2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

FarDetect3D: Enhancing Long-Range Object Detection in Multi-View 3D Systems

Songyan Liu✉, Chaoyi Luo, Tong Xiao, Xiaofei Liu✉, and Jing Chai

School of Information Science and Engineering, Yunnan University,
Kunming 650091, China.
liusongyan@ynu.edu.cn

Abstract. In this paper, we propose a new multi-view 3D detection paradigm, named FarDetect3D, to enhance the detection of long-range objects.FarDetect3D improves the existing sparse query-based multi-view 3D object detection framework by introducing two modules: Remote Detection Denoising (ReDN) and Long-range Feature Attention (LrFA). In ReDN, we utilize the fake long-range depth information to generate sparse 3D queries, which improves the performance of the long-range detection. In LrFA, we enhance the central features and capture the contextual relationships between the distant pixels, further boosting the detection accuracy. Experimental results show that our approach outperforms the state-of-the-art camera-based multiview 3D detection methods, which can provide a robust solution for safe autonomous driving in complex environments.

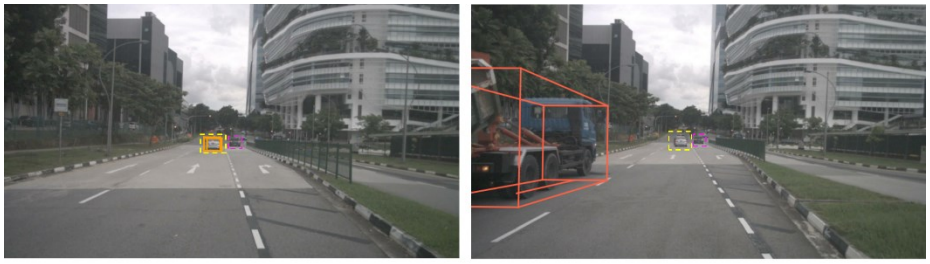
Keywords: Multi-view 3D Object Detection, Denoising, Long-range Feature Attention.

1 Introduction

Multi-view 3D object detection[1–5] is an essential step for comprehending the 3D environments in autonomous driving systems, which aims to identify and classify the objects around the vehicle by surround-view cameras. Compared with LIDAR-based methods, camera-based methods are more economical and can provide valuable visual cues for long-range distance detection and vision-only element identification.

Current multi-view 3D object detection methods can generally be divided into two main categories: dense Bird's Eye View (BEV) and sparse query-based approaches. BEV-based techniques [1–3] often use a view transformer [6] to transform perspective features into BEV features, followed by the generation of 3D bounding boxes using a 3D detector head. On the other hand, sparse query-based methods [4, 5] focus on learning 3D global object queries from representative training data and produce detection results by integrating DETR [7] style image features.

✉ Corresponding Author



(a) (b)

Fig. 1. (a) and (b) show the prediction results of StreamPETR[5] for the same scene at different moments. Some vehicles are successfully detected in (a) but missed in (b), as highlighted in the yellow dashed boxes. In addition, as shown in the pink dashed box, some vehicles are missed in both frames. This highlights the challenge of maintaining detection consistency across time, with some long-range objects being missed.

However, the BEV-based and sparse query-based methods are skilled at short-range detection but struggle with long-range detection. The detection results for long-range objects are usually inaccurate and inconsistent. As shown in Fig. 1. (a) and (b) show the prediction results of StreamPETR[5] for the same scene at different moments. Some vehicles are successfully detected in (a) but missed in (b), as highlighted in the yellow dashed boxes. In addition, as shown in the pink dashed box, some vehicles are missed in both frames. This highlights the challenge of maintaining detection consistency across time, with some long-range objects being missed., some long-range objects cannot be detected in both the two frames, while other objects may be detected in one frame but missed in another. These unfavorable results have different causes for different categories of detection methods. In the BEV-based methods, the computational complexity is too high to obtain a large enough BEV feature map for long-range detection, especially in real-time applications. Besides, the fixed number of queries in sparse query-based methods limits their ability to effectively capture the dynamic environment changes, resulting in missed detections of long-range objects.

