The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)

Far3D: Expanding the Horizon for Surround-View 3D Object Detection

Xiaohui Jiang *[1†], **Shuailin Li** *[2], **Yingfei Liu**[2], **Shihao Wang**[1†], **Fan Jia**[2], **Tiancai Wang**[2‡], **Lijin Han**[1], **Xiangyu Zhang**[2]

[1]Beijing Institute of Technology
[2]MEGVII Technology

Abstract

Recently 3D object detection from surround-view images has made notable advancements with its low deployment cost. However, most works have primarily focused on close perception range while leaving long-range detection less explored. Expanding existing methods directly to cover long distances poses challenges such as heavy computation costs and unstable convergence. To address these limitations, this paper proposes a novel sparse query-based framework, dubbed Far3D. By utilizing high-quality 2D object priors, we generate 3D adaptive queries that complement the 3D global queries. To efficiently capture discriminative features across different views and scales for long-range objects, we introduce a perspective-aware aggregation module. Additionally, we propose a range-modulated 3D denoising approach to address query error propagation and mitigate convergence issues in long-range tasks. Significantly, Far3D demonstrates SoTA performance on the challenging Argoverse 2 dataset, covering a wide range of 150 meters, surpassing several LiDAR-based approaches. The code is available at https://github.com/megvii-research/Far3D.

Introduction

3D object detection plays an important role in understanding 3D scenes, aiming to provide accurate object localization and category around the ego vehicle. Surround-view methods (Huang and Huang 2022; Li et al. 2023; Liu et al. 2022b; Li et al. 2022c; Yang et al. 2023; Park et al. 2022; Wang et al. 2023a), with their advantages of low cost and wide applicability, have achieved remarkable progress. However, most of them focus on close-range perception (e.g., $\sim$50 meters on nuScenes (Caesar et al. 2020)), leaving the long-range detection field less explored. Detecting distant objects is essential for real-world driving to maintain a safe distance, especially at high speeds or complex road conditions.

Existing surround-view methods can be categorized into two groups based on the intermediate representation, dense Bird's-Eye-View (BEV) based methods and sparse query-based methods. BEV based methods (Huang et al. 2021; Huang and Huang 2022; Li et al. 2023, 2022c; Yang et al. 2023) usually convert perspective features to BEV features by employing a view transformer (Philion and Fidler 2020), then utilizing a 3D detector head to produce the 3D bounding boxes. However, dense BEV features come at the cost of high computation even for the close-range perception, making it more difficult to scale up to long-range perception. Instead, following DETR (Carion et al. 2020) style, sparse query-based methods (Wang et al. 2022; Liu et al. 2022a,b; Wang et al. 2023a) adopt learnable global queries to represent 3D objects, and interact with surround-view image features to update queries. Although sparse design can avoid the squared growth of query numbers, its global fixed queries cannot adapt to dynamic scenarios and usually miss targets in long-range detection. We adopt the sparse query design to maintain detection efficiency and introduce 3D adaptive queries to address the inflexibility weaknesses.

(a) 3D predicted boxes (b) 2D predicted boxes

Recall	0-50m	50-100m	100-150m	Average
R@3D	0.46	0.17	0.06	0.32
R@2D	0.92	0.68	0.46	0.78

(c) Recall comparison

Figure 1: Peformance comparisons on Argoverse 2 between 3D detection and 2D detection. (a) and (b) demonstrate predicted boxes of StreamPETR and YOLOX, respectively. (c) imply that 2D recall is notably better than 3D recall and can act as a bridge to achieve high-quality 3D detection. Note that 2D recall does not represent 3D upper bound due to different recall criteria.

*Equal contribution.
†Work done during the internship at MEGVII Technology.
‡Corresponding author

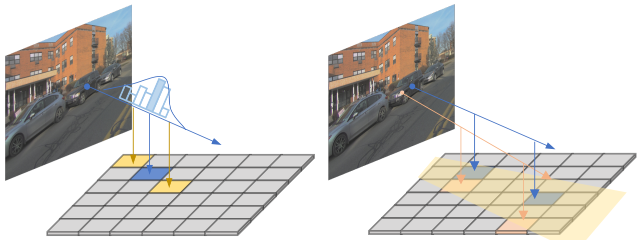

Figure 2: Different cases of transformimg 2D points into 3D space. The blue dots indicate the centers of 3D objects in images. (left) shows the redundant prediction with the wrong depth, which is in yellow. (right) illustrates the error propagation problem dominated by different ranges.

To employ the sparse query-based paradigm for long-range detection, the primary challenge lies in poor recall performance. Due to the query sparsity in 3D space, assignments between predictions and ground-truth objects are affected, generating only a small amount of matched positive queries. As illustrated in Fig. 1, 3D detector recalls are pretty low, yet recalls from the existing 2D detector are much higher, showing a significant performance gap between them. Motivated by this, leveraging high-quality 2D object priors to improve 3D proposals is a promising approach, for enabling accurate localization and comprehensive coverage. Although previous methods like SimMOD (Zhang et al. 2023) and MV2D (Wang et al. 2023b) have explored using 2D predictions to initialize 3D object proposals, they primarily focus on close-range tasks and discard learnable object queries. Moreover, as depicted in Fig. 2, directly introducing 3D queries derived from 2D proposals for long-range tasks encounters two issues: 1) inferior redundant predictions due to uncertain depth distribution along the object rays, and 2) larger deviations in 3D space as the range increases due to frustum transformation. These noisy queries can impact the training stability, requiring effective denoising ways to optimize. Furthermore, within the training process, the model exhibits a tendency to overfit on densely populated close objects while disregarding sparsely distributed distant objects.

