Refine3D: Scene-Adaptive Reference Point Refinement for Sparse 3D Object Detection

Fan Li[*2], Jing Lu[*2], Yunlu Xu[*1, 2 †], Changhong Wu[2], Tao Xu[2], Zhaoyi Xiang[2], Yi Niu[2]

[1]Nanjing University
[2]Hikvision Research Institute
{lifan37, lujing6, xuyunlu, wuchanghong, xutao26, xiangzhaoyi, niuyi}@hikvision.com

Abstract

Sparse query-based detectors have emerged as the dominant paradigm in camera-only 3D object detection, owing to their exceptional performance and computational efficiency. A central component of these approaches is the use of reference points, which serve as learnable spatial anchors to guide queries in localizing target objects. However, existing methods typically employ a unified set of reference points across all scenes, a design we find suboptimal for handling complex scenarios with highly imbalanced object distributions, such as road intersections or occluded environments. In this paper, we investigate the adaptability of reference points and propose Refine3D, an adaptive refinement mechanism that achieves scene-level alignment between the distribution of reference points and ground-truth objects. In particular, we introduce a novel Reference Point Distribution Loss (RPD-Loss) to ensure reference points converge globally toward object positions, and a Scene-Adaptive Refinement head (SAR-Head) that predicts dynamic offsets for each reference point. Both components can be seamlessly integrated into mainstream sparse detectors. Extensive experiments on two challenging autonomous driving datasets demonstrate that Refine3D outperforms the state-of-the-art with improved detection accuracy and robustness.

Introduction

Camera-only 3D object detection has become a compelling solution in autonomous driving, thanks to its low cost, easy deployment, and long perception range. Existing approaches can be broadly categorized into Bird's-Eye-View (BEV)-based and query-based methods. BEV-based methods (Huang et al. 2022; Li et al. 2023b) project multi-view image features into a unified BEV space. Although effective, those typically incur substantial memory and computational overhead due to dense BEV grid prediction. Recently, query-based methods (Liu et al. 2022b; Wang et al. 2022) model potential objects as a sparse set of queries, which not only improves efficiency by focusing on foreground predictions but also achieves remarkable performance.

The query-based detection paradigm generally employs a set of learnable 3D reference points that represent 3D locations, which are embedded into high-dimensional positional embeddings to provide spatial clues for further detection. These positional embeddings are combined with learnable content embeddings to construct object queries, and are then iteratively refined through transformer-based decoding blocks to produce final results. Previous mainstream methods (Wang et al. 2023; Liu et al. 2024) adopt a unified set of reference points across varied driving scenarios and keep them fixed throughout all transformer decoding blocks, regardless of changes in object distributions. As shown in Fig. 1(a), such a static design leads to reference points being relatively uniformly distributed across the 3D space without

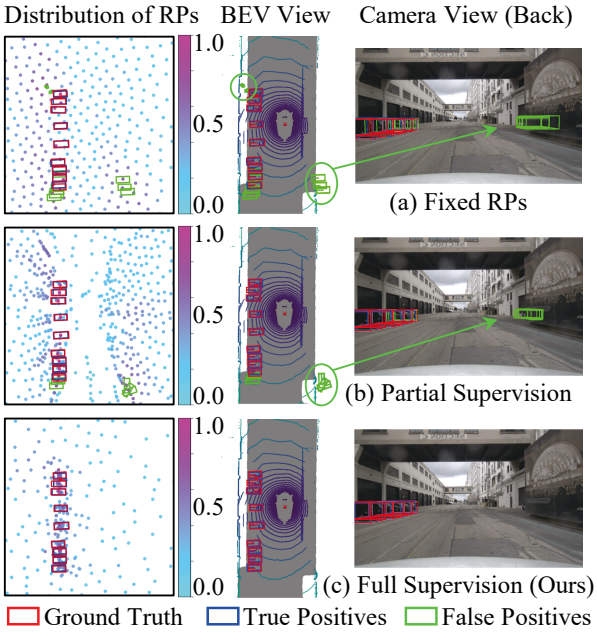


Figure 1: Comparison of reference point (RP) distributions and detection results under different refinement strategies: (a) fixed reference points (RPs); (b) partially supervised refinement; (c) fully supervised refinement (ours). Each row shows: BEV view of RP distribution (colored by confidence); BEV detection results; Perspective camera image. Our method concentrates RPs around objects, thus better suppressing false positives. Best viewed by zooming in.

[*]These authors contributed equally.
[†]Corresponding author.

differentiation between foreground and background. Consequently, many reference points fall into background regions, increasing the risk of high-confidence false positives (marked as green rectangles). This not only limits model's adaptability but also results in the inefficient use of queries.

To alleviate this issue, recent efforts such as the Sparse4D series (Lin et al. 2023a,b,c) have proposed dynamically updating reference points across different transformer decoding blocks.[1] This layer-wise refinement strategy encourages reference points to iteratively move toward potential object centers. However, this refinement is guided solely by the detection loss applied at the final output stage. As a result, only a small subset of reference points falling near ground-truth objects receive meaningful gradient signals that guide them toward the corresponding objects. The majority, particularly those initialized in background regions, remain unsupervised. As in Fig. 1(b), the reference points exhibit a nearly symmetric horizontal distribution after iterative updates, but the actual object distribution is horizontally asymmetric, which still leads to undesired false detections. This reflects the inadequacy of sparse supervision signals.