In this paper, we propose a new multi-view 3D detection framework, FarDetect3D, based on traditional sparse query-based methods to overcome these challenges. Compared with the existing methods, our proposed FarDetect3D can boost the detection of long-range objects in the following two aspects: First, we introduce the Remote Detection Denoising (ReDN) module, which presents a distance-based denoising mechanism. Given the real depth information of a long-range object, ReDN generates different fake depths and then obtains the corresponding 3D queries. In this way, ReDN improves the number of long-range 3D queries, then enables the model to learn the long-range depth information and reduces the noise from depth errors and object scale variations, which can help to improve the detection accuracy of long-range objects. Second, we introduce the Long-range Feature Attention (LrFA) module, which refines and enhances the features relevant to long-range objects. LrFA works by focusing on the central features of the 3D objects, capturing contextual relationships between distant pixels, and aggregating information over long distances. This process improves the system's ability to detect and classify objects more holistically, taking into account not only the immediate surroundings but also the broader environment. By incorporating long-range contextual information, LrFA effectively mitigates the challenges posed by dynamic environments, where objects may appear and disappear rapidly, ensuring better accuracy and fewer missed detections.

In summary, the contributions of this paper are as follows:

- We propose the ReDN module to generate the long-range fake queries and the LrFA module to enhance the features relevant to long-range objects; both of them are helpful in detecting long-range objects.
- By introducing ReDN and LrFA into the existing sparse query-based multi-view 3D detection model, we propose a novel detection framework, FarDetect3D.
- Experimental results show that FarDetect3D achieves state-of-the-art in camera-based multi-view 3D object detection methods. To be specific, we obtain 56.7% mAP and 64.7% NDS on the nuScenes dataset with V2-99 as the backbone.

2 Related Work**Multi-View 3D Detection**

Multi-view 3D detection involves predicting 3D bounding boxes in a global framework using multiple cameras. Most prior methods [8–10] are extensions of monocular 3D object detection and fail to fully leverage the geometric details present in multi-view images. Recently, some studies [2, 6] have explored the use of explicit bird's eye view (BEV) mapping for object perception within a global system. LSS [6] performs view transformation by estimating depth distributions and converting images into BEV. BEVFormer [2] combines spatial and temporal information by employing BEV queries arranged in a predefined grid pattern. BEVDepth incorporates a point cloud as a depth supervisor to encode camera parameters into a depth subnetwork.

Following the DETR [7] framework, several approaches developed implicit BEV features. These approaches generally start with 3D sparse object queries and interact with 2D features using an attention mechanism to conduct 3D object detection directly. For instance, DETR3D [11] extracts 2D features from projected 3D reference points and updates the query using local cross-attention. PETR [12] introduces 3D positional

embedding in a global context and updates the query through global cross-attention. PETRv2 [4] extends PETR with temporal modeling and integrates self-motion into the positional embedding concept.

2.2 Denoising in Object Detection

In the realm of 2D object detection, models such as DETR [7] and its variations [13–15] frequently struggle with convergence, often due to the instability of bipartite graph matching and inconsistent optimization objectives during the early stages of training. DN-DETR [16] tackles this issue by employing the concept of input noise, incorporating ground truth bounding boxes into the Transformer decoder. This trains the model to reconstruct the original, clean boxes through a process known as denoising. Here, noisy ground truth bounding boxes are used to initialize queries, termed denoising queries. DINO [17] builds upon this method by creating both positive and negative denoising queries for each ground truth bounding box, with the negative queries containing additional noise to improve the model's ability to reject incorrect predictions.

In 3D object detection, DETR-style methods [4, 5, 18] use denoising techniques by generating denoising queries for each ground truth bounding box, distinguishing between objects and "no object" based on noise levels. However, these methods primarily focus on short-range sensing and struggle with detecting long-range objects. This limitation is particularly evident in multi-view 3D object detection, where the model performs poorly on long-range objects due to insufficient training samples and challenges like depth blurring and feature degradation.

