Interactive 3D Object Detection with Prompts

Ruifei Zhang[1,2⋆], Xiangru Lin[4⋆], Wei Zhang[4], Jincheng Lu[4], Xuekuan Wang[4], Xiao Tan[4], Yingying Li[4], Errui Ding[4], Jingdong Wang[4], and Guanbin Li[3,5⋆⋆]

[1] The Chinese University of Hong Kong, Shenzhen, China
[2] Shenzhen Research Institute of Big Data, Shenzhen, China
[3] Sun Yat-sen University, Guangzhou, China
[4] Department of Computer Vision Technology (VIS), Baidu Inc., China
[5] Peng Cheng Laboratory
ruifeizhang@link.cuhk.edu.cn, liguanbin@mail.sysu.edu.cn,
{linxiangru,zhangwei99,lujincheng01,wangxuekuan,tanxiao01,
liyingying05,dingerrui,wangjingdong}@baidu.com

Abstract. The evolution of 3D object detection hinges not only on advanced models but also on effective and efficient annotation strategies. Despite this progress, the labor-intensive nature of 3D object annotation remains a bottleneck, hindering further development in the field. This paper introduces a novel approach, incorporated with "prompt in 2D, detect in 3D" and "detect in 3D, refine in 3D" strategies, to 3D object annotation: multi-modal interactive 3D object detection. Firstly, by allowing users to engage with simpler 2D interaction prompts (e.g., clicks or boxes on a camera image or a bird's eye view), we bridge the complexity gap between 2D and 3D spaces, reimagining the annotation workflow. Besides, Our framework also supports flexible iterative refinement to the initial 3D annotations, further assisting annotators in achieving satisfying results. Evaluation on the nuScenes dataset demonstrates the effectiveness of our method. And thanks to the prompt-driven and interactive designs, our approach also exhibits outstanding performance in open-set scenarios. This work not only offers a potential solution to the 3D object annotation problem but also paves the way for further innovations in the 3D object detection community.

Keywords: Interactive 3D Object Detection · Prompt Learning

1 Introduction

3D object detection is pivotal for autonomous driving, gaining significant traction in recent years. Current research often derives 3D object information from sources such as monocular [24, 39, 51, 54] or multi-view camera images [11, 12, 18, 21], point cloud [17, 46, 53], and multi-modal sensors [1, 5, 45]. This has led to remarkable progress in detection performance. A major catalyst behind these advancements is the availability of high-quality 3D datasets like KITTI [9],

⋆ Equal contribution.
⋆⋆ Corresponding author.

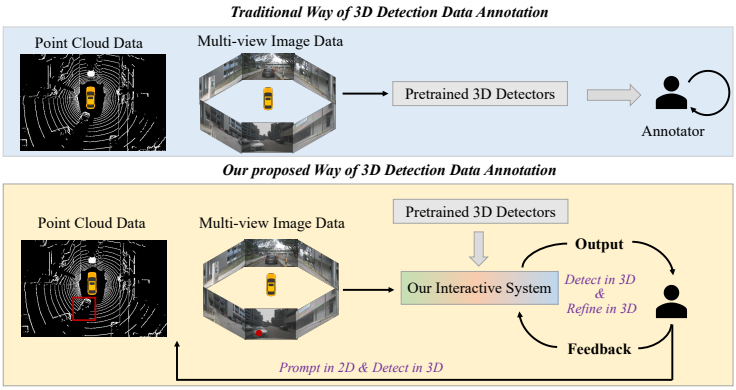

Fig. 1: We introduce an interactive 3D object detection concept. Rather than the traditional method where a model produces 3D object detection boxes followed by user refinement in 3D space, our strategy is "prompt in 2D, detect in 3D, refine in 3D". This method streamlines the annotation process, reducing the need for user refinements.

nuScenes [2], and Waymo [36]. These datasets fuel the progression of data-driven algorithms in 3D object detection. However, the intricate nature of 3D object annotation often necessitates significant human resources and incurs high labor costs. On average, annotating an hour of driving data takes hundreds of hours [27]. This annotation challenge potentially restricts the scale and diversity of datasets, possibly hindering further progression in this field.

Addressing this annotation challenge is the central theme of our paper. While weakly-supervised [25, 27–29] and semi-supervised [38, 43, 50] object detection methods exist, they typically suffer from reduced performance, especially in practical scenarios. This does not substantially address the overarching issue of the data-hungry nature of the field. With the recent emergence of SAM [13], both industry and academia are now recognizing the potential of interactive annotation approaches.

Our work aims to integrate this concept into 3D object detection, thus simplifying the annotation process, as presented in Fig. 1. Specifically, we observe that within the traditional process of annotating 3D detection data, annotators tend to concentrate on addressing two challenges stemming from imperfect pretrained 3D detectors: **1) supplementing missing closed-set objects and incorporating additional required open-set ones; 2) rectifying existing erroneous annotations.** Both scenarios necessitate annotators to operate within the 3D point cloud space, which is extremely time-consuming and labor-intensive.

To circumvent this, we create an interactive 3D object detection system enhanced by two strategies: ***prompt in 2D, detect in 3D***, and ***detect in 3D, refine in 3D***. The first principle enables users to interact with a 2D camera image or bird's eye view (BEV) of the point cloud, prompting the model to generate corresponding 3D bounding boxes, instead of encountering the 3D point cloud

directly. Furthermore, benefiting from the prompt-driven annotation paradigm, our method inherently possesses the capability to annotate **open-set** objects. The latter ensures annotators refine the initial 3D bounding box at the fastest pace with minimal effort, which significantly streamlines operations in the 3D space.

In light of these principles, we present our multi-modal interactive 3D object detection task and algorithm. Firstly, it employs prompts in the form of clicks or boxes on a 2D camera or BEV view, guiding the model to generate the relevant 3D bounding box. Our framework is reminiscent of the DETR-series detectors [3, 20, 26, 40] but comes with distinctive features. We integrate user inputs as prompt tokens and design a learnable component that investigates multi-modal features of indicated objects, informed by the prompts. Furthermore, we incorporate a global-to-local strategy to supply fine-grained object features, enhancing detection accuracy. Secondly, we empower our model with iterative refinement capability by incorporating the predicted 3D bounding box encoding as a new prompt into the query sequence. For those inaccurate 3D predictions, our framework allows annotators to adjust a few wrong attributes, e.g. dragging the center point position or adjusting the orientation angle of the 3D bounding box, and then perform self-correction on other characters.