To address the aforementioned challenges, we design a novel 3D detection paradigm to expand the perception horizon. Despite the 3D global query that was learned from the dataset, our approach also incorporates auxiliary 2D proposals into 3D adaptive query generation. Specifically, we first produce reliable pairs of 2D object proposals and corresponding depths then project them to 3D proposals via spatial transformation. We compose 3D adaptive queries with the projected positional embedding and semantic context, which would be refined in the subsequent decoder. In the decoder layers, perspective-aware aggregation is employed across different image scales and views. It learns sampling offsets for each query and dynamically enables interactions with favorable features. For instance, distant object queries are beneficial to attend large-resolution features, while the opposite is better for close objects in order to capture high-level context. Lastly, we design a range-modulated 3D denoising technique to mitigate query error propagation and slow convergence. Considering the different regression difficulties for various ranges, noisy queries are constructed based on ground-truth (GT) as well as referring to their distances and scales. Our method feeds multi-group noisy proposals around GT into the decoder and trains the model to a) recover 3D GT for positive ones and b) reject negative ones, respectively. The inclusion of query denoising also alleviates the problem of range-level unbalanced distribution.

Our proposed method achieves remarkable performance advancements over state-of-the-art (SoTA) approaches on challenging long-range Argoverse 2 dataset, and surpasses the prior arts of LiDAR-based methods. To evaluate the generalization capability, we further validate its results on nuScenes and demonstrate SoTA metrics. In summary:

- We propose a novel sparse query-based framework to expand the perception range in 3D detection, by incorporating high-quality 2D object priors into 3D adaptive queries.
- We develop perspective-aware aggregation that captures informative features from diverse scales and views, as well as a range-modulated 3D denoising technique to address query error propagation and convergence problems.
- On the challenging long-range Argoverse 2 datasets, our method surpasses surround-view methods and outperforms several LiDAR-based methods. The generalization of our method is validated on the nuScenes dataset.

Related Work

Surround-view 3D Object Detection

Recently 3D object detection from surround-view images has attracted much attention and achieved great progress, due to its advantages of low deployment cost and rich semantic information. Based on feature representation, existing methods (Wang et al. 2021, 2022; Liu et al. 2022a; Huang and Huang 2022; Li et al. 2023, 2022b; Jiang et al. 2023; Liu et al. 2022b; Li et al. 2022c; Yang et al. 2023; Park et al. 2022; Wang et al. 2023a; Zong et al. 2023; Liu et al. 2023) can be largely classified into BEV-based methods and sparse-query based methods.

Extracting image features from surround views, BEV-based methods (Huang et al. 2021; Huang and Huang 2022; Li et al. 2023, 2022c) transform features into BEV space by leveraging estimated depths or attention layers, then a 3D detector head is employed to predict localization and other properties of 3D objects. For instance, BEVFormer (Li et al. 2022c) leverages both spatial and temporal features by interacting with spatial and temporal space through predefined grid-shaped BEV queries. BEVDepth (Li et al. 2023) propose a 3D detector with a trustworthy depth estimation, by introducing a camera-aware depth estimation module. On the other hand, sparse query-based paradigms (Wang et al. 2022; Liu et al. 2022a) learn global object queries from the representative data, then feed them into the decoder to predict 3D bounding boxes during inference. This line of work has the advantage of lightweight computing.