To overcome these challenges, we propose Remote Detection Denoising (ReDN), to incorporate the knowledge of the 3D structure of the scene. Different from the traditional denoising queries, which focus on nearby objects, ReDN generates separate negative denoising queries for long-range objects, which are spatially distinct from those used for nearby objects. This enables the model to better handle the detection of long-range objects by accounting for depth inaccuracies and scale variations.

3 Methodology

The overall architecture of FarDetect3D is shown in Fig. 2, multi-view images are input into the backbone network to obtain the feature maps, and then enhance the feature of long-range objects by the LrFA module. Meanwhile, ReDN generates the fake depths and corresponding 3D queries of the objects. Finally, the generated fake queries and the original real queries are mixed and input into the Transformer Decoder to predict the detection results. Next, we introduce ReDN and LrFA in detail.

3.1 Implementation Details

We denote the center of the ground truth bounding box $B_{GT} = (x, y, z, w, h, l)$ as $C_{GT} = (x, y, z, 1)$ and its size as $S_{GT} = (w, h, l)$, with all coordinates expressed in the camera's coordinate system. Using C_{GT}, we generate reference points for remote detection

denoising queries. The center of projection in the camera's truncated cone is represented as$C_{GT} = (u \times d, v \times d, d, 1)$, where (u, v) are the pixel coordinates and d denotes the depth value.

With the known depth of the center of the ground truth object, we can generate fake depth noise information based on the true depth:

$$d_{fake} = d + \beta_i \cdot g(S_{GT}) \tag{1}$$

In this context, g represents the function to calculate the average scale of the actual object, given by $g(S_{GT}) = k \cdot \frac{(w+h+l)}{6}$. Here, the radius k determines the effective distribution range of the reference points, and β_i denotes the offset for the i -th reference point.

By the above method, we obtain the fake depth information d of the object, while the real coordinates of the object is (X, Y), where $X = u \times$ d and $Y = v \times d$. Here, u and v are the pixel coordinates, and d represents the depth value. According to the fake

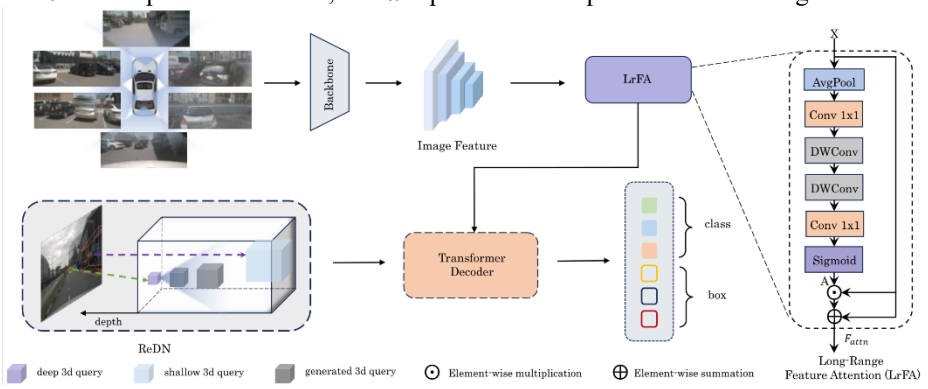


Fig. 2. Overview of FarDetect3D. Surround-view images are fed into the backbone to obtain image features, which are augmented with the LrFA module to refine and enhance the features relevant to long-range objects. Using the ReDN method, we generate the fake depths of the objects and the corresponding 3D queries. The generated 3d queries are then mixed with the original 3d queries and iteratively improved by the decoder layer to predict the 3D bounding box.

depth information, the fake coordinates X_{fake} and Y_{fake} of the object can be calculated by the following formula:

$$X_{fake} = \frac{X \times d_{fake}}{d} \tag{2}$$

$$Y_{fake} = \frac{Y \times d_{fake}}{d} \tag{3}$$

Then we use X_{fake}, Y_{fake} and d_{fake} to compute the 3D proposal center c_{3d}.

$$c_{3d} = (X_{fake}, Y_{fake}, d_{fake}, 1) \tag{4}$$

Finally, we encode c_{3d} as a generated 3D query as follows:

$$Q = Embed(c_{3d}) \tag{5}$$

Where $Embed$ denotes an embedding layer consisting of a sinusoidal transformation and an MLP.