To rigorously evaluate our approach, we devise two evaluation configurations for practical annotation scenarios. The first, regarding **Annotation from Raw Data**, leverages user 2D prompts and produces corresponding 3D predictions. Besides, our framework is also compatible with starting from the results of existing State-of-the-Art (SOTA) 3D detection models, termed as **Annotation from SOTA Model** and iteratively refining them. Experimental results on the nuScenes dataset highlight the effectiveness of our method in various scenarios.

In summary, this paper has the following contributions:

- We introduce a new concept of multi-modal interactive 3D object detection, proposing a substantial solution to prevailing annotation challenges. Our work is expected to provide critical insights for the 3D object detection community across academic and industrial domains.
- We unveil a prompt-driven interactive 3D object detection methodology with self-correction capability, leveraging it to establish an interactive system tailored for practical annotation scenarios.
- Comprehensive experiments in both annotation scenarios and evaluation settings attest to our method's efficacy.

2 Related Work

2.1 3D Object Detection

3D object detection techniques have been developed across a range of input modalities. Our approach emphasizes the advantages of interactive inputs, setting it apart from traditional methods:

Camera-only. Monocular 3D object detection techniques [24, 39, 51, 54] enjoy advantages such as low cost and ease of deployment. However, it also faces the inherent challenge of limited depth information. Multi-view methods utilize data from multiple camera perspectives [11, 12, 18, 21, 22, 30, 41], which enhances the model's ability to perceive the 3D environment, thereby improving detection accuracy.
LiDAR-only. Benefiting from the robustness and high precision, numerous studies utilize LiDAR data for 3D object detection. Point-based methods [17, 33, 48] like PointNet [31] directly engage with raw LiDAR data, while grid-based strategies structure the data into 3D voxels [46, 53] or feature pillars [14].
Multi-modality. Capitalizing on the strengths of camera images and point clouds, multi-modal 3D object detection has seen significant advancements [1, 5, 45]. Early methods combine camera images and point clouds at the input or result stages [5, 34]. While, current models, like TransFusion [1] and CMT [45], emphasize more intricate feature-level fusion. Uniquely, our method embeds interactive capacities, ensuring richer, user-tailored annotations.

2.2 Interactive Object Detection/Segmentation

Interactive methods have historically aimed to reduce the annotation workload while maintaining, or even enhancing, model performance:
Interactive Object Detection. Early attempts, like the incremental learning strategy [49], require user corrections to update detectors. More recent models like C3Det [15] have substantially eased the annotation process. Tools such as LATTE [37] and iDet3D [7] further simplify the 3D object annotation process. Our framework, however, brings in the versatility of prompt formats and modalities, offering users a broadened interaction spectrum coupled with robust detection capabilities.
Interactive Object Segmentation. Interactive segmentation has matured significantly over the years [4, 35, 44]. Models like SAM [13] have empowered users with tools like clicks, boxes, and referring expressions to guide the segmentation process. Diverging from interactive segmentation confined to 2D spaces, our work addresses the challenge of integrating 2D interactive prompts with 3D object detection, mitigating the spatial discrepancy between 2D annotations and 3D environments.

2.3 Weakly-supervised Object Detection

The notion of harnessing weak signals, like clicks [29] or extreme-points [25, 28], for supervising detectors has been a topic of intrigue. Some methods [27] have extended this idea to 3D object detection, using center-click based strategies. These approaches, while commendable for their balance between performance and annotation overhead, sometimes fail to harness the full potential of the detectors. Our methodology, on the other hand, incorporates strong 3D bounding box supervisions in an interactive setting, streamlining the data acquisition process without compromising on detection performance.

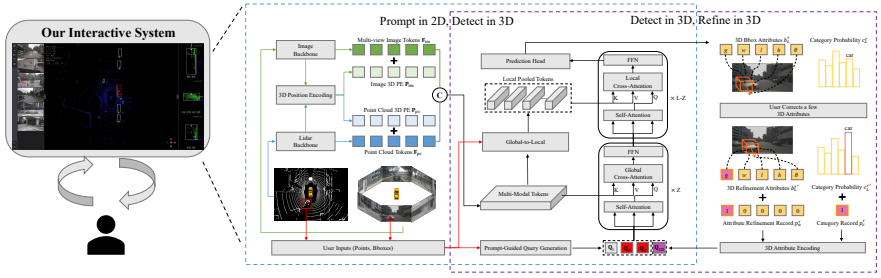

Fig. 2: An overview of our proposed interactive 3D object detection system/method. Taking multi-modal data with user prompts as input, we explicitly add 3D information $\mathbf{P}_{im}$, $\mathbf{P}_{pc}$ into features $\mathbf{F}_{im}$, $\mathbf{F}_{pc}$ by **3D Position Encoding**. Object queries consist of a learnable token $\mathbf{Q}_l$ and prompt ones $\mathbf{Q}_p$, which are generated by **Prompt-Guided Query generation**. We also introduce a **Global-to-Local** enhancement strategy to dynamically adjust features integrated with queries from global representations to local RoI details. Besides, our framework allows users to refine any of 3D attributes directly, and encode the refined 3D bounding box $b_L^r{}'$ and category information $c_L^r{}'$, along with attribute refinement record p_a^r and p_c^r, into a new prompt token $\mathbf{Q}_{3D}$ by **3D Attribute Encoding**. g, w, l, h, θ denotes the gravity center, width, length, height and yaw angle respectively. This iterative refinement propels our model to automatically rectify other attributes, ultimately delivering satisfactory results.

3 Methodology

3.1 Problem Definition

We present an interactive multi-modal 3D object detection algorithm. Given a multi-view set of camera images and associated point cloud data, the objective is to predict both the class labels and the 3D bounding boxes, driven by user prompts either in the form of a *point* or a *box*. Further, our model also supports iteratively self-correction with limited human assistance. To supervise the iteration process, the Ground Truth (c, b) is employed. Here, c signifies the category label, and b represents the 3D attributes. We adopt a deep supervision strategy for our multi-layer Transformer decoder, which guides the predictions across each decoder layer of each iteration, denoted as $\{(c_i^r, b_i^r), i = 1, 2, ...L, r = 0, 1, ...M\}$, with L and M being the total number of decoder layers and iterative refinements.