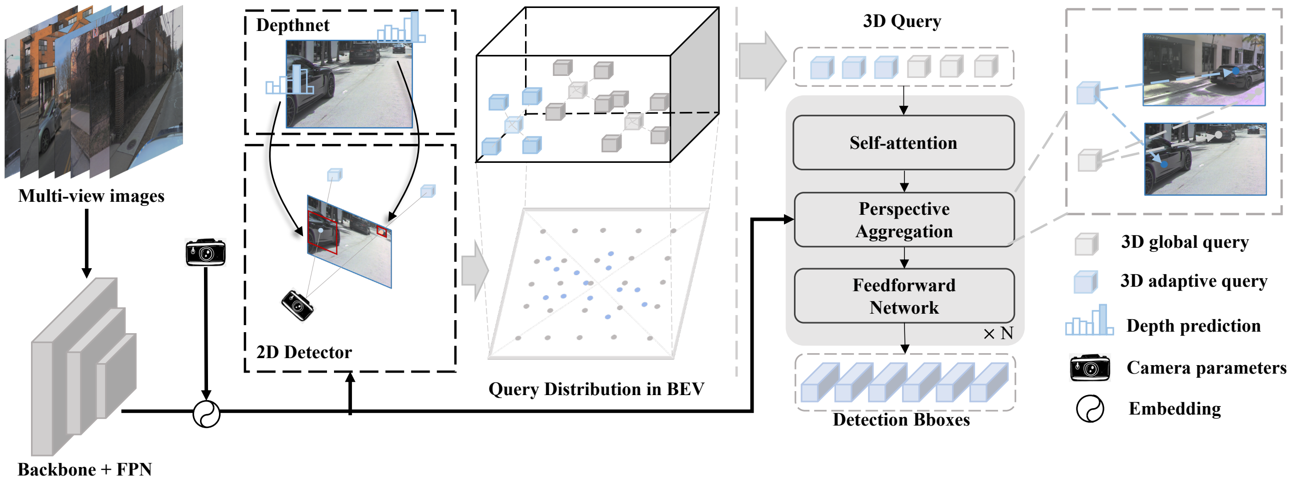

Figure 3: The overview of our proposed Far3D. Feeding surround-view images into the backbone and FPN neck, we obtain 2D image features and encode them with camera parameters for perspective-aware transformation. Utilizing a 2D detector and DepthNet, we generate reliable 2D box proposals and their corresponding depths, which are then concatenated and projected into 3D space. The generated 3D adaptive queries, combined with the initial 3D global queries, are iteratively refined by the decoder layers to predict 3D bounding boxes. Furthermore, temporal modeling is equipped through long-term query propagation.

Furthermore, temporal modeling for surround-view 3D detection can improve detection performance and decrease velocity errors significantly, and many works (Huang and Huang 2022; Liu et al. 2022b; Park et al. 2022; Wang et al. 2023a; Lin et al. 2022, 2023) aim to extend a single-frame framework to multi-frame design. BEVDet4D (Huang and Huang 2022) lifts the BEVDet paradigm from the spatial-only 3D space to the spatial-temporal 4D space, via fusing features with the previous frame. PETRv2 (Liu et al. 2022b) extends the 3D position embedding in PETR for temporal modeling through the temporal alignment of different frames. However, they use only limited history. To leverage both short-term and long-term history, SOLOFusion (Park et al. 2022) balances the impacts of spatial resolution and temporal difference on localization potential, then use it to design a powerful temporal 3D detector. StreamPETR (Wang et al. 2023a) develops an object-centric temporal mechanism in an online manner, where long-term historical information is propagated through object queries.

2D Auxiliary Tasks for 3D Detection

3D detection from surround-view images can be improved through 2D auxiliary tasks, and some works (Xie et al. 2022; Zhang et al. 2023; Wang, Jiang, and Li 2022; Yang et al. 2023; Wang et al. 2023b) aim to exploit its potential. There are several approaches including 2D pertaining, auxiliary supervision, and proposal generation. SimMOD (Zhang et al. 2023) exploits sample-wise object proposals and designs a two-stage training manner, where perspective object proposals are generated and followed by iterative refinement in DETR3D-style. Focal-PETR (Wang, Jiang, and Li 2022) performs 2D object supervision to adaptively focus the attention of 3D queries on discriminative foreground regions. BEVFormerV2 (Yang et al. 2023) presents a two-stage BEV detector where perspective proposals are fed into the BEV head for final predictions. MV2D (Wang et al. 2023b) designs a 3D detector head that is initialized by RoI regions of 2D predicted proposals.

Compared to the above methods, our framework differs in the following aspects. Firstly, we aim to resolve the challenges of long-range detection with surrounding views, which are less explored in previous methods. Besides learning 3D global queries, we explicitly leverage 2D predicted boxes and depths to build 3D adaptive queries, utilizing positional prior and semantic context simultaneously. Furthermore, the designs of perspective-aware aggregation and 3D denoising are integrated to address task issues.

Method

Overview

Fig. 3 shows the overall pipeline of our sparse query-based framework. Feeding surround-view images $\mathbf{I} = \{\mathbf{I^1}, ..., \mathbf{I^n}\}$, we extract multi-level images features $\mathbf{F} = \{\mathbf{F^1}, ..., \mathbf{F^n}\}$ by using the backbone network (e.g. ResNet, ViT) and a FPN (Lin et al. 2017) neck. To generate 3D adaptive queries, we first obtain 2D proposals and depths using a 2D detector head and depth network, then filter reliable ones and transform them into 3D space to generate 3D object queries. In this way, informative object priors from 2D detections are encoded into the 3D adaptive queries.

In the 3D detector head, we concatenate 3D adaptive queries and 3D global queries, then input them to transformer decoder layers including self-attention among queries and perspective-aware aggregation between queries and features. We propose perspective-aware aggregation to efficiently capture rich features in multiple views and scales by considering the projection of 3D objects. Besides, range-modulated 3D denoising is introduced to alleviate query error propagation and stabilize the convergence, when training with long-range and imbalanced distributed objects. Sec depicts the denoising technique in detail.

Adaptive Query Generation

Directly extend existing 3D detectors from short range (e.g. ~50m) to long range (e.g. ~150m) suffers from several problems: heavy computation costs, inefficient convergence and declining localization ability. For instance, the query number is supposed to grow at least squarely to cover possible objects in a larger range, yet such a computing disaster is unacceptable in realistic scenarios. Besides that, small and sparse distant objects would hinder the convergence and even hurt the localization of close objects. Motivated by the high performance of 2D proposals, we propose to generate adaptive queries as objects prior to assist 3D localization. This paradigm compensates for the weakness of global fixed query design and allows the detector to generate adaptive queries near the ground-truth (GT) boxes for different images. In this way, the model is equipped with better generalization and practicality.