To better adapt reference points to scene-level object distributions under complex real-world driving occasions, we propose the scene-adaptive refinement approach, *i.e.*, Refine3D, which introduces explicit and dense supervision on reference point distributions. Specifically, we predict the offset of each reference point with a well-designed position-aware Scene-Adaptive Refinement Head (SAR-Head). It takes the query feature as input and outputs the 3D offset vector to iteratively update the reference point location. At its core lies the layer-wise embedding of the reference points' 3D coordinates injected into the head. This design enables the head to regress reference points in a discriminative manner based on their spatial priors. Then, we propose the Reference Point Distribution Loss (RPD-Loss) for dense regression constraints. The scene-level target distribution, derived from ground-truth object center points, provides supervisory signals to regulate reference point displacements at a global scale. A dynamic re-weighting mechanism is also used to increase the penalty for reference points drifting farther from their corresponding objects. To this end, Refine3D encourages explicit distribution alignment between reference points and scene-level object locations, providing layer-wise constraints for better reference point updates.

As shown in Fig. 1(c), our approach effectively concentrates reference points around foreground regions and significantly reduces false activations in cluttered or sparsely populated scenes. Moreover, given its modular plug-and-play design, our method can be seamlessly integrated into existing query-based detection frameworks.

To summarize, our contributions are threefold:

- We propose a novel Reference Point Refinement mechanism that adaptively aligns the reference point distribution to object locations across varied scenarios, showing particular effectiveness in handling imbalanced scenes.
- We propose the Reference Point Distribution Loss (RPD-Loss) as an explicit alignment constraint, together with a Scene-Adaptive Regression Head (SAR-Head) that predicts the offset of reference points. Moreover, both components can be seamlessly integrated into existing query-based detectors.
- Extensive validations on two challenging autonomous driving datasets demonstrate that our approach further improves the performance of state-of-the-art methods, particularly in scenarios with uneven or imbalanced object distributions.

[1] In the Sparse4D series, the term *3D anchor boxes* denotes 11-dimensional vectors. In this work, we focus on their positional components and refer to them as *reference points*, following the PETR-style terminology.

Related Work

Multi-View 3D Object Detection

Recently, 3D object detection from surround-view images has gained significant traction, due to its advantages in low deployment cost and rich semantic information. From a feature representation perspective, existing methods (Huang et al. 2022; Li et al. 2023b; Huang and Huang 2022; Yang et al. 2023; Liu et al. 2023a; Li et al. 2024; Wang et al. 2022; Liu et al. 2022b; Wang et al. 2023; Jiang et al. 2024; Liu et al. 2024; Lin et al. 2023c) can be broadly categorized into BEV-based and sparse query-based methods.

BEV-based methods transform multi-camera image features into a unified top-down representation, leveraging geometric consistency across views for 3D object detection. Building on this intermediate space, temporal modeling are introduced in a natural extension. BEVDet4D (Huang and Huang 2022) directly concatenates adjacent BEV features, while BEVFormer (Yang et al. 2023) adopts temporal self-attention within the BEV space. SOLOFusion (Park et al. 2022) further incorporates long-term memory based on BEVStereo (Li et al. 2023a) to capture extended temporal dependencies. Beyond architectural design, auxiliary supervision are also explored. BEVDepth (Li et al. 2023b) leverages depth ground truth to guide feature transformation, while IA-BEV (Jiao et al. 2024) employs 2D detection labels for additional semantic guidance. Despite these advances, BEV-based pipelines remain constrained by dense representations and the quality of view transformation, motivating the development of query-based detectors that operate directly on image features via sparse attention mechanisms.

Query-based methods address these limitations by directly interacting with image features through a sparse set of learnable queries, avoiding dense BEV construction. DETR3D (Wang et al. 2022) pioneers this approach by employing learnable 3D object queries to sample multi-view image features through transformer attention. Building upon this concept, PETR (Liu et al. 2022b) and PETRv2 (Liu et al. 2023b) introduce explicit 3D positional encodings to enhance spatial reasoning capability. Recent efforts further extend query-based frameworks with temporal modeling. StreamPETR (Wang et al. 2023) proposes a streaming architecture that efficiently integrates historical features, significantly improving detection accuracy and temporal consistency. Sparse4D series (Lin et al. 2023a,b,c) extend the sparse query concept to spatiotemporal queries, capturing

richer motion dynamics through sparse attention across temporal dimensions. Far3D (Jiang et al. 2024) introduces a frame-aware refinement mechanism, dynamically adjusting queries based on temporal correlations. To improve localization under weak supervision, RayDN (Liu et al. 2024) exploits depth priors through implicit geometric modeling. These advancements collectively highlight the effectiveness of query-based approaches in reducing computational complexity and enhancing temporal modeling, paving the way for efficient and accurate 3D object detection.

Optimizing Reference Point Distribution

In query-based detection frameworks, the impact of reference points has been extensively studied in 2D detection (Meng et al. 2021; Liu et al. 2022a), where each query typically comprises two components: a content query that encodes semantic information, and a positional query derived from reference points, which encodes spatial clues. These reference points are designed as learnable 2D coordinates and directly inform the positional queries, playing a crucial role in guiding attention mechanisms.

In 3D query-based detection, this design is extended by adapting reference points and their encoding processes into 3D space. Similar to 2D detection, 3D reference points serve as critical positional anchors that guide the attention mechanism within transformer architectures, thus significantly influencing detection accuracy. Given their central role, it is relatively intuitive to optimize the distribution of reference points to enhance detector performance. To address this, DETR3D (Wang et al. 2022) proposes a layer-wise refinement approach that progressively shifts reference points toward the centers of objects across transformer layers. FUTR3D (Chen et al. 2023) and Sparse4D (Lin et al. 2023a) improve this method by initializing reference points using statistical priors derived from k-means clustering, which alleviates training complexity. Far3D (Jiang et al. 2024) further incorporates auxiliary information from 2D detection and depth prediction modules, strategically positioning reference points near potential objects to ensure a more informative and effective spatial distribution.