3.2 Overview

The architecture underpinning our interactive 3D object detection method is visualized in Fig. 2, which could be roughly decomposed into two stages, i.e. "*prompt in 2D, detect in 3D*" and "*detect in 3D, refine in 3D*". In the first phase, two separate backbones are utilized to extract multi-view image features $\mathbf{F}_{im} = \{\mathbf{F}_{im}^t, t = 1, 2, ..., T\}$ and point cloud features $\mathbf{F}_{pc}$ respectively. To enhance these features with 3D-awareness, following PETR [21] and CMT [45], we introduce 3D position encoding. Specifically, for the multi-view images, this encoding is

represented as $\mathbf{P}_{im} = \{\mathbf{P}_{im}^t, t = 1, 2, ..., T\}$, and for the point cloud, it is denoted as $\mathbf{P}_{pc}$. For object queries, denoted as $\mathbf{Q}$, we utilize a unique blending of a learnable token, $\mathbf{Q}_l$, and prompt-specific tokens, $\mathbf{Q}_p$, for every detected object. The prompt queries, $\mathbf{Q}_p$, play a pivotal role in guiding the learnable token, $\mathbf{Q}_l$, by leveraging the insights from the multi-modal features. Furthermore, to enhance the detection accuracy across various object scales, we incorporate a global-to-local strategy within the decoder. This strategy ensures that the object queries are continually enriched by granular local features, enabling finer detections. Ultimately, the learnable token predicts the object's class and its 3D bounding box (c_L^0, b_L^0).

Subsequently, the model enters an interactive refinement stage. This stage reuses the multi-modal features but updates the object queries $\mathbf{Q}$ by inserting the encoding of the refined 3D prediction from the current iteration. Specifically, in the r-th iteration, we enable annotators to fine-tune the current prediction $(c_L^r,\ b_L^r)$ and encode the refined result $(c_L^{r\prime},\ b_L^{r\prime})$ and corresponding attribute refinement record $(p_c^r,\ p_a^r)$ into a new prompt $\mathbf{Q}_{3D}$. $\mathbf{Q}_{3D}$ and $\mathbf{Q}_p$ collectively drive $\mathbf{Q}_l$ to generate more accurate predictions in the next iteration. Note that we omit the iteration superscript of $\mathbf{Q}_{3D}$ and $\mathbf{Q}$ for simplicity.

3.3 3D Position Encoding

Multi-view Camera Images. The core idea of embedding 3D position for multi-view image features includes two steps: 1) modeling each pixel coordinate as a series of points along a ray in the camera frustum space, and 2) projecting camera frustum coordinates into 3D space. Specifically, for a given camera image feature, denoted as $\mathbf{F}_{im}^t \in \mathbb{R}^{W_{im} \times H_{im} \times C}$, each pixel can be visualized as tracing a ray within the camera frustum space. This perspective allows us to sample D discrete points along the depth dimension. This process is mathematically captured by:

$$\mathbf{P}_{cf}(u, v, j) = (u \times d_j, v \times d_j, d_j, 1)^T, j = 1, 2, ..., D \tag{1}$$

Here, $\mathbf{P}_{cf}$ indicates the camera's frustum coordinates. Pixel coordinates in the image are denoted by (u, v), and d_j represents the j-th depth value.

To relate these coordinates to a 3D context, we employ an inverse 3D projection method:

$$\mathbf{P}_{3d}^t(u, v, j) = \mathbf{K}_t^{3D} \mathbf{K}_t^{-1} \mathbf{P}_{cf}(u, v, j) \tag{2}$$

where $\mathbf{K}_t^{3D} \in \mathbb{R}^{4\times4}$ represents the transformation matrix that links 3D space with the t-th camera's coordinate space. $\mathbf{K}_t \in \mathbb{R}^{4\times4}$ is the intrinsic matrix of t-th camera. To ensure consistent value scales, we normalize the 3D coordinates to the range $[0, 1]$. The final 3D position encoding for the t-th camera image, represented as $\mathbf{P}_{im}^t \in \mathbb{R}^{W_{im} \times H_{im} \times C}$, is derived via a multi-layer perceptron (MLP): $\mathbf{P}_{im}^t = \mathrm{MLP}_{im}(\mathbf{P}_{3d}^t)$.

Point Cloud. In contrast to the approach for camera image features, where points are assembled along the depth axis, the LiDAR point set necessitates sampling along the height axis. Specifically, given a BEV feature map, denoted as

$\mathbf{F}_{pc} \in \mathbb{R}^{W_{pc} \times H_{pc} \times C}$, we can compute the corresponding 3D position information, represented as $\mathbf{P}_{pc} \in \mathbb{R}^{W_{pc} \times H_{pc} \times C}$, using the following steps: First, we determine the 3D position based on the feature map's coordinates:

$$\mathbf{P}_{3d}(u, v, j) = (u \times u_s, v \times v_s, h_j, 1)^T, j = 1, 2, ..., H \tag{3}$$

Here, (u, v) specifies the coordinate within the point cloud feature map. The variables (u_s, v_s) represent the dimensions of the feature grid. H is the number of sampled points along the height axis and h_j is the j-th height value. Next, we utilize a multi-layer perceptron (MLP) to compute the 3D position encoding from the derived 3D position: $\mathbf{P}_{pc} = \text{MLP}_{pc}(\mathbf{P}_{3d})$. Note that as in CMT, only one point is sampled along the height axis, thus simplifying the process into standard 2D position encoding. Consequently, in our methodology, we have adopted a sinusoidal embedding to cater to this behavior.

3.4 Prompt-guided Query Generation

Our approach provides an avenue for user interaction, distinct from traditional DETR-based detectors that often rely on multiple queries to discern image objects. By merging learnable tokens, $\mathbf{Q}_l$, with user-directed prompt tokens, $\mathbf{Q}_p$, we achieve object detection that is not only efficient but also offers a new angle on object localization.

