.; Yang, T.; Zhang, X.; Zhang, W.; and Sun, J. 2021. Points as queries: Weakly semi-supervised object detection by points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8823–8832.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15750–15758.

Chen, Z.; Li, Z.; Zhang, S.; Fang, L.; Jiang, Q.; and Zhao, F. 2022a. Autoalignv2: Deformable feature aggregation for dynamic multi-modal 3d object detection. *arXiv preprint arXiv:2207.10316*.

Chen, Z.; Li, Z.; Zhang, S.; Fang, L.; Jiang, Q.; and Zhao, F. 2022b. Graph-DETR3D: rethinking overlapping regions for multi-view 3D object detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5999–6008.

Chen, Z.; Li, Z.; Zhang, S.; Fang, L.; Jiang, Q.; Zhao, F.; Zhou, B.; and Zhao, H. 2022c. Autoalign: Pixel-instance feature aggregation for multi-modal 3d object detection. *arXiv preprint arXiv:2201.06493*.

Fan, L.; Pang, Z.; Zhang, T.; Wang, Y. X.; Zhao, H.; Wang, F.; Wang, N.; and Zhang, Z. 2021. Embracing Single Stride 3D Object Detector with Sparse Transformer. *arXiv e-prints*.

Ge, Y.; Zhou, Q.; Wang, X.; Shen, C.; Wang, Z.; and Li, H. 2023. Point-teaching: weakly semi-supervised object detection with point annotations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 667–675.

Kim, B.; Jeong, J.; Han, D.; and Hwang, S. J. 2023. The Devil is in the Points: Weakly Semi-Supervised Instance Segmentation via Point-Guided Mask Representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11360–11370.

Lang, A. H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; and Beijbom, O. 2019. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12697–12705.

Lee, J.; Kim, E.; Lee, S.; Lee, J.; and Yoon, S. 2019. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5267–5276.

Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L. M.; and Zhang, L. 2022. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13619–13627.

Li, Q.; Arnab, A.; and Torr, P. H. 2018. Weakly-and semi-supervised panoptic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 102–118.

Liang, H.; Jiang, C.; Feng, D.; Chen, X.; Xu, H.; Liang, X.; Zhang, W.; Li, Z.; and Van Gool, L. 2021. Exploring geometry-aware contrast and clustering harmonization for self-supervised 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3293–3302.

Liang, T.; Xie, H.; Yu, K.; Xia, Z.; Lin, Z.; Wang, Y.; Tang, T.; Wang, B.; and Tang, Z. 2022. Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in Neural Information Processing Systems*, 35: 10421–10434.

Meng, Q.; Wang, W.; Zhou, T.; Shen, J.; Van Gool, L.; and Dai, D. 2020. Weakly supervised 3d object detection from lidar point cloud. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII*, 515–531. Springer.

Papandreou, G.; Chen, L.-C.; Murphy, K. P.; and Yuille, A. L. 2015. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, 1742–1750.

Park, J.; Xu, C.; Zhou, Y.; Tomizuka, M.; and Zhan, W. 2022. Detmatch: Two teachers are better than one for joint 2d and 3d semi-supervised object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, 370–389. Springer.

Peng, L.; Yan, S.; Wu, B.; Yang, Z.; He, X.; and Cai, D. 2022. Weakm3d: Towards weakly supervised monocular 3d object detection. *arXiv preprint arXiv:2203.08332*.

Su, H.; Deng, J.; and Fei-Fei, L. 2012. Crowdsourcing annotations for visual object detection. In *Workshops at the twenty-sixth AAAI conference on artificial intelligence*. Citeseer.

Tang, P.; Wang, X.; Bai, X.; and Liu, W. 2017. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2843–2851.

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

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, H.; Cong, Y.; Litany, O.; Gao, Y.; and Guibas, L. J. 2021. 3dioumatch: Leveraging iou prediction for semi-supervised 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14615–14624.

Wang, Y.; Guizilini, V. C.; Zhang, T.; Wang, Y.; Zhao, H.; and Solomon, J. 2022. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, 180–191. PMLR.

Wang, Y.; and Solomon, J. M. 2021. Object dgcnn: 3d object detection using dynamic graphs. *Advances in Neural Information Processing Systems*, 34: 20745–20758.

Xie, Y.; Xu, C.; Rakotosaona, M.-J.; Rim, P.; Tombari, F.; Keutzer, K.; Tomizuka, M.; and Zhan, W. 2023. SparseFusion: Fusing Multi-Modal Sparse Representations for Multi-Sensor 3D Object Detection. *arXiv preprint arXiv:2304.14340*.

Xu, X.; Wang, Y.; Zheng, Y.; Rao, Y.; Zhou, J.; and Lu, J. 2022. Back to reality: Weakly-supervised 3d object detection with shape-guided label enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8438–8447.

Yang, K.; Li, D.; and Dou, Y. 2019. Towards precise end-to-end weakly supervised object detection network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8372–8381.

Yang, Z.; Chen, J.; Miao, Z.; Li, W.; Zhu, X.; and Zhang, L. 2022. Deepinteraction: 3d object detection via modality interaction. *Advances in Neural Information Processing Systems*, 35: 1992–2005.

Yin, J.; Fang, J.; Zhou, D.; Zhang, L.; Xu, C.-Z.; Shen, J.; and Wang, W. 2022. Semi-supervised 3D object detection with proficient teachers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, 727–743. Springer.

Yin, T.; Zhou, X.; and Krahenbuhl, P. 2021. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11784–11793.

Zhang, S.; Yu, Z.; Liu, L.; Wang, X.; Zhou, A.; and Chen, K. 2022. Group R-CNN for weakly semi-supervised object detection with points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9417–9426.

Zhang, W.; Wang, Z.; and Loy, C. C. 2023. Improving Data Augmentation for Multi-Modality 3D Object Detection. In *International Conference on Learning Representations Workshop on Scene Representations for Autonomous Driving*, 1–10.

Zhao, N.; Chua, T.-S.; and Lee, G. H. 2020. Sess: Self-ensembling semi-supervised 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11079–11087.

Zhou, Y.; and Tuzel, O. 2018. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4490–4499.