Despite the progress, reference points are either kept fixed throughout transformer decoding blocks or only a few queries receive sparse ground-truth supervision. Instead, we investigate how global reference point distributions can be further optimized with dense scene-level representations, thus enhancing the 3D detection performance in complex scenes with diverse object distributions.

Methodology

Preliminary

Query-Based Detectors. Given V surround-view RGB images $\boldsymbol{I} = \left\{I_i \in \mathbb{R}^{3\times h\times w}, i = 1, \ldots, V\right\}$, a convolutional neural network is employed to extract multi-scale features, yielding $\boldsymbol{F} = \left\{F_i \in \mathbb{R}^{C\times H\times W}, i = 1, \ldots, V\right\}$, where C, H, and W represent the channel and spatial dimensions of the feature maps.

A transformer decoder $\mathbb{D} = \left\{\mathbb{L}^{(l)}, l = 1, ..., M\right\}$ with M attention blocks follows, as shown in Fig. 2(a). The typical structure of one block comprises a self- and cross-attention layer, together with a feedforward module for output projection. In the self-attention stage, a fixed set of N learnable 3D reference points $\mathbf{P} \in \mathbb{R}^{3\times N}$ is sampled in world coordinates and passed through an MLP to obtain positional embeddings $\mathbf{Q}_{\text{pos}} \in \mathbb{R}^{D\times N}$, where D is the embedding dimension. These are combined with content embeddings $\mathbf{Q}_{\text{con}} \in \mathbb{R}^{D\times N}$ (initialized to zeros) to form the input queries $\mathbf{Q}_{\text{in}} = \mathbf{Q}_{\text{pos}} + \mathbf{Q}_{\text{con}}$. During cross-attention, each query interacts with multi-view features $\boldsymbol{F}$, selectively aggregating information from spatially relevant image regions. The output is fed into the feedforward layer to update the content component of queries, denoted as $\mathbf{Q}_{\text{out}}$, which are then passed into to the subsequent attention block. The final updated queries $\mathbf{Q}_{\text{out}}$ from the last attention block are passed through a task-specific prediction head to predict object category scores (cls), 3D location offsets (x, y, z), object sizes (d_x, d_y, d_z), orientation (θ), and velocities (v_x, v_y).

The Effect of Reference Points. Reference points play a dual role in query-based frameworks: (1) they serve as positional anchors that inject spatial priors into the object queries via positional embeddings, guiding the attention mechanism to reason in the 3D space; (2) they act as geometric anchors during cross-attention, where their projected locations in camera views indicate the regions for image feature aggregation. Thus, the design and distribution of reference points are critical to detection performance. Intuitively, reference points should be encouraged to cluster around object-dense regions to capture richer contextual and geometric information for precise localization. Meanwhile, reducing their presence in object-sparse areas can help suppress false positives and enhance detection robustness.

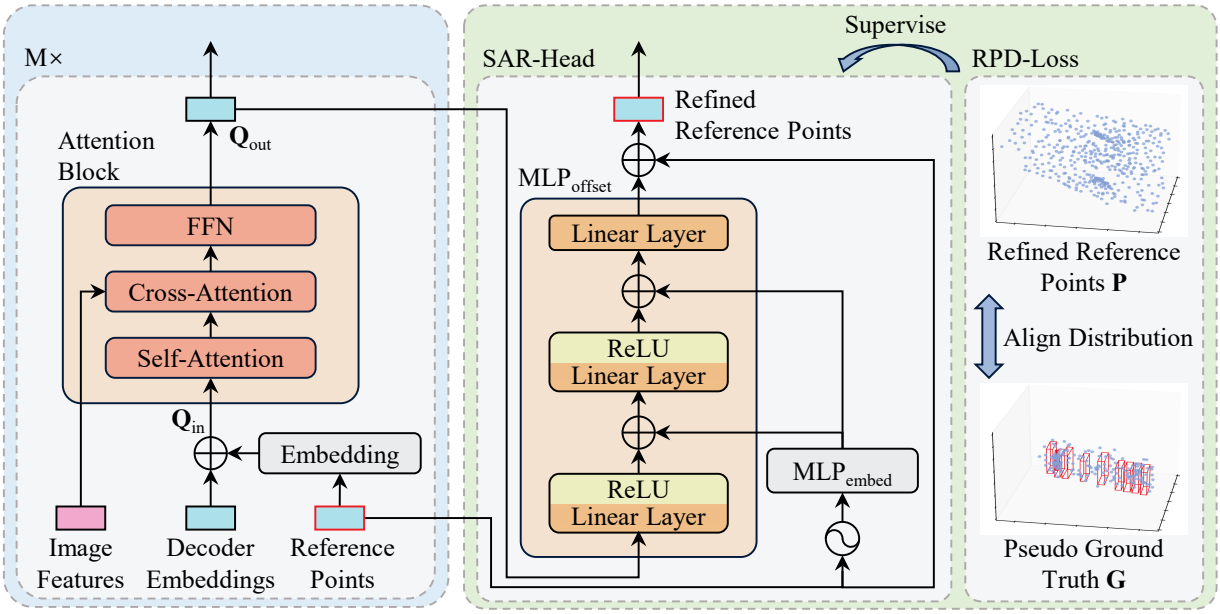
However, in most existing query-based detectors such as PETR (Liu et al. 2022b) and StreamPETR (Wang et al. 2023), reference points are shared across different scenes and remain fixed throughout all transformer decoding layers, limiting their adaptability to diverse scene layouts and object densities. While DETR3D (Wang et al. 2022) and Sparse4Dv3 (Lin et al. 2023c) introduce progressive refinement, only those reference points that are near ground-truth objects receive effective supervision, often leading to suboptimal alignment. These limitations motivate the development of our scene-adaptive refinement mechanism.