Learnable Query $\mathbf{Q}_l$: Every object is associated with a unique learnable token that facilitates feature probing, and subsequent prediction of class labels and 3D bounding boxes. In line with prior research [20, 40], we conceptualize the learnable query as an anchor point, initiating it with the positional prior of the user prompt. The initialization process unfolds over three phases:

1). *2D Center Position Computation with User Prompt.* A user might provide a bounding box $box = (x_1, y_1, x_2, y_2)$ or a singular point $point = (x, y)$. Depending on the input type, the 2D center position (c_x, c_y) is determined as:

$$(c_x, c_y) = \begin{cases} (\frac{x_1+x_2}{2}, \frac{y_1+y_2}{2}) & \text{if bounding box} \\ (x, y) & \text{if point} \end{cases} \tag{4}$$

2). *2D Center Transformation into 3D Coordinate.* Given a user prompt on the BEV map, the 2D space naturally omits height data. We address this by designating $c_z = 0.5$, the median value of height, resulting in a 3D center labeled as $center = (c_x, c_y, c_z)$. Alternatively, for a user prompt within 2D multi-view images, we transform the center coordinates (c_x, c_y) into a set of 3D points along the depth axis, using Eqs. (1) and (2). The median of the ray is then selected as the 3D center, represented as $center = (c'_x, c'_y, c'_z)$. The 3D center serves as a reference point for bounding box prediction and we prioritize the center from BEV if the user provides prompts in both modalities.

3). *3D Position Encoding Extraction from Both Modalities.* The 3D *center* is projected onto both the camera image and BEV plane, yielding their 3D position encodings $\mathbf{Q}_{im} \in \mathbb{R}^{1 \times C}$ and $\mathbf{Q}_{pc} \in \mathbb{R}^{1 \times C}$, respectively. The learnable query is initialized by combining information from both modalities: $\mathbf{Q}_l = \mathbf{Q}_{im} + \mathbf{Q}_{pc}$.

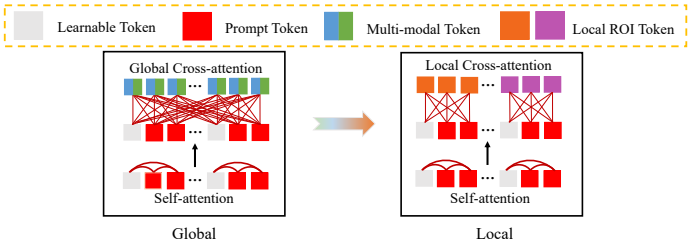

Fig. 3: Illustration of Global-to-Local Enhancement. Our queries interact with unmodified multi-modal features and its object's local features respectively in the global and local phases. Self-Attention is confined within queries of identical objects in both stages.

Prompt Query $\mathbf{Q}_p$: It is designed to embed user guidance by merging position encoding and context embedding: $\mathbf{Q}_p = \mathbf{Q}_{pe} + \mathbf{Q}_{ce}$. This synthesis is constructed for both box and point prompt formats: For the box format, the box position is delineated using its center point. From our 3D position encoding, the position encoding for box center, $\mathbf{Q}_{pe} \in \mathbb{R}^{1\times C}$, is derived. The context embedding, $\mathbf{Q}_{ce} \in \mathbb{R}^{1\times C}$, is secured via average pooling of the feature region bounded by the prompt box. For the point format, the same technique determines the position encoding, $\mathbf{Q}_{pe} \in \mathbb{R}^{1\times C}$. The feature at the indicated prompt point gives the context embedding, $\mathbf{Q}_{ce} \in \mathbb{R}^{1\times C}$.

3.5 Global-to-Local Enhancement

Given multi-modal features $\mathbf{F}_{pc}$ and $\mathbf{F}_{im}$ with their respective 3D position embeddings $\mathbf{P}_{pc}$ and $\mathbf{P}_{im}$, along with object queries $\mathbf{Q} \in \mathbb{R}^{N\times C}$, where N represents the cumulative number of queries from $\mathbf{Q}_l$ and $\mathbf{Q}_p$ across all objects, we employ a Transformer-based decoder to amalgamate these features for prediction.
Self-Attention: The formulation for the decoder's self-attention mechanism is:

$$\mathbf{Q}'_i = \text{softmax}(\mathbf{M}_{sa} + \mathbf{Q}_i\mathbf{Q}_i^T)\mathbf{Q}_i \tag{5}$$

where $\mathbf{Q}_i$ corresponds to the queries from the i-th layer, and $\mathbf{Q}_0 = \mathbf{Q}$. The attention mask $\mathbf{M}_{sa}$ ensures interaction is confined within queries of identical objects:

$$\mathbf{M}_{sa}(q_1, q_2) = \begin{cases} 0, & \text{if } q_1, q_2 \text{ belong to the same object,} \\ -\infty, & \text{otherwise.} \end{cases} \tag{6}$$

Cross-Attention: For cross-attention in the decoder, we express it as:

$$\mathbf{Q}_{i+1} = \text{softmax}(\mathbf{M}_{ca} + \mathbf{Q}'_i\mathbf{K}^T)\mathbf{V} \tag{7}$$

where $\mathbf{K}$, $\mathbf{V}$ come from multi-modal features, $\mathbf{M}_{ca}$ is attention mask for cross-attention. Influenced by traditional two-stage detectors [32] and recent DETR-centric techniques [6, 52], we adopt a **global-to-local enhancement** strategy.

This approach adjusts $\mathbf{K}$ and $\mathbf{V}$ dynamically to augment our model's local perceptiveness, as illustrated in Fig. 3.

For the initial Z layers among a total of L decoder layers, queries access unmodified multi-modal features:

$$\mathbf{K} = [f_k(\mathbf{F}_{pc} + \mathbf{P}_{pc}) : f_k(\mathbf{F}_{im} + \mathbf{P}_{im})], \tag{8}$$

$$\mathbf{V} = [f_v(\mathbf{F}_{pc}) : f_v(\mathbf{F}_{im})], \tag{9}$$

where $\mathbf{K}$, $\mathbf{V} \in \mathbb{R}^{(H_{pc}W_{pc}+H_{im}W_{im}T)\times C}$ are original multi-modal features under transformation $f_k(\cdot)$, $f_v(\cdot)$ and concatenation $[\cdot : \cdot]$. $\mathbf{M}_{ca} = 0$ during the global cross-attention phase.