Specifically, given image features after FPN neck, we feed them into the anchor-free detector head from YOLOX (Ge et al. 2021) and a light-weighted depth estimation net, outputting 2D box coordinates, scores and depth map. 2D detector head follows the original design, while the depth estimation is regarded as a classification task by discretizing the depth into bins (Reading et al. 2021; Zhang et al. 2022). We then make pairs of 2D boxes and corresponding depths. To avoid the interference of low-quality proposals, we set a score threshold τ (e.g. 0.1) to leave only reliable ones. For each view i, box centers $(\mathbf{c}_w, \mathbf{c}_h)$ from 2D predictions and depth $\mathbf{d}_{wh}$ from depth map are combined and projected to 3D proposal centers $\mathbf{c}_{3d}$.

$$\mathbf{c_{3d}} = K_i^{-1} I_i^{-1} [\mathbf{c}_w * \mathbf{d}_{wh}, \mathbf{c}_h * \mathbf{d}_{wh}, \mathbf{d}_{wh}, \mathbf{1}]^T \quad (1)$$

where K_i, I_i denote camera extrinsic and intrinsic matrices.

After obtaining projected 3D proposals, we encode them into 3D adaptive queries as follows,

$$\mathbf{Q}_{pos} = PosEmbed(\mathbf{c}_{3d}) \quad (2)$$

$$\mathbf{Q}_{sem} = SemEmbed(\mathbf{z}_{2d}, \mathbf{s}_{2d}) \quad (3)$$

$$\mathbf{Q} = \mathbf{Q}_{pos} + \mathbf{Q}_{sem} \quad (4)$$

where $\mathbf{Q}_{pos}, \mathbf{Q}_{sem}$ denote positional embedding and semantic embedding, respectively. $\mathbf{z}_{2d}$ sampled from $\mathbf{F}$ corresponds to the semantic context of position $(\mathbf{c}_w, \mathbf{c}_h)$, and $\mathbf{s}_{2d}$ is the confidence score of 2D boxes. $PosEmbed(\cdot)$ consists of a sinusoidal transformation (Vaswani et al. 2017) and a MLP, while $SemEmbed(\cdot)$ is another MLP.

Lastly, the proposed 3D adaptive queries are concatenated with initialized global queries, and fed to subsequent transformer layers in the decoder.

Perspective-aware Aggregation

Existing sparse query-based approaches usually adopt one single-level feature map for computation effectiveness (e.g. StreamPETR). However, the single feature level is not optimal for all object queries of different ranges. For example, small distant objects require large-resolution features for precise localization, while high-level features are better suited for large close objects. To overcome the limitation, we propose perspective-aware aggregation, enabling efficient feature interactions on different scales and views.

Inspired by the deformable attention mechanism (Zhu et al. 2020), we apply a 3D spatial deformable attention consisting of 3D offsets sampling followed by view transformation. Formally, we first equip image features $\mathbf{F}$ with the camera information including intrinsic $\mathbf{I}$ and extrinsic parameters $\mathbf{K}$. A squeeze-and-excitation block (Hu, Shen, and Sun 2018) is used to explicitly enrich the features. Given enhanced feature $\mathbf{F}'$, we employ 3D deformable attention instead of global attention in PETR series (Liu et al. 2022a,b; Wang et al. 2023a). For each query reference point in 3D space, the model learns M sampling offsets around and projects these references into different 2D scales and views.

$$\mathbf{P}_q^{2d} = \mathbf{I} \cdot \mathbf{K} \cdot (\mathbf{P}_q^{3d} + \Delta\mathbf{P}_q^{3d}) \quad (5)$$

where $\mathbf{P}_q^{3d}, \Delta\mathbf{P}_q^{3d}$ are 3D reference point and learned offsets for query q, respectively. $\mathbf{P}_q^{2d}$ stands for the projected 2d reference point of different scales and views. For simplicity, we omit the subscripts of scales and views.

Next, 3D object queries interact with multi-scale sampled features from $\mathbf{F}^{'}$, according to the above 2D reference points $\mathbf{P}_q^{2d}$. Then features from various views and scales are aggregated into 3D queries considering their relative importance.

Range-modulated 3D Denoising

3D object queries at different distances have different regression difficulties, which is different from 2D queries that are usually treated equally for existing 2D denoising methods such as DN-DETR (Li et al. 2022a). The difficulty discrepancy comes from query density and error propagation. On the one hand, queries corresponding to distant objects are less matched compared to close ones. On the other hand, small errors of 2D proposals can be amplified when introducing 2D priors to 3D adaptive queries, illustrated in Fig. 2, not to mention which effect increases along with object distance. As a result, some query proposals near GT boxes can be regarded as noisy candidates, whereas others with notable deviation should be negative ones. Therefore we aim to recall those potential positive ones and directly reject solid negative ones, by developing a method called range-modulated 3D denoising.

Concretely, we construct noisy queries based on GT objects by simultaneously adding positive and negative groups. For both types, random noises are applied according to object positions and sizes to facilitate denoising learning in long-range perception. Formally, we define the position of noisy queries as:

$$\tilde{\mathbf{P}} = \mathbf{P}_{GT} + \alpha f_p(\mathbf{S}_{GT}) + (1-\alpha) f_n(\mathbf{P}_{GT}) \quad (6)$$

where $\alpha \in \{0, 1\}$ corresponds to the generation of negative and positive queries, respectively. $\mathbf{P}_{GT}, \mathbf{S}_{GT} \in \mathbb{R}^3$ represents 3D center (x, y, z) and box scale (w, l, h) of GT, and $\tilde{\mathbf{P}}$ is noisy coordinates. We use functions f_p and f_n to encode position-aware noise for positive and negative samples.