Overall Architecture of Refine3D

Refine3D is designed to be integrated into any existing query-based 3D detector. As shown in Fig. 2(b), our method is built upon a standard transformer-based architecture and introduces a plug-and-play refinement mechanism that adaptively adjusts the positions of reference points across layers.

Scene-Adaptive Refinement Head

Previous studies (Lin et al. 2023c) predict reference point offsets using detection heads directly, coupling reference point refinement closely with object localization task. As a result, the movement of reference points is biased toward a small subset of foreground objects. Instead, SAR-Head is specifically designed to handle the offset regression of ref-



(a) Query-based Detectors (b) Refine3D

Figure 2: (a) Architecture of query-based transformer decoder. (b) Proposed Refine3D with Scene-Adaptive Refinement Head (SAR-Head) and Reference Point Distribution Loss (RPD-Loss), that can be easily plugged into existing query-based detectors.

erence points, regardless of whether they are located in the vicinity of objects or background areas.

Specifically, at each attention block $\mathbb{L}^{(l)}$, we first leverage output features $\mathbf{Q}_{\text{out}}^{(l)} \in \mathbb{R}^{N\times D}$ to predict reference point offsets using a lightweight MLP, denoted as $\text{MLP}_{\text{offset}}$:

$$\Delta\mathbf{P}^{(l)} = \text{MLP}_{\text{offset}}^{(l)}(\mathbf{Q}_{\text{out}}^{(l)}) \tag{1}$$

where $\text{MLP}_{\text{offset}}^{(l)}$ consists of two linear layers, each followed by a ReLU (Glorot, Bordes, and Bengio 2011), and a final linear layer. The reference points are then updated via:

$$\mathbf{P}^{(l+1)} = \mathbf{P}^{(l)} + \Delta\mathbf{P}^{(l)} \tag{2}$$

While each query captures local image features, it lacks positional awareness of the global distribution of reference points in 3D space. To address this, we introduce an auxiliary embedding branch to explicitly inject spatial priors into SAR-Head. As shown in Fig. 2(b), we first encode the original reference point coordinates $\mathbf{P}^{(l)}$ with sine-cosine encoding to obtain $\mathbf{e}_{\text{pos}}^{(l)}$, which is then passed through a three-layer MLP with a SiLU (Elfwing, Uchibe, and Doya 2018) activation to produce a position-aware latent embedding:

$$\mathbf{z}_{\text{pos}}^{(l)} = \text{MLP}_{\text{embed}}(\mathbf{e}_{\text{pos}}^{(l)}). \tag{3}$$

where $\text{MLP}_{\text{embed}}$ shares across all layers. The resulting embedding $\mathbf{z}_{\text{pos}}^{(l)}$ is layer-wise combined with each ReLU activation output of $\text{MLP}_{\text{offset}}^{(l)}$ via summation, allowing the regression procedure to differentially process each reference point according to its current location.

By jointly leveraging query features from transformer outputs and spatial priors from reference point coordinates, our SAR-Head can adaptively refine reference points to match varied object distributions under complex real-world environments. We simply replace the original reference points with the refined ones, and the updated query features and reference points are then passed into the next decoder block. This plug-and-play design enables seamless and flexible integration into any existing query-based method. Moreover, both $\text{MLP}_{\text{offset}}$ and $\text{MLP}_{\text{embed}}$ are lightweight, and $\text{MLP}_{\text{embed}}$ shares the same parameters across all decoder blocks, making the SAR-Head highly parameter-efficient.

To fully exploit the potential of our refinement mechanism, the Reference Point Distribution Loss (RPD-Loss) is introduced in the next section, which directly imposes a constraint on the alignment between the reference point distribution and object layouts.

Reference Point Distribution Loss

We propose a novel and effective Reference Point Distribution Loss (RPD-Loss) to provide explicit supervision for the movement of all reference points, encouraging them to converge toward nearby objects, while maintaining diversity and coverage across the 3D scene.

We begin by constructing a pseudo ground-truth set $\mathbf{G}$ that serves as the regression target for reference point refinement. Specifically, for each 3D ground-truth bounding box, we take its geometric center as the mean and define a diagonal Gaussian distribution whose variances along each axis is proportional to the object's length, width, and height. Let there be K ground-truth objects in the current frame, and we sample a total of N points from the K Gaussian distributions. That is, we uniformly select a box indexed by $k \sim \mathcal{U}\{1, 2, \ldots, K\}$, then sample a point from corresponding Gaussian distribution:

$$\mathbf{g}_i \sim \mathcal{N}(\mathbf{c}_k, \text{diag}(l_k^2, w_k^2, h_k^2)), i = 1, \ldots, N \tag{4}$$

where $\mathbf{c}_k$ denotes the center of the selected bounding box, and l_k, w_k, and h_k are its length, width and height. This stochastic sampling strategy yields a spatially dispersed yet object-centric target set $\mathbf{G} = \{\mathbf{g}_i, i = 1, 2, \ldots, N\}$ (shown in Fig. 2(b)), whose cardinality matches the number of reference points, enabling dense and robust supervision for refinement.