3.2 Implementation Details

Directly scaling an existing 3D detector from short-range to long-range presents several challenges, including high computational demands, inefficient convergence, and declining localization accuracy. For instance, to adequately cover a broader range of potential objects, the number of queries must increase at least quadratically, which is impractical in real-world applications. Furthermore, the presence of small and sparse long-range objects can obstruct convergence and negatively affect the localization of nearby objects.

To address these issues, we introduce the Long-range Feature Attention (LrFA) module designed to process image features derived from the backbone. LrFA module focuses on enhancing central features while also capturing the contextual relationships among distant pixels, as depicted in Fig. 2. It employs average pooling and 1×1 convolution to extract features from local regions:

$$x_{pool} = Conv_{1\times1}\left(P_{avg}(x)\right) \tag{6}$$

Where P_{avg} denotes the average pooling operation,x denotes the image features extracted by the backbone network.

We utilize two depth-wise strip convolutions as an approximation of a standard large-kernel depth-wise convolution:

$$x_w = DWConv_{1\times k_b}(x_{pool}) \tag{7}$$

$$x_h = DWConv_{k_b\times1}(x_w) \tag{8}$$

We chose the deep strip convolution based on one main consideration: its lightweight nature. Compared to the traditional $k_b \times k_b$ 2D depth convolution, the deep strip convolution achieves a similar effect with $\frac{k_b}{2}$ fewer parameters, making it more computationally efficient.

Finally, our LrFA module produces an attention weight A, which is further used to enhance the output of the image features:

$$A = Sigmoid\big(Conv_{1\times1}(x_h)\big) \tag{9}$$

$$x_{attn} = \lambda_1(A \odot x) \oplus \lambda_2 x \tag{10}$$

Here, the Sigmoid function ensures that the attention graph A is in the range $(0,1)$, $\odot$ denotes element-by-element multiplication, $\otimes$ denotes element-by-element summation, and x_{attn} is an augmented feature.

4 Experiment

4.1 Dataset and Metrics

Our model's effectiveness is assessed using the nuScenes dataset [19] and Argoverse 2 dataset[20]. The nuScenes dataset contains 1000 video sequences, categorized into a

training set (700 videos), a validation set (150 videos), and a test set (150 videos). Each video lasts around 20 seconds and is annotated every 0.5 seconds. This collection includes 1.4 million annotated 3D bounding boxes spanning ten object categories. The evaluation metrics encompass mean Average Precision (mAP) along with five accurate positive metrics: mean Average Translation Error (mATE), mean Average Scale Error (mASE), mean Average Orientation Error (mAOE), mean Average Velocity Error (mAVE), and mean Average Attribute Error (mAAE), which evaluate errors in translation, scale, orientation, velocity, and attributes, respectively. The nuScenes Detection Score (NDS) combines these metrics to evaluate overall performance.

The Argoverse 2 dataset contains 1000 unique scenes, each 15 seconds long, annotated at 10 Hz. The scenes are divided into 700 for training, 150 for validation, and 150 for testing. The evaluation covers 26 categories within a 150-meter range, addressing long-range perception tasks. Metrics include the mAP and the Composite Detection Score (CDS), which integrates three key true positive metrics: mean Average Translation Error (mATE), mean Average Scale Error (mASE), and mean Average Orientation Error (mAOE).

4.2 Implementation Details

Our experimental setup utilizes ResNet50, ResNet101[21], and V2-99[22] as the backbone, each initialized with different pre-training strategies. We thoroughly assess the performances of ResNet50 and ResNet101, both pre-trained on nuImages[19], using the nuScenes validation dataset. To illustrate the extensibility of our approach, we also include results on the nuScenes test set employing V2-99, which is initialized with DD3D[23] checkpoints. The optimization of the model is performed using the AdamW optimizer with a batch size of 16. For models trained solely on the training set, the learning rate is set at $4 \times 10{-}4$,whereas for those trained on both the training and validation sets, it is adjusted to $3 \times 10{-}4$. A cosine annealing schedule is applied to adjust the learning rate. To benchmark against top-performing methods, the model is trained for 60 epochs without a Class-Balanced Group. Sampling (CBGS)[24] and for 24 epochs during the ablation study on the nuScenes. Our implementation is predominantly based on the StreamPETR[5] framework.