In the succeeding $L - Z$ layers, local region-of-interest (RoI) features are extracted by RoI Align [10] via either the user-provided prompt box or the predicted 2D bounding box based on prompt point:

$$\mathbf{K} = [f_k(\mathrm{RoI}(\mathbf{F}_{pc} + \mathbf{P}_{pc})) : f_k(\mathrm{RoI}(\mathbf{F}_{im} + \mathbf{P}_{im}))], \tag{10}$$

$$\mathbf{V} = [f_v(\mathrm{RoI}(\mathbf{F}_{pc})) : f_v(\mathrm{RoI}(\mathbf{F}_{im}))], \tag{11}$$

where $\mathbf{K}$, $\mathbf{V} \in \mathbb{R}^{(H_oW_o+H_oW_o)\times C}$ now signify local multi-modal features. (H_o, W_o) is the output size of RoI Align. With $\mathbf{M}_{ca}$ defined similarly to $\mathbf{M}_{sa}$, it ensures each query is exclusively attentive to its object's local features. Following the decoding process, we derive the object query $\mathbf{Q}_L$ from the last decoding layer, and subsequently produce the initial prediction via two feed-forward networks (FFNs):

$$c_L^0 = \phi_c(\mathbf{Q}_L), \quad b_L^0 = \phi_b(\mathbf{Q}_L) \tag{12}$$

Here, c_L^0 and b_L^0 represent the class and 3D bounding box predictions, respectively. While ϕ_c and ϕ_b are FFNs designated for classification and regression tasks. At this point, the first phase of *"prompt in 2D, detect in 3D"* is completed.

3.6 3D Attribute Encoding

For those inaccurate results of the initial predictions, we endow our model with iteratively self-correction capability under limited human interaction. Specifically, we decompose the 3D bounding box into (*gravity center*, *width*, *length*, *height* and *yaw angle*) five attributes, along with category information, and assume users to adjust a few wrong attributes, e.g. "dragging the center point of the bounding box". Given the prediction b_L^r and c_L^r of the r-th iteration, the adjusted 3D bounding box and category are denoted as $b_L^r{}'$ and $c_L^r{}'$. We also introduce two binary attribute refinement records $p_a^r \in \mathbb{R}^5$ and $p_c^r \in \mathbb{R}^1$ to explicitly aid the model in distinguishing between corrected attributes and those yet to be predicted.

Finally, the 3D attribute query Q_{3D} for each object is formed by the combination of both attribute and record embedding:

$$\mathbf{Q}_{3D} = \mathrm{MLP}_{att}(b_L^r{}') + \mathrm{MLP}_{rec1}(p_a^r) + \mathrm{MLP}_{cat}(c_L^r{}') + \mathrm{MLP}_{rec2}(p_c^r) \tag{13}$$

And $\mathbf{Q}_{3D}$ is inserted into the corresponding query sequence of the respective object, collaborating with the original prompt tokens to jointly drive the model to provide more precise predictions. Note that in the following decoder process, $\mathbf{M}_{sa}$ and $\mathbf{M}_{ca}$ are also dynamically adapted according to Eq. (6).

3.7 Label Assignment and Losses

Due to the intrinsic one-to-one correspondence between our learnable queries and prompt objects, our proposed interactive 3D object detection task does not require label assignment like Hungarian Algorithm. To this end, following the methodologies in [21,45], we employ the Focal loss [19] for the classification task and the L1 loss for regression. To bolster the training of individual decoder layers, a deep supervision strategy is implemented. The 3D loss function amalgamates these predictions and is represented as:

$$\mathcal{L}_{3D} = \sum_{r=0}^{M} \sum_{i=1}^{L} \lambda_{cls} \mathcal{L}_{cls}(c_i^r, c) + \lambda_{reg} \mathcal{L}_{reg}(b_i^r, b) \tag{14}$$

Here, c and b are the ground truths, c_i^r and b_i^r denote the predictions of the i-th decoder layer and r-th iteration, which are generated by the shared MLP as Eq. (12). λ_{cls} and λ_{reg} are weights for balancing loss items. Besides, we also utilize L1 loss to supervise the 2D bounding box prediction on the image or BEV modality, which is produced by another MLP and serves as the local region for point prompts to perform the global-to-local strategy. The overarching loss function is the combination of both 3D and 2D losses.

4 Experiments

4.1 Experiment Settings

Dataset splits: The nuScenes dataset [2] is adopted in our work for training and evaluation. This dataset encompasses 1000 driving scenes, capturing data via six cameras, one LiDAR, and five radars. Following the official partition, we employ 750 scenes for training and 150 for validation. Contrasting prior work, our interactive 3D object detection approach leverages user prompts to guide the model's predictions. Similar to common interactive algorithms [4, 47], we reconstruct the evaluation benchmark by simulating user prompts and refinements from Ground Truth. Notably, since the nuScenes test set's Ground Truth remains inaccessible, we **exclude** it from our evaluations.

Training Protocol: Throughout the training process, we leverage Ground Truth to simulate practical user prompts and iterative refinements. Detailed training procedures for ***Prompt in 2D, Detect in 3D*** and ***Detect in 3D, Refine in 3D*** phases are provided in the supplementary materials.

Evaluation Protocol: We evaluate our method through human annotation comparisons and simulated experiments, under two distinct annotation scenarios: (1) ***Annotation from Raw Data*** and (2) ***Annotation from SOTA***

Table 1: Comparisons of **the average and standard deviation time cost** required for annotating an image extracted from the nuScenes validation set, with four annotation strategies, i.e. manual-only, Ours, SOTA model CMT with manual correction (CMT+M) and Ours (CMT+Ours) respectively.