In each transformer block $\mathbb{L}^l$, given the refined reference points $\mathbf{P}^{(l)}$ produced by our SAR-Head, the RPD-Loss is computed as:

$$\begin{aligned}\mathcal{L}_{\text{RPD}}(\mathbf{P}^{(l)}, \mathbf{G}) = \alpha \frac{1}{|\mathbf{P}^{(l)}|} \sum_{p \in \mathbf{P}^{(l)}} w_p \min_{g \in \mathbf{G}} ||p - g||_2^2 \\ + \beta \frac{1}{|\mathbf{G}|} \sum_{g \in \mathbf{G}} w_g \min_{p \in \mathbf{P}^{(l)}} ||g - p||_2^2\end{aligned} \tag{5}$$

where $|| \cdot ||_2^2$ denotes the squared Euclidean distance. The first term penalizes predicted points that are far from any ground-truth, while the second ensures all ground-truth points are well covered by predictions. In our implementation, we empirically set hyper-parameters $\alpha : \beta = 4 : 1$. Additionally, since objects close to the ego vehicle are more important in safety-critical driving scenarios, we assign distance-based weights to imposes stronger penalty weights on near-field obstacles. The weight w decays exponentially with its BEV distance to the ego vehicle, *i.e.*, $w_i = exp(-||(x_i, z_i) - (0, 0)||_2^2), i \in \{\mathbf{P}, \mathbf{G}\}$, where x_i and z_i are normalized lateral and longitudinal coordinates of the i-th point in $\mathbf{P}$ or $\mathbf{G}$.

By minimizing $\mathcal{L}_{\text{RPD}}$, the model learns to guide reference points to concentrate around true objects while suppressing activations in the background. This mechanism strengthens spatial supervision, improves query localization, and leads to enhanced detection robustness in complex driving scenes.

Experiment

Dataset and Metrics

We conduct evaluations on two standard autonomous driving benchmarks for 3D object detection: nuScenes (Caesar et al. 2020) and Argoverse 2 (Wilson et al. 2023).

nuScenes consists of 1000 20-second driving sequences with annotations at 2 Hz. It is split into 700 training, 150 validation, and 150 test scenes. Each frame includes six camera views covering a full 360° field of view, with 3D bounding boxes annotated for 10 object categories, totaling 1.4 million instances. Evaluation follows the official nuScenes protocol, which reports mean Average Precision (mAP) along with five true positive metrics: Average Translation Error (ATE), Scale Error (ASE), Orientation Error (AOE), Velocity Error (AVE), and Attribute Error (AAE). The nuScenes Detection Score (NDS) combines these into a single comprehensive performance metric.

Argoverse 2 comprises 1000 driving scenes, each lasting 15 seconds and annotated at 10 Hz. The data is divided into 700 training, 150 validation, and 150 test scenes. Each scene provides seven high-resolution cameras covering a 360° field of view, with annotations spanning 26 object classes and a sensing range up to 150 meters. Performance is evaluated using mean Average Precision (mAP) and the Composite Detection Score (CDS), the latter being a weighted metric that combines three true positive components: Average Translation Error (ATE), Scale Error (ASE), and Orientation Error (AOE). Overall, it contains a relatively larger data scale, longer perception range and more diversified scenarios than nuScenes.

Implementation Details

We instantiate our scene-adaptive reference point refinement mechanism on a representative query-based baseline, *i.e.*, RayDN (Liu et al. 2024). Experiments are conducted using backbone architectures including ResNet50 (He et al. 2016), ResNet101, and V2-99 (Lee et al. 2019), each initialized with different pre-training schemes. Specifically, ResNet50 and ResNet101 are pre-trained on the nuImages dataset (Caesar et al. 2020), and their performance is evaluated on the nuScenes validation set. To demonstrate the scalability of our method, we further report results on the nuScenes test set using V2-99, initialized from the DD3D checkpoint (Park et al. 2021). All models are optimized using the AdamW optimizer (Loshchilov and Hutter 2019) with a batch size of 16 and a base learning rate of 4×10^{-4}, following a cosine annealing schedule. For comparisons with state-of-the-art approaches, models are trained for 60 epochs without using the CBGS augmentation strategy (Zhu et al. 2019), while those in the ablation study are trained for 24 epochs. On the Argoverse 2 dataset, we train all models for 6 epochs under the same optimization settings.

Comparison with State-of-the-Art Methods

We compare the proposed method with other state-of-the-art multi-view 3D object detectors on the nuScenes validation and test sets, as well as the validation set of the Argoverse 2. Our method is evaluated under standard conditions without applying test-time augmentation (TTA), which serves to highlight its intrinsic performance. The experiments aim to assess both the accuracy and generalization capacity of our methods across diverse datasets and conditions.

nuScenes Validation Set. Table 1 presents a comparison against state-of-the-art methods on the nuScenes validation set. We evaluate the performance of our method with both ResNet50 and ResNet101 backbones, which are widely adopted for 3D object detection. Using ResNet50 as the backbone at an input resolution of 256×704, our method achieves 46.9% mAP and 57.4% NDS, surpassing the previous best result from RayDN by 1.1% NDS. When we switch to the more robust ResNet101 backbone and increase the image resolution to 512 × 1408, the performance further improves, reaching 53.9% mAP and 62.4% NDS, showing marginal improvements over the previous state-of-the-art method, Sparse4Dv3, by 0.2% mAP and 0.1% NDS.

nuScenes Test Set. Table 2 provides the results evaluated on the nuScenes test set, assessed through the official test server. For this evaluation, our model is trained on both the training and validation sets. Our method achieves 57.7% mAP and 65.8% NDS, surpassing the baseline, RayDN, by 1.2% mAP and 1.3% NDS, and exceeding the previous state-of-the-art method, Sparse4Dv3, by 0.7% mAP and 0.2%