4.3 Comparison with State-of-the-Art Methods

We compare the FarDetect3D model with other state-of-the-art multi-view 3D object detectors on the validation and test sets of the nuScenes dataset, as well as the validation set of the Argoverse 2 dataset.

nuScenes Validation Set. Table 1 presents a comparison with state-of-the-art methods on the nuScenes validation set. We evaluated both ResNet-50 and ResNet-101 backbones. Using ResNet-50 as the backbone with an image size of 256×704,we achieved 46.8% mAP and 56.6% NDS, improving upon the previous state-of-the-art method StreamPETR by 1.8% mAP and 1.6% NDS. With the more powerful ResNet-101 backbone and an increased image size of 512×1408,our performance reached

51.6% mAP and 60.3% NDS, surpassing StreamPETR by 1.2% mAP and 1.1% NDS with the similar speed (FPS 6.3 v.s. 6.4).

Long-range Results. We introduce a new test set specifically designed for 3D long-range car detection on the nuScenes validation dataset. In our experiments, we define the objects with a depth farther than 50 meters from the camera as the long-range. The 3D long-range car detection evaluation results are presented in Table 2. Our method outperforms StreamPETR [5] with a notable improvement in both mAP and NDS. Specifically, we achieve an mAP of 34.3% and an NDS of 48.5%, compared to StreamPETR's mAP of 31.7% and NDS of 45.6%. These results for StreamPETR were obtained by running the original author's code.

Argoverse 2 Validation Set. To evaluate the generalization ability of the model, we performed additional experiments on the Argoverse 2 dataset, as shown in

Table **4**. Our method clearly outperforms previous StreamPETR, achieving a 2.3% mAP and 1.9% CDS improvement on the validation set. These metrics emphasize the generalization ability of our method.

Table 1. Evaluation on the nuScenes validation set. † Marks techniques that gain advantages from pre-training in a perspective view. FPS is measured on RTX3090 with fp32.

Methods	Backbone	Image Size	mAP	NDS	mATE	mASE	mAOE	mAVE	mAAE	FPS
BevDet4D [1]	ResNet50	256x704	32.2	45.7	0.703	0.278	0.495	0.354	0.206	16.7
PETRv2 [4]	ResNet50	256x704	34.9	45.6	0.700	0.275	0.58	0.437	0.187	16.7
BEVDepth [25]	ResNet50	256x704	35.1	47.5	0.629	0.267	0.479	0.428	0.198	15.7
BEVStereo [26]	ResNet50	256x704	37.2	50	0.598	0.270	0.438	0.367	0.190	12.2
BEVFormerv2 [3] †	ResNet50	-	42.3	52.9	0.618	0.273	0.413	0.333	0.188	-
SOLOFusion[27]	ResNet50	256x704	42.7	53.4	0.567	0.274	0.511	0.252	0.181	11.4
SparseBEV[18]†	ResNet50	256x704	44.8	55.8	0.581	0.271	0.373	0.247	0.190	-
StreamPETR[5]†	ResNet50	256x704	45	55	0.613	0.267	0.413	0.265	0.198	31.7
DVPE [28]	ResNet50	256x704	46.6	55.9	0.608	0.271	0.386	0.274	0.202	-
FarDetect3D (Ours)	ResNet50	256x704	**46.8**	**56.6**	0.556	0.261	0.412	0.252	0.196	30.6
BEVDepth [25]	ResNet101	512x1408	41.2	53.5	0.565	0.266	0.358	0.331	0.190	-
PETRv2 [4]	ResNet101	640x1600	42.1	52.4	0.681	0.267	0.357	0.377	0.186	-
Sparse4D [29]	ResNet101	900x1600	43.6	54.1	0.633	0.279	0.363	0.317	0.177	4.3
SOLOFusion [27]	ResNet101	512x1408	48.3	58.2	0.503	0.264	0.381	0.246	0.207	-
SparseBEV [18] †	ResNet101	512x1408	50.1	59.2	0.562	0.265	0.321	0.243	0.195	-
StreamPETR [5] †	ResNet101	512x1408	50.4	59.2	0.569	0.262	0.315	0.257	0.199	6.4
FarDetect3D (Ours)	ResNet101	512x1408	**51.6**	**60.3**	0.541	0.260	0.314	0.236	0.198	6.3

Table 2. Comparing the performance of StreamPETR and ReDN on 3D long-range car detection.