Methods	Backbone	Modality	Image/Voxel Size	mAP↑	CDS↑	mATE↓	mASE↓	mAOE↓
BEVStereo‡	VoV-99	Camera	960 × 640	0.146	0.104	0.847	0.397	0.901
SOLOFusion‡	VoV-99	Camera	960 × 640	0.149	0.106	0.934	0.425	0.779
PETR	VoV-99	Camera	960 × 640	0.176	0.122	0.911	0.339	0.819
Sparse4Dv2	VoV-99	Camera	960 × 640	0.189	0.134	0.832	0.343	0.723
StreamPETR	VoV-99	Camera	960 × 640	0.203	0.146	0.843	0.321	0.650
Far3D (Ours)	VoV-99	Camera	960 × 640	**0.244**	**0.181**	**0.796**	**0.304**	**0.538**
CenterPoint	-	Lidar	(0.2, 0.2, 0.2)	0.274	0.210	0.548	0.362	0.781
FSD	-	Lidar	(0.2, 0.2, 0.2)	0.291	0.233	**0.468**	**0.299**	0.740
VoxelNeXt	-	Lidar	(0.1, 0.1, 0.2)	0.307	0.225	0.431	0.291	1.157
Far3D (Ours)	ViT-L	Camera	1536 × 1536	**0.316**	**0.239**	0.732	0.303	**0.459**

Table 1: Comparisons on the Argoverse 2 `val` set. We evaluate 26 object categories with a range of 150 meters. Far3D outperform previous surround-view methods with a large margin, and surpass several SoTA LiDAR-based methods. Surround-view methods except for PETR are with temporal modeling. ‡ are reproduced by ourselves.

Methods	Backbone	Split	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
PETR	ResNet101	`val`	0.366	0.441	0.717	0.261	0.412	0.834	0.190
SOLOFusion	ResNet101	`val`	0.483	0.582	**0.503**	0.264	0.381	0.246	0.207
StreamPETR*	ResNet101	`val`	0.504	0.592	0.569	0.262	**0.315**	0.257	0.199
Sparse4Dv2*	ResNet101	`val`	0.505	0.594	0.548	0.268	0.348	0.239	**0.184**
Far3D (Ours)*	ResNet101	`val`	**0.510**	**0.594**	0.551	**0.258**	0.372	**0.238**	0.195
SOLOFusion	ConvNeXt-B	`test`	0.540	0.619	0.453	0.257	0.376	0.267	0.148
Sparse4Dv2	VoV-99	`test`	0.556	0.638	0.462	0.238	0.328	0.264	**0.115**
StreamPETR	ViT-L	`test`	0.620	0.676	0.470	0.241	**0.258**	0.236	0.134
Far3D (Ours)	ViT-L	`test`	**0.635**	**0.687**	**0.432**	**0.237**	0.278	**0.227**	0.130

Table 2: Comparison on the nuScenes `val` and `test` splits. Far3D achieves the highest performance compared to prior-arts, validating its generalization ability. *Benefited from the perspective-view pre-training. We employ the resolution 512 × 1408 for `val` and 1536 × 1536 for `test` split.

For positive noisy samples, we set $f_p(\mathbf{S}_{GT})$ as a linear function of 3D box scale with a random variable. We incorporate the offset constraint within GT boxes to guide the model in accurately reconstructing the GT from positive queries, while ensuring clear distinction from surrounding adjacent boxes. For negative samples, the offsets are supposed to be relevant to their position range, thus we propose several implementations. For some examples, $f_n(\mathbf{P}_{GT})$ can be in forms of $log(\mathbf{P}_{GT})$, $\lambda_2\mathbf{P}_{GT}$ or $\sqrt{\mathbf{P}_{GT}}$. We show these attempts in Sec. . Moreover, multi-group samples are generated for each GT object to enhance query diversity. Each group comprises one positive sample and K negative samples. This approach serves as an imitation of noisy positive candidates and false positive candidates during training.

Experiment

Datasets and Metrics

We use the large-scale Argoverse 2 dataset (Wilson et al. 2023) and nuScenes dataset (Caesar et al. 2020) to explore and evaluate the effectiveness of our approach.

Argoverse 2 is a dataset for perception and prediction studies in autonomous driving domain. It contains 1000 scenes with 15 seconds duration and 10Hz annotation frequency. And these total scenes are divided into 700 for training, 150 for validation, and 150 for testing. Seven high-resolution ring cameras are provided with a combined 360° field of view. We evaluate it with 26 categories in a 150-meter range, satisfying the need for long-range tasks. In addition to the mean Average Precision (mAP), we evaluate the methods with the metrics that Argoverse 2 proposed: the Composite Detection Score (CDS), which is the main metric combining all factors in Argoverse 2, and three true positive metrics, including ATE, ASE, and AOE.

nuScenes is one of the most trustworthy datasets for multi-camera 3D object detection containing 1000 driving scenes in total. Each scene, approximately 20 seconds long, is annotated in 10 categories with 3D bounding boxes for sampled keyframes. We conduct experiments on it and compare the results with other methods using the following metrics: mAP and the nuScenes Detection Score (NDS).