Methods	Backbone	Image Size	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
PETRv2 (Liu et al. 2023b)	ResNet50	256×704	0.456	0.349	0.700	0.725	0.580	0.437	0.187
BevDet4D (Huang and Huang 2022)	ResNet50	256×704	0.457	0.322	0.703	0.728	0.495	0.354	0.206
BEVDepth (Li et al. 2023b)	ResNet50	256×704	0.475	0.351	0.629	0.667	0.479	0.428	0.198
SOLOFusion (Park et al. 2022)	ResNet50	256×704	0.508	0.427	0.567	0.274	0.511	0.252	**0.181**
StreamPETR (Wang et al. 2023)†	ResNet50	256×704	0.550	0.450	0.613	0.267	0.413	0.265	0.198
SparseBEV (Liu et al. 2023a)	ResNet50	256×704	0.558	0.448	0.581	0.271	**0.373**	0.247	0.190
Sparse4Dv3 (Lin et al. 2023c)†	ResNet50	256×704	0.561	**0.469**	0.553	0.274	0.476	**0.227**	0.200
RayDN (Liu et al. 2024)†	ResNet50	256×704	0.563	**0.469**	0.579	0.264	0.433	0.256	0.187
Ours (with RayDN) †	ResNet50	256×704	**0.574**	**0.469**	**0.547**	**0.261**	0.378	0.231	0.191
PETRv2 (Liu et al. 2023b)	ResNet101	640×1600	0.524	0.421	0.681	0.267	0.357	0.377	0.186
BEVDepth (Li et al. 2023b)	ResNet101	512×1408	0.535	0.412	0.565	0.266	0.358	0.331	0.190
SOLOFusion (Park et al. 2022)	ResNet101	512×1408	0.582	0.483	0.503	0.264	0.381	0.246	0.207
SparseBEV (Liu et al. 2023a)	ResNet101	512×1408	0.592	0.501	0.562	0.265	0.321	0.243	0.195
StreamPETR (Wang et al. 2023)†	ResNet101	512×1408	0.592	0.504	0.569	0.262	0.315	0.257	0.199
Far3D (Jiang et al. 2024)†	ResNet101	512×1408	0.594	0.510	0.551	0.258	0.372	0.238	0.195
Sparse4Dv3 (Lin et al. 2023c)†	ResNet101	512×1408	0.623	0.537	0.511	**0.255**	0.306	**0.194**	0.192
RayDN (Liu et al. 2024)†	ResNet101	512×1408	0.604	0.518	0.541	0.260	0.315	0.236	0.200
Ours (with RayDN)†	ResNet101	512×1408	**0.624**	**0.539**	**0.484**	0.258	**0.293**	0.219	0.202

Table 1: Comparison on the nuScenes validation set. † Indicates methods that benefit from perspective-view pre-training.

Methods	Backbone	Image Size	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
DETR3D (Wang et al. 2022)	V2-99	900×1600	0.479	0.412	0.641	0.255	0.394	0.845	0.133
PETRv2 (Liu et al. 2023b)	V2-99	640×1600	0.582	0.490	0.561	0.243	0.361	0.343	0.120
SparseBEV (Liu et al. 2023a)	V2-99	640×1600	0.627	0.543	0.524	0.244	0.324	0.251	0.126
StreamPETR (Wang et al. 2023)	V2-99	640×1600	0.636	0.550	0.479	0.239	0.317	0.241	0.119
Sparse4Dv3 (Lin et al. 2023c)	V2-99	640×1600	0.656	0.570	**0.412**	0.236	**0.312**	**0.210**	0.117
RayDN (Liu et al. 2024)	V2-99	640×1600	0.645	0.565	0.461	0.241	0.322	0.239	0.114
Ours (with RayDN)	V2-99	640×1600	**0.658**	**0.577**	0.417	**0.231**	0.319	0.224	**0.112**

Table 2: Comparison on the nuScenes test set.

NDS. These gains highlight the strong generalization ability of our approach and underscore the effectiveness of the proposed reference point refinement mechanism in handling unseen driving scenes.

Argoverse 2 Validation Set. As shown in Table 3, our method significantly outperforms the baseline, achieving a remarkable 3.5% mAP improvement and a 2.5% increase in CDS, further demonstrating stronger performance gains across more diverse scenarios and longer perception ranges. We surpass the previous best-performing method, Far3D (Jiang et al. 2024), by 1.5% mAP and 0.5% CDS.

Ablation Study & Analysis

This section presents the ablation studies conducted using the validation set from the nuScenes dataset (Caesar et al. 2020). Unless specified otherwise, the experiments are carried out with a ResNet50 backbone. The model is trained for 24 epochs without employing CBGS.

Effectiveness of Each Component. We perform ablation studies to evaluate the effectiveness of each proposed component and summarize the results in Table 4. Starting from the RayDN baseline, we incrementally introduce our modules under controlled settings. First, integrating the SAR-Head without the position embedding branch yields notable improvements, raising mAP by 3.3% and NDS by 2.3%. This confirms that directly refining reference points with transformer-informed offsets is already beneficial. Adding the embedding branch further enhances performance, contributing an additional 0.5% mAP and 1.6% NDS improvement. The embedding branch captures position-dependent information from the initial reference point coordinates, helping produce more adaptive and accurate offsets. Finally, introducing the proposed RPD-Loss provides consistent gains across all metrics. Compared to the RayDN baseline, our full method achieves 4.6% higher mAP and 4.2% higher NDS, while also reducing most of error metrics. These results validate the effectiveness of explicitly supervising reference point movement and the synergy of all components in enhancing 3D detection accuracy.