Time cost (s)	*Annotation from Raw Data*		*Annotation from SOTA Model*	
	Manual-only	Ours	CMT+M	CMT + Ours
nuScenes-val	124±28	**55±17**	40±12	**24±7**

Model. We describe the simulated experimental settings as follows: in scenario (1), we employ 2D projections of 3D Ground Truth with perturbations to simulate user prompts, while in scenario (2), SOTA 3D detection models provide initial prediction results to initiate our algorithm. The attributes from Ground Truth are utilized to mimic subsequent user refinements in both scenarios. We explore various orders of adjustments to bounding box attributes and assess our model both **without** and **with** corrected category priors, denoted as **Ours** and **Ours***, respectively. * indicates the pre-calibration of category data. Unless specified, we prioritize using 2D bounding boxes as prompts and following the adjustment sequence of [gravity center, yaw angle, height, width, length]. Experimental results for alternative prompt formats (e.g., point prompts) and refinement orders are detailed in the supplementary materials. Furthermore, in addition to conventional **closed-set** evaluation, we also conduct extended experiments on challenging **open-set** scenarios. The nuScenes official metrics, i.e. nuScenes Detection Score (NDS), mean Average Precision (mAP) are adopted to evaluate the approaches.

4.2 Human Annotation Comparisons

To showcase the efficacy of our proposed method, we perform annotation comparisons across two scenarios employing four strategies: Manual-only, Ours, CMT+M (manual correction), and CMT+Ours. Ten experienced annotators are recruited to annotate the same set of twenty randomly selected nuScenes validation scenes using each of these strategies. Statistical analysis in Table 1 demonstrates that benefiting from the interactive design, our method achieves a speedup of approximately 56% and 40% compared to manual-only and CMT+M, respectively.

4.3 Closed-Set 3D Object Detection

Annotation from SOTA Model Evaluation. Results from our experiments on the nuScenes validation set are tabulated in Table 2. In practical applications, annotators often streamline their efforts by initiating annotations from the 3D predictions generated by SOTA models, such as CMT. Our model seamlessly integrates into this workflow, harnessing its robust capabilities to effectively aid in annotation tasks. Specifically, given the 3D predictions from CMT, Table 2 illustrates the interactive refinement results, without employing category

Table 2: Comparisons of annotation performance on nuScenes `val` set. "M" means manual correction. (Ri) denotes the i-th iteration refinement. "C" and "L" are camera and LiDAR modalities respectively.

Method	Modality	NDS↑	mAP↑
Traditional 3D Detectors			
UVTR [16]	CL	70.2	65.4
BEVFusion [23]	CL	71.4	68.5
MetaBEV [8]	CL	71.5	68.0
SpaseFusion [42]	CL	72.8	70.4
CMT [45]	CL	72.9	70.3
Annotation from SOTA Model			
CMT+M (R1)	CL	81.3	81.6
CMT+M (R2)	CL	83.8	81.6
CMT+M (R3)	CL	84.3	81.6
CMT+M (R4)	CL	85.3	81.6
CMT+M (R5)	CL	86.3	81.6
CMT+Ours (R1)	CL	83.4 (+2.1)	86.3 (+4.7)
CMT+Ours (R2)	CL	86.3 (+2.5)	86.3 (+4.7)
CMT+Ours (R3)	CL	87.4 (+3.1)	87.0 (+5.4)
CMT+Ours (R4)	CL	89.3 (+4.0)	89.1 (+7.5)
CMT+Ours (R5)	CL	90.4 (+4.1)	89.4 (+7.8)
Annotation from Raw Data			
Ours	CL	75.0	76.2
Ours (R1)	CL	84.4 (+9.4)	88.0 (+11.8)
Ours (R2)	CL	87.2 (+12.2)	88.0 (+11.8)
Ours (R3)	CL	88.2 (+13.2)	88.4 (+12.2)
Ours (R4)	CL	89.5 (+14.5)	89.5 (+13.3)
Ours (R5)	CL	90.5 (+15.5)	89.6 (+13.4)

correction priors. As a comparative reference, we perform identical manual attribute adjustments on the initial results predicted by CMT. It is evident that, in five consecutive iterations, our approach consistently outperforms CMT+M by margins of 2.1%, 2.5%, 3.1%, 4.0%, and 4.1%, respectively, in terms of NDS. Additionally, CMT maintains a constant mAP after the initial refinement of the center position, lacking self-correction for object categories. In contrast, our approach iteratively refines predictions with minimal human intervention, resulting in sustained performance improvements. In addition, we also provide comparisons with corrected category priors, denoted as CMT+M* and CMT+Ours* in Fig. 4. Through the rectification of object categories, our model exhibits a significantly enhanced self-correction capability, outperforming CMT+M* by a considerable margin.

Annotation from Raw Data Evaluation. In cases involving missed detections by existing 3D detectors, our method is expected to have the capability of annotating from raw data. As shown in Table 2, by leveraging a box as the prompt format, our model registers an NDS of 75.0% and an mAP of 76.2%. Subsequent iterations of refinements contribute to continued performance gains, and the final model attains an NDS score of 90.5% and mAP of 89.6%. Some visualization results are presented in Fig. 5.

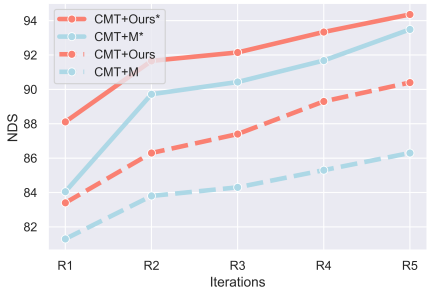

Fig. 4: Performance comparisons of annotation from SOTA model. R1-R5 follows a fixed refinement order. * indicates the utilization of category correction.

Table 3: Categories and quantities of open-set objects extracted from nuScenes validation set.

Category	Num	Category	Num
personal_mobility	15	stroller	103
animal	32	bicycle_rac	246
pushable pullable	2925	debris	514
ambulance	30	police	64

Table 4: Open-set evaluation.

Method	Modality	NDS↑	mAP↑
Annotation from Raw Data			
Ours*	CL	89.3	91.9
Ours* (R1)	CL	94.1	99.5
Ours* (R2)	CL	96.6	99.5
Ours* (R3)	CL	97.3	99.5
Ours* (R4)	CL	98.1	99.5
Ours* (R5)	CL	99.2	99.5

Table 5: Ablation study of prompt queries. "PE", "CE" are position encoding and context encoding. "It." denotes the iteration.