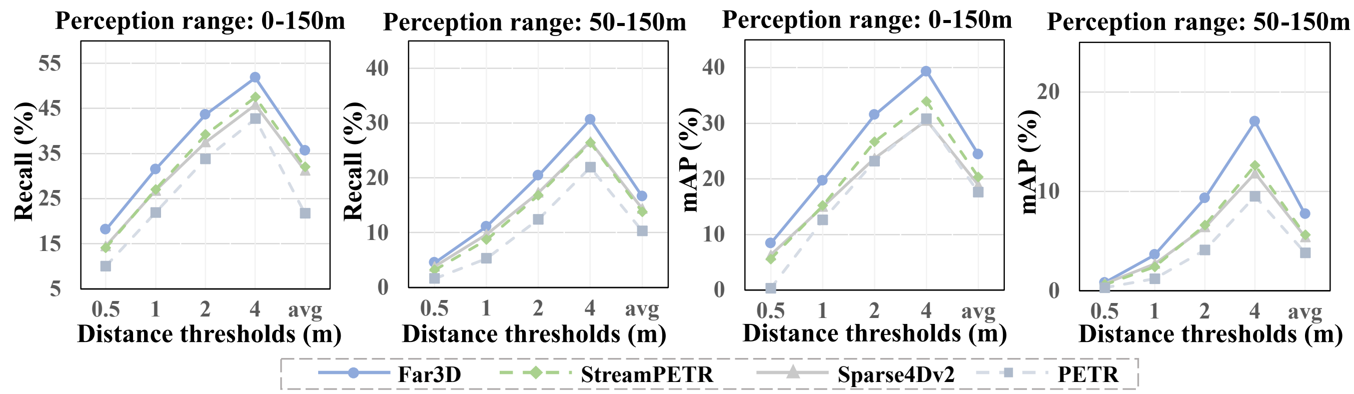

Figure 4: 3D Recall and AP of each method with different distance thresholds. Metrics of different ranges show that our approach consistently achieves a better result.

#	Adaptive Query	PA	3D Denoising	mAP[%]↑	CDS[%]↑
1				20.3	14.6
2	✔			22.4	16.1
3	✔	✔		23.4	17.3
4	✔	✔	✔	**24.4** (+4.1)	**18.1** (+3.5)

Table 3: Ablation of our components on Argoverse 2 `val` set. StreamPETR is employed as the baseline, and we add the adaptive query, perspective-aware aggregation (PA) and range-modulated 3D denoising in order.

τ	mAP[%]↑	CDS[%]↑	mATE↓	mASE↓	mAOE↓
0.01	23.1	17.2	0.807	0.307	0.531
0.05	23.4	17.3	0.806	0.312	0.531
0.1	**24.4**	**18.1**	**0.796**	**0.304**	0.538
0.2	23.7	17.6	0.802	0.307	**0.530**
0.3	23.5	17.4	0.799	0.307	0.577

Table 4: Ablation of different threshold τ for 2D proposals.

Implementation Details

With StreamPETR (Wang et al. 2023a) as our baseline, Far3D is composed of a backbone, an FPN neck, a 2D proposal head, and a 3D detection head. We adopt VoVNet-99 (Lee et al. 2019) pre-trained with FCOS3D (Wang et al. 2021) on nuScenes as the backbone to conduct main experiments. ViT-Large (Dosovitskiy et al. 2020) pre-trained by Objects365 (Shao et al. 2019) and COCO (Lin et al. 2014) dataset is used to scale up our model. By default, the FPN gives 4-level feature maps with sizes of 1/8, 1/16, 1/32, and 1/64. The perception range is set as 152.4m $\times$ 152.4m.

We use AdamW (Loshchilov and Hutter 2017) optimizer with a weight decay of 0.01. The total batch size is 8 and the learning rate is set to 2e-4. The models are totally trained for 6 epochs, following the previous method (Chen et al. 2023). Since the resolution of the front-view image is different from other views in Argoverse 2 dataset, we first resize the front image to a consistent resolution, then do the same image data augmentation as other images do. We do not use any BEV data augmentation on Argoverse 2 dataset. On the nuScenes dataset, we set the batch size as 32 and use the ResNet101 (He et al. 2016) backbone to train our method for 60 epochs. Other settings keep in line with StreamPETR.

Main Results

Argoverse 2 Dataset. We compare the proposed framework with the existing state-of-the-arts on Argoverse 2 `val` set. As shown in Tab. 1, when adopting VoV-99 backbone and 960$\times$640 input size, our method demonstrates a substantial superiority over other methods, achieving an impressive margin of 4.1% mAP and 3.5% CDS. Besides the listed sparse query-based methods, we also conduct experiments on dense BEV-based methods, BEVStereo (Li et al. 2022b) and SOLOFusion (Park et al. 2022). The results are barely satisfactory and we suppose that is because of the greater difficulty of depth estimation. We also reproduce MV2D (Wang et al. 2023b) but it can hardly converge here. The reason is mainly the generated anchors lack accurate depth estimation, leading to large localization deviations over long distances. To sum up, the convergence problem in long-range detection is severe for the above methods, and we believe that our depth estimation and 3D denoising play key roles to solve it. More explanations are in the supplementary.