Advantages under Imbalanced Object Distributions. To assess the robustness of Refine3D under challenging scenarios, we construct a subset of the nuScenes validation set where the object distribution is highly imbalanced. Specifically, we divide the detection space into four ego-centered quadrants in the BEV plane. A frame is included if the number of objects in one quadrant exceeds that in another by over

Methods	Backbone	Image Size	mAP	CDS	mATE	mASE	mAOE
PETR (Liu et al. 2022b)	V2-99	900×640	0.176	0.122	0.911	0.339	0.819
Sparse4Dv2 (Lin et al. 2023b)	V2-99	900×640	0.189	0.134	0.832	0.343	0.723
StreamPETR (Wang et al. 2023)	V2-99	900×640	0.203	0.146	0.843	0.321	0.650
Far3D (Jiang et al. 2024)	V2-99	900×640	0.244	0.181	0.796	0.304	0.538
RayDN (Liu et al. 2024)	V2-99	900×640	0.223	0.161	0.825	0.325	0.629
Ours (with RayDN)	V2-99	900×640	**0.259**	**0.186**	**0.788**	**0.284**	**0.426**

Table 3: Comparisons on the Argoverse 2 validation set. Performance is evaluated over 26 object categories within a 150-meter perception range.

Methods	NDS↑	mAP↑	mATE↓
Baseline (Liu et al. 2024)	0.509	0.409	0.684
+ $\text{MLP}_{\text{offset}}$	0.532	0.442	0.629
+ $\text{MLP}_{\text{embed}}$ (SAR-Head)	0.548	0.447	0.589
+ RPD-Loss (Ours)	0.551	0.455	0.595

Table 4: Ablation experiments on nuScenes validation set.

Test Set	Methods	NDS↑	mAP↑	mATE↓
Full Set	Baseline	0.563	0.469	0.579
	+Ours	+1.1%	±0.0%	-3.2%
Imbalanced Subset	Baseline	0.537	0.433	0.608
	+Ours	+3.0%	+2.4%	-4.7%

Table 5: Advantages on imbalanced scenes from the nuScenes validation set.

five times, or if any quadrant contains no foreground objects. This results in a harder subset of approximately 2000 frames with imbalanced object layouts. We re-evaluate the detection results obtained using a ResNet50 backbone, on this subset without additional inference. As shown in Table 5, the baseline exhibits a notable performance drop on these imbalanced scenes. In contrast, our method consistently achieves higher relative gains across both the full set and the imbalanced subset, with significantly larger margins in the latter. Specifically, Refine3D outperforms the baseline by 3.0% NDS, 2.4% mAP, and -4.7% mATE on the imbalanced subset. These results underscore the effectiveness of our scene-adaptive refinement strategy in handling localized imbalances. Qualitative comparisons on this subset are provided in the supplementary material.

Cost of Refine3D. We assess the computational overhead of Refine3D by comparing it with the RayDN baseline. The additional cost per block mainly arises from two lightweight MLPs ($\text{MLP}_{\text{offset}}$ and $\text{MLP}_{\text{embed}}$), where $\text{MLP}_{\text{embed}}$ is shared across all layers. This design effectively suppresses the increase in parameter count and memory consumption, preventing noticeable overhead. On the nuScenes dataset with a ResNet50 backbone and an input resolution of 256×704, trained on eight GeForce RTX 4090 GPUs and evaluated on a single RTX 4090 GPU, Refine3D exhibits computational efficiency comparable to RayDN. Specifically, the total training times of Refine3D and RayDN are 13 hours and 11.5 hours, respectively, and the inference speeds reach 10.24 FPS and 11.85 FPS.

Methods	mAP↑	NDS↑	mATE↓
PETR	0.370	0.387	0.753
+Ours	0.387	0.408	0.750
StreamPETR	0.378	0.490	0.678
+Ours	0.437	0.530	0.617
Sparse4Dv3	0.397	0.504	0.610
+Ours	0.442	0.541	0.534

Table 6: Plug-and-play integration and improvements across mainstream detectors.

Plug-and-Play Deployment on Mainstream Detectors. To demonstrate the generalizability and compatibility of our method, we integrate it into several representative query-based 3D detectors, including PETR (Liu et al. 2022b), StreamPETR (Wang et al. 2023), and Sparse4Dv3 (Lin et al. 2023c). These models cover both static and temporally enhanced query-based architectures, providing a comprehensive evaluation of our approach under diverse design paradigms. For each baseline, we retain its original architecture and training settings, simply augmenting them with our proposed scene-adaptive refinement mechanism. As shown in Table 6, our method consistently improves performance across all detectors, with significant gains in both mAP and NDS. In particular, on Sparse4Dv3, which already employs refinement strategy, our method still provides additional gains, by 4.5% mAP and 3.7% NDS.

Conclusion

In this work, we propose Refine3D, a plug-and-play scene-adaptive reference point refinement mechanism for query-based 3D object detection. To guide reference points toward regions with dense object distributions, we introduce a Reference Point Distribution Loss (RPD-Loss) and a Scene-Adaptive Regression Head (SAR-Head), enabling globally informed and spatially adaptive refinement. Extensive results on the nuScenes and Argoverse 2 benchmarks validate the effectiveness and generality of our approach, also showing consistent performance gains when deployed on various mainstream detectors. We believe that our studies can further enhance the robustness of 3D detection when confronting increasingly complex and dynamic driving conditions.