It.	$\mathbf{Q}_l$	$\mathbf{Q}_p$		$\mathbf{Q}_{3D}$			NDS↑	mAP↑
		PE	CE	$b_L^r{}'$	p_a^r			
0	✓						73.5	75.5
0	✓	✓					74.5	75.7
0	✓	✓	✓				**75.0**	**76.2**
1	✓	✓	✓	✓			83.2	85.9
1	✓	✓	✓	✓	✓		**84.4**	**88.0**

Table 6: Ablation study of Global-to-Local strategy (G2L). "L", "G" denotes utilizing local RoI or global multi-modal features to perform Cross-Attention.

It.	Attention	NDS↑	mAP↑
0	G	73.7	75.1
0	L	74.3	75.2
0	G2L	**75.0**	**76.2**

4.4 Open-Set 3D Object Detection

To delve deeper into the 3D object annotation capabilities guided by 2D prompts, we extract eight novel categories from the nuScenes dataset, which are not employed in the training process. Table 3 outlines their distribution on the validation set. Since our primary goal is to validate the model's generalization ability based on user-provided prompts, as opposed to venturing into open-vocabulary recognition, we ignore the class prediction and assume that the class information is provided manually together with 2D prompts, denoted as Ours*. The experimental results in Table 4 demonstrate that our model possesses the capability to detect 3D objects based on 2D priors. Furthermore, with manual interaction for refinement, the detection performance improves progressively.

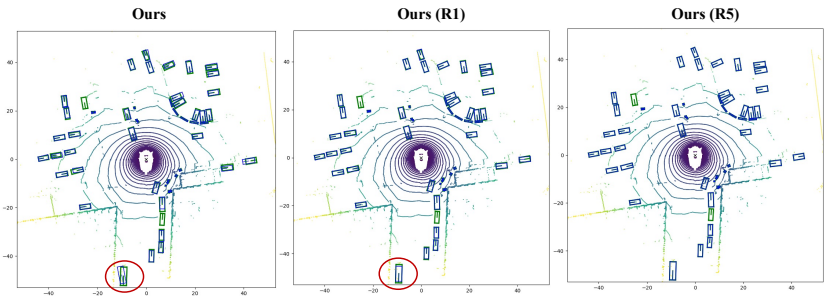

Fig. 5: Visualizations of the 3D detection results from our method are presented. Ground Truth and our predictions are exhibited in green and blue, respectively. The outcomes highlighted by red circles illustrate that, with a single round of manual adjustment focusing solely on the gravity center, our model automatically rectified other attributes such as yaw angles, thus yielding more accurate 3D results.

4.5 Ablation Study

Ablation Study of Object Queries. The pivotal role of prompt tokens is evident from the results in Table 5. When solely introducing prompt queries with position encoding, the model attains a nuanced understanding of the object's spatial location, registering an NDS of 74.5%—a significant boost of 1.0%. Moreover, integrating context embedding, which encapsulates appearance attributes of objects, further escalates performance, culminating in a gain of 0.5%. Besides, we also delve into an analysis of the effectiveness of Q_{3D} during the first iteration refinement stage. Clear performance improvements are evident with the explicit inclusion of the current attribute refinement record p_a^r.

Ablation Study of Global-to-Local Strategy. The proposed global-to-local strategy's efficacy is substantiated by the results displayed in Table 6. Contrasted with models relying exclusively on either global or local feature extraction, our hybrid approach ensures a comprehensive global perspective while meticulously capturing local nuances, leading to superior performance outcomes.

5 Conclusions

This paper introduces a new method for 3D object detection that tackles the daunting task of 3D object annotation. In consideration of the necessity for either creating annotations from raw data or refining existing imperfect annotations within practice annotation scenarios, two dedicated strategies are proposed. "Prompt in 2D, detect in 3D" principle leverages straightforward 2D interactions, like clicks or boxes, to streamline the transition between 2D images and 3D object annotations. "Detect in 3D, refine in 3D" strategy further endows our model with self-correction capability. Evaluations on the nuScenes dataset validate our approach's superiority. Beyond a mere advancement, our work can serve as a cornerstone for future 3D object detection endeavors.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (NO. 62322608), in part by the Fundamental Research Funds for the Central Universities under Grant 22lgqb25, and in part by the Open Project Program of the Key Laboratory of Artificial Intelligence for Perception and Understanding, Liaoning Province (AIPU, No. 20230003).

References

1. Bai, X., Hu, Z., Zhu, X., Huang, Q., Chen, Y., Fu, H., Tai, C.L.: Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1090–1099 (2022)
2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11621–11631 (2020)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision. pp. 213–229. Springer (2020)
4. Chen, X., Zhao, Z., Zhang, Y., Duan, M., Qi, D., Zhao, H.: Focalclick: Towards practical interactive image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1300–1309 (2022)
5. Chen, X., Ma, H., Wan, J., Li, B., Xia, T.: Multi-view 3d object detection network for autonomous driving. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1907–1915 (2017)
6. Chen, Y., Chen, Q., Sun, P., Chen, S., Wang, J., Cheng, J.: Enhancing your trained detrs with box refinement. arXiv preprint arXiv:2307.11828 (2023)
7. Choi, D., Cho, W., Kim, K., Choo, J.: idet3d: Towards efficient interactive object detection for lidar point clouds. arXiv preprint arXiv:2312.15449 (2023)
8. Ge, C., Chen, J., Xie, E., Wang, Z., Hong, L., Lu, H., Li, Z., Luo, P.: Metabev: Solving sensor failures for bev detection and map segmentation. arXiv preprint arXiv:2304.09801 (2023)
9. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3354–3361. IEEE (2012)
10. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2961–2969 (2017)
11. Huang, J., Huang, G.: Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. arXiv preprint arXiv:2203.17054 (2022)
12. Huang, J., Huang, G., Zhu, Z., Ye, Y., Du, D.: Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. arXiv preprint arXiv:2112.11790 (2021)
13. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)

14. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12697–12705 (2019)
15. Lee, C., Park, S., Song, H., Ryu, J., Kim, S., Kim, H., Pereira, S., Yoo, D.: Interactive multi-class tiny-object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14136–14145 (2022)
16. Li, Y., Chen, Y., Qi, X., Li, Z., Sun, J., Jia, J.: Unifying voxel-based representation with transformer for 3d object detection. Advances in Neural Information Processing Systems **35**, 18442–18455 (2022)
17. Li, Z., Wang, F., Wang, N.: Lidar r-cnn: An efficient and universal 3d object detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7546–7555 (2021)
18. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: European Conference on Computer Vision. pp. 1–18. Springer (2022)
19. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988 (2017)
20. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: Dab-detr: Dynamic anchor boxes are better queries for detr. arXiv preprint arXiv:2201.12329 (2022)
21. Liu, Y., Wang, T., Zhang, X., Sun, J.: Petr: Position embedding transformation for multi-view 3d object detection. In: European Conference on Computer Vision. pp. 531–548. Springer (2022)
22. Liu, Y., Yan, J., Jia, F., Li, S., Gao, A., Wang, T., Zhang, X., Sun, J.: Petrv2: A unified framework for 3d perception from multi-camera images. arXiv preprint arXiv:2206.01256 (2022)
23. Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D.L., Han, S.: Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 2774–2781. IEEE (2023)
24. Lu, Y., Ma, X., Yang, L., Zhang, T., Liu, Y., Chu, Q., Yan, J., Ouyang, W.: Geometry uncertainty projection network for monocular 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3111–3121 (2021)
25. Maninis, K.K., Caelles, S., Pont-Tuset, J., Van Gool, L.: Deep extreme cut: From extreme points to object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 616–625 (2018)
26. Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., Sun, L., Wang, J.: Conditional detr for fast training convergence. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3651–3660 (2021)
27. Meng, Q., Wang, W., Zhou, T., Shen, J., Jia, Y., Van Gool, L.: Towards a weakly supervised framework for 3d point cloud object detection and annotation. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(8), 4454–4468 (2021)
28. Papadopoulos, D.P., Uijlings, J.R., Keller, F., Ferrari, V.: Extreme clicking for efficient object annotation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4930–4939 (2017)

29. Papadopoulos, D.P., Uijlings, J.R., Keller, F., Ferrari, V.: Training object class detectors with click supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6374–6383 (2017)
30. Philion, J., Fidler, S.: Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16. pp. 194–210. Springer (2020)
31. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 652–660 (2017)
32. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems **28** (2015)
33. Shi, S., Wang, X., Li, H.: Pointrcnn: 3d object proposal generation and detection from point cloud. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–779 (2019)
34. Shin, K., Kwon, Y.P., Tomizuka, M.: Roarnet: A robust 3d object detection based on region approximation refinement. In: 2019 IEEE Intelligent Vehicles Symposium (IV). pp. 2510–2515. IEEE (2019)
35. Sofiiuk, K., Petrov, I.A., Konushin, A.: Reviving iterative training with mask guidance for interactive segmentation. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 3141–3145. IEEE (2022)
36. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2446–2454 (2020)
37. Wang, B., Wu, V., Wu, B., Keutzer, K.: Latte: accelerating lidar point cloud annotation via sensor fusion, one-click annotation, and tracking. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC). pp. 265–272. IEEE (2019)
38. Wang, K., Zhuang, J., Li, G., Fang, C., Cheng, L., Lin, L., Zhou, F.: De-biased teacher: Rethinking iou matching for semi-supervised object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2573–2580 (2023)
39. Wang, T., Zhu, X., Pang, J., Lin, D.: Fcos3d: Fully convolutional one-stage monocular 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 913–922 (2021)
40. Wang, Y., Zhang, X., Yang, T., Sun, J.: Anchor detr: Query design for transformer-based detector. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2567–2575 (2022)
41. Wang, Y., Guizilini, V.C., Zhang, T., Wang, Y., Zhao, H., Solomon, J.: Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In: Conference on Robot Learning. pp. 180–191. PMLR (2022)
42. Xie, Y., Xu, C., Rakotosaona, M.J., Rim, P., Tombari, F., Keutzer, K., Tomizuka, M., Zhan, W.: Sparsefusion: Fusing multi-modal sparse representations for multi-sensor 3d object detection. arXiv preprint arXiv:2304.14340 (2023)
43. Xu, M., Zhang, Z., Hu, H., Wang, J., Wang, L., Wei, F., Bai, X., Liu, Z.: End-to-end semi-supervised object detection with soft teacher. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3060–3069 (2021)
44. Xu, N., Price, B., Cohen, S., Yang, J., Huang, T.S.: Deep interactive object selection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 373–381 (2016)

45. Yan, J., Liu, Y., Sun, J., Jia, F., Li, S., Wang, T., Zhang, X.: Cross modal transformer via coordinates encoding for 3d object dectection. arXiv preprint arXiv:2301.01283 (2023)
46. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors **18**(10), 3337 (2018)
47. Yang, J., Zeng, A., Li, F., Liu, S., Zhang, R., Zhang, L.: Neural interactive keypoint detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15122–15132 (2023)
48. Yang, Z., Sun, Y., Liu, S., Jia, J.: 3dssd: Point-based 3d single stage object detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11040–11048 (2020)
49. Yao, A., Gall, J., Leistner, C., Van Gool, L.: Interactive object detection. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3242–3249. IEEE (2012)
50. Zhang, J., Lin, X., Zhang, W., Wang, K., Tan, X., Han, J., Ding, E., Wang, J., Li, G.: Semi-detr: Semi-supervised object detection with detection transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23809–23818 (2023)
51. Zhang, R., Qiu, H., Wang, T., Guo, Z., Xu, X., Qiao, Y., Gao, P., Li, H.: Monodetr: depth-guided transformer for monocular 3d object detection. arXiv preprint arXiv:2203.13310 (2022)
52. Zhao, J., Sun, L., Li, Q.: Recursivedet: End-to-end region-based recursive object detection. arXiv preprint arXiv:2307.13619 (2023)
53. Zhou, Y., Tuzel, O.: Voxelnet: End-to-end learning for point cloud based 3d object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4490–4499 (2018)
54. Zhou, Y., Zhu, H., Liu, Q., Chang, S., Guo, M.: Monoatt: Online monocular 3d object detection with adaptive token transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17493–17503 (2023)