We further compare it with LiDAR-based SoTAs, CenterPoint (Yin, Zhou, and Krahenbuhl 2021), FSD (Fan et al. 2022), and VoxelNeXt (Chen et al. 2023). With a ViT-L backbone and 1536$\times$1536 resolution, our method outperforms them, showcasing the great potential of surround-view methods. In detail, LiDAR-based methods have a lower localization error (i.e. ATE) due to accurate depth information, while surround-view ones identify orientation properties (i.e. AOE) better.

As shown in Fig. 4, we present the 3D recall and mAP results with different distances of 0-150m and 50-150m. Far3D consistently outperforms other methods. For distant objects, Far3D has a greater improvement when comparing recall and mAP with thresholds of 2m and 4m.

nuScenes Dataset. To evaluate the generalization ability

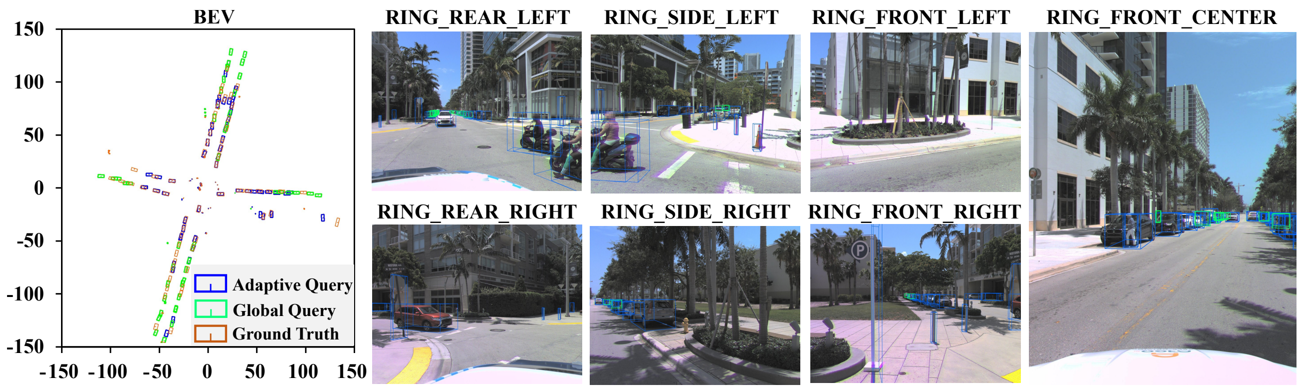

Figure 5: Visualization results on Argoverse 2 dataset. We show 3D bounding boxes predicted both in multi-camera images and bird's eye view. The view of the front center is distinguished from other six views. The detection boxes predicted from 3D adaptive queries and 3D global queries are drawn in blue and green respectively. The GTs in orange are presented in BEV only.

of our approach, we conducted additional comparisons on nuScenes dataset, as shown in Tab. 2. Notably, our method outperforms previous SoTA methods with impressive results, achieving 51.0% mAP and 59.4% NDS on the `val` set and 63.5% mAP and 68.7% NDS on the `test` set. These superior metrics specifically highlight its effectiveness.

Ablation Study & Analysis

In this section, we present a comprehensive analysis of the essential components of our model. As shown in Tab. 3, we start from StreamPETR as the baseline in #1 and add each module to verify its effect.

Adaptive Query. Comparing #1 and #2 in Tab. 3, we can observe that adaptive query brings an improvement of 2.1% mAP and 1.5% CDS. Adaptive queries are insensitive to object range due to the robustness of 2D detectors in images, thus it is more suitable for general detection scenarios. To choose the optimal score threshold of 2D proposals, we conduct experiments shown in Tab. 4. Besides, we visualize the detection results in Fig. 5 and distinguish the boxes predicted from 3D adaptive queries and 3D global queries. The predictions from 3D adaptive queries cover a larger range, showing their indispensable significance.

Perspective-aware Aggregation. Adding the perspective-aware aggregation contributes a gain of 1.0% mAP and 1.2% CDS. Distant objects only occupy a few pixels on the image, therefore employing multi-level scales and views brings rich features according to different object locations.

Range-modulated 3D Denoising. 3D denoising brings an improvement of 1.0% mAP and 0.8% CDS. Penalizing negative samples flexibly alleviates the challenge of false proposals and helps localize 3D objects, by taking the object range into consideration. We present experiments on different noising designs and numbers of negative samples, shown in Tab. 5. The results imply that the logarithm function and two negative samples are optimal settings.

Effect of the Global Query. We also design the experiment to investigate the effect of global query in Tab. 6. 3D global queries and adaptive queries coexist in our framework and compensate for each other. As a baseline, StreamPETR suffers from the convergence problem when using a small number of global queries (e.g. 100), and only works for a sufficient amount. In contrast, our method showcases distinctive robustness. As the number of global queries decreases, our performance shows a slight decline.

# Negative sample	Method	mAP[%]↑	CDS[%]↑
0	–	23.4	17.3
1	$log(\cdot)$	24.0	17.7
2	$log(\cdot)$	**24.4**	**18.1**
3	$log(\cdot)$	24.3	18.0
2	$linear$	24.1	17.9
2	$sqrt$	24.0	17.7
2	$fixed$	23.7	17.6

Table 5: Performance Comparison of negative denoising samples with different designs and numbers.