References

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, X.; Zhang, T.; Wang, Y.; Wang, Y.; and Zhao, H. 2023. Futr3D: A unified sensor fusion framework for 3D detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 172–181.

Elfwing, S.; Uchibe, E.; and Doya, K. 2018. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107: 3–11.

Glorot, X.; Bordes, A.; and Bengio, Y. 2011. Deep sparse rectifier neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 315–323. JMLR Workshop and Conference Proceedings.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Huang, J.; and Huang, G. 2022. BEVDet4D: Exploit temporal cues in multi-camera 3D object detection. arXiv:2203.17054.

Huang, J.; Huang, G.; Zhu, Z.; Ye, Y.; and Du, D. 2022. BEVDet: High-performance multi-camera 3D object detection in bird-eye-view. arXiv:2112.11790.

Jiang, X.; Li, S.; Liu, Y.; Wang, S.; Jia, F.; Wang, T.; Han, L.; and Zhang, X. 2024. Far3D: Expanding the horizon for surround-view 3D object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2561–2569.

Jiao, Y.; Jie, Z.; Chen, S.; Cheng, L.; Chen, J.; Ma, L.; and Jiang, Y.-G. 2024. Instance-aware multi-camera 3D object detection with structural priors mining and self-boosting learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2598–2606.

Lee, Y.; Hwang, J.-w.; Lee, S.; Bae, Y.; and Park, J. 2019. An energy and GPU-computation efficient backbone network for real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

Li, P.; Shen, W.; Huang, Q.; and Cui, D. 2024. DualBEV: Unifying dual view transformation with probabilistic correspondences. In *Proceedings of the European Conference on Computer Vision*, 286–302.

Li, Y.; Bao, H.; Ge, Z.; Yang, J.; Sun, J.; and Li, Z. 2023a. BEVStereo: Enhancing depth estimation in multi-view 3D object detection with temporal stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1486–1494.

Li, Y.; Ge, Z.; Yu, G.; Yang, J.; Wang, Z.; Shi, Y.; Sun, J.; and Li, Z. 2023b. BEVDepth: Acquisition of reliable depth for multi-view 3D object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1477–1485.

Lin, X.; Lin, T.; Pei, Z.; Huang, L.; and Su, Z. 2023a. Sparse4D: Multi-view 3D object detection with sparse spatial-temporal fusion. arXiv:2211.10581.

Lin, X.; Lin, T.; Pei, Z.; Huang, L.; and Su, Z. 2023b. Sparse4D v2: Recurrent temporal fusion with sparse model. arXiv:2305.14018.

Lin, X.; Pei, Z.; Lin, T.; Huang, L.; and Su, Z. 2023c. Sparse4D v3: Advancing end-to-end 3D detection and tracking. arXiv:2311.11722.

Liu, F.; Huang, T.; Zhang, Q.; Yao, H.; Zhang, C.; Wan, F.; Ye, Q.; and Zhou, Y. 2024. Ray Denoising: Depth-aware hard negative sampling for multi-view 3D object detection. In *Proceedings of the European Conference on Computer Vision*, 200–217.

Liu, H.; Teng, Y.; Lu, T.; Wang, H.; and Wang, L. 2023a. SparseBEV: High-performance sparse 3D object detection from multi-camera videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18580–18590.

Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; and Zhang, L. 2022a. DAB-DETR: Dynamic anchor boxes are better queries for DETR. arXiv:2201.12329.

Liu, Y.; Wang, T.; Zhang, X.; and Sun, J. 2022b. PETR: Position embedding transformation for multi-view 3D object detection. In *Proceedings of the European Conference on Computer Vision*, 531–548.

Liu, Y.; Yan, J.; Jia, F.; Li, S.; Gao, A.; Wang, T.; and Zhang, X. 2023b. PETRv2: A unified framework for 3D perception from multi-camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3262–3272.

Loshchilov, I.; and Hutter, F. 2019. Decoupled weight decay regularization. arXiv:1711.05101.

Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; and Wang, J. 2021. Conditional DETR for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3651–3660.

Park, D.; Ambrus, R.; Guizilini, V.; Li, J.; and Gaidon, A. 2021. Is pseudo-lidar needed for monocular 3D object detection? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3142–3152.

Park, J.; Xu, C.; Yang, S.; Keutzer, K.; Kitani, K.; Tomizuka, M.; and Zhan, W. 2022. Time Will Tell: New outlooks and a baseline for temporal multi-View 3D object detection. arXiv:2210.02443.

Wang, S.; Liu, Y.; Wang, T.; Li, Y.; and Zhang, X. 2023. Exploring object-centric temporal modeling for efficient multi-view 3D object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3621–3631.

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 *Proceedings of the Conference on Robot Learning*, 180–191.

Wilson, B.; Qi, W.; Agarwal, T.; Lambert, J.; Singh, J.; Khandelwal, S.; Pan, B.; Kumar, R.; Hartnett, A.; Pontes, J. K.; Ramanan, D.; Carr, P.; and Hays, J. 2023. Argoverse 2: Next generation datasets for self-driving perception and forecasting. arXiv:2301.00493.

Yang, C.; Chen, Y.; Tian, H.; Tao, C.; Zhu, X.; Zhang, Z.; Huang, G.; Li, H.; Qiao, Y.; Lu, L.; et al. 2023. BEVFormer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17830–17839.

Zhu, B.; Jiang, Z.; Zhou, X.; Li, Z.; and Yu, G. 2019. Class-balanced grouping and sampling for point cloud 3D object detection. arXiv:1908.09492.