# Global query	StreamPETR			Far3D (Ours)		
	100	300	644	100	300	644
mAP[%]↑	1.5	16.9	20.5	23.5	23.6	24.4
CDS[%]↑	0.9	11.8	14.8	17.4	17.5	18.1

Table 6: Impact of global query number. StreamPETR suffers from convergence problem. In contrast, our framework shows robust performance even with only adaptive queries.

Conclusion

In this paper, we present a sparse query-based method for 3D long-range detection. Our approach incorporates 3D adaptive queries derived from 2D object priors, yielding high-quality proposals for the decoder. To improve training efficacy, we introduce a perspective-aware aggregation and range-modulated 3D denoising technique. Experimental results demonstrate the promising performance of our method, indicating its great potential for practical applications.

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.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Chen, Y.; Liu, J.; Zhang, X.; Qi, X.; and Jia, J. 2023. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21674–21683.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fan, L.; Wang, F.; Wang, N.; and ZHANG, Z.-X. 2022. Fully sparse 3d object detection. *Advances in Neural Information Processing Systems*, 35: 351–363.

Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.

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.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Huang, J.; and Huang, G. 2022. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*.

Huang, J.; Huang, G.; Zhu, Z.; Ye, Y.; and Du, D. 2021. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*.

Jiang, Y.; Zhang, L.; Miao, Z.; Zhu, X.; Gao, J.; Hu, W.; and Jiang, Y.-G. 2023. Polarformer: Multi-camera 3d object detection with polar transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1042–1050.

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*, 0–0.

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

Li, Y.; Bao, H.; Ge, Z.; Yang, J.; Sun, J.; and Li, Z. 2022b. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with dynamic temporal stereo. *arXiv preprint arXiv:2209.10248*.

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

Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Qiao, Y.; and Dai, J. 2022c. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, 1–18. Springer.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Lin, X.; Lin, T.; Pei, Z.; Huang, L.; and Su, Z. 2022. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. *arXiv preprint arXiv:2211.10581*.

Lin, X.; Lin, T.; Pei, Z.; Huang, L.; and Su, Z. 2023. Sparse4D v2: Recurrent Temporal Fusion with Sparse Model. *arXiv preprint arXiv:2305.14018*.

Liu, J.; Wang, T.; Liu, B.; Zhang, Q.; Liu, Y.; and Li, H. 2023. Towards Better 3D Knowledge Transfer via Masked Image Modeling for Multi-view 3D Understanding. *arXiv preprint arXiv:2303.11325*.

Liu, Y.; Wang, T.; Zhang, X.; and Sun, J. 2022a. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, 531–548. Springer.

Liu, Y.; Yan, J.; Jia, F.; Li, S.; Gao, Q.; Wang, T.; Zhang, X.; and Sun, J. 2022b. Petrv2: A unified framework for 3d perception from multi-camera images. *arXiv preprint arXiv:2206.01256*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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 preprint arXiv:2210.02443*.

Philion, J.; and Fidler, S. 2020. 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*, 194–210. Springer.

Reading, C.; Harakeh, A.; Chae, J.; and Waslander, S. L. 2021. Categorical depth distribution network for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8555–8564.

Shao, S.; Li, Z.; Zhang, T.; Peng, C.; Yu, G.; Zhang, X.; Li, J.; and Sun, J. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of* *the IEEE/CVF international conference on computer vision*, 8430–8439.

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

Wang, S.; Jiang, X.; and Li, Y. 2022. Focal-PETR: Embracing Foreground for Efficient Multi-Camera 3D Object Detection. *arXiv preprint arXiv:2212.05505*.

Wang, S.; Liu, Y.; Wang, T.; Li, Y.; and Zhang, X. 2023a. Exploring Object-Centric Temporal Modeling for Efficient Multi-View 3D Object Detection. *arXiv preprint arXiv:2303.11926*.

Wang, T.; Zhu, X.; Pang, J.; and Lin, D. 2021. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 913–922.

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

Wang, Z.; Huang, Z.; Fu, J.; Wang, N.; and Liu, S. 2023b. Object as query: Equipping any 2d object detector with 3d detection ability. *arXiv preprint arXiv:2301.02364*.

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

Xie, E.; Yu, Z.; Zhou, D.; Philion, J.; Anandkumar, A.; Fidler, S.; Luo, P.; and Alvarez, J. M. 2022. M2BEV: Multi-Camera Joint 3D Detection and Segmentation with Unified Birds-Eye View Representation. *arXiv preprint arXiv:2204.05088*.

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.

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

Zhang, R.; Qiu, H.; Wang, T.; Guo, Z.; Xu, X.; Qiao, Y.; Gao, P.; and Li, H. 2022. MonoDETR: depth-guided transformer for monocular 3D object detection. *arXiv preprint arXiv:2203.13310*.

Zhang, Y.; Zheng, W.; Zhu, Z.; Huang, G.; Lu, J.; and Zhou, J. 2023. A Simple Baseline for Multi-Camera 3D Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3507–3515.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.

Zong, Z.; Jiang, D.; Song, G.; Xue, Z.; Su, J.; Li, H.; and Liu, Y. 2023. Temporal Enhanced Training of Multi-view 3D Object Detector via Historical Object Prediction. *arXiv preprint arXiv:2304.00967*.