Methods	Backbone	mAP↓	NDS↓	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
StreamPETR [5]	ResNet50	31.7	45.6	0.713	0.275	0.493	0.332	0.203
FarDetect3D (Ours)	ResNet50	**34.3**	**48.5**	0.631	0.257	0.489	0.311	0.196

Table 3. Evaluation on the nuScenes test dataset.

Methods	Backbone	Image Size	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
DETR3D [11]	V2-99	900×1600	41.2	47.9	0.641	0.255	0.394	0.845	0.133
MV2D [30]	V2-99	640×1600	46.3	51.4	0.542	0.247	0.403	0.857	0.127
BEVFormer [2]	V2-99	900×1600	48.1	56.9	0.582	0.256	0.375	0.378	0.126
PETRv2 [4]	V2-99	640×1600	49.0	58.2	0.561	0.243	0.361	0.343	0.120
PolarFormer [31]	V2-99	900×1600	49.3	57.2	0.556	0.256	0.364	0.439	0.127
BEVStereo [26]	V2-99	900×1600	52.5	61.0	0.431	0.246	0.358	0.357	0.138
HoP [32]	V2-99	640×1600	52.8	61.2	0.491	0.242	0.332	0.343	0.109
SparseBEV [18]	V2-99	640×1600	54.3	62.7	0.502	0.244	0.324	0.251	0.126
StreamPETR [5]	V2-99	640×1600	55.0	63.6	0.479	0.239	0.317	0.241	0.119
DualBEV [33]	V2-99	640×1600	55.2	63.4	0.414	0.245	0.377	0.252	0.129
FarDetect3D (Ours)	V2-99	640×1600	**56.7**	**64.7**	0.461	0.240	0.310	0.239	0.114

Table 4. Comparisons on the Argoverse 2 validation set.

Methods	**Backbone**	**Resolution**	**mAP↑(%)**	**CDS↑(%)**	**mATE↓**	**mASE↓**	**mAOE↓**
PETR [12]	V2-99	900×640	17.6	12.2	0.911	0.339	0.819
Sparse4Dv2 [34]	V2-99	900×640	18.6	13.4	0.832	0.343	0.723
StreamPETR [5]	V2-99	900×640	20.3	14.6	0.843	0.321	0.650
FarDetect3D (Ours)	V2-99	900×640	**22.6**	**16.5**	0.825	0.319	0.624

Table 5. Ablation studies of LrFA and ReDN in FarDetect3D.

LrFA	ReDN	mAP↑(%)	NDS↑(%)	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
		42.1	51.0	0.703	0.278	0.455	0.354	0.206
✓		42.7	51.9	0.700	0.275	0.431	0.351	0.185
	✓	43.0	52.1	0.630	0.267	0.430	0.420	0.190
✓	✓	**43.9**	**52.5**	0.670	0.290	0.452	0.345	0.184

4.4 Ablation Studies

This part delves into the ablation studies carried out using validation sets from the nuScenes. Unless specified otherwise, our experiments on nuScenes use a ResNet50 backbone that has been pre-trained on the nuImages dataset. We take 8frames as input, each with a resolution of 256 × 704 pixels. The decoder uses 428 queries, and the model undergoes training for 24 epochs without utilizing CBGS.

We conduct an ablation study to evaluate the contributions of Long-range Feature Attention (LrFA) and Remote Detection Denoising (ReDN) in improving long-range 3D object detection. The results are summarized in Table 5.

LrFA Ablation. We first evaluate the impact of LrFA by comparing the performance of our model with and without the LrFA module. As shown in Row #1 and #2 of Table 5, the inclusion of LrFA results in an improvement of 0.6% mAP and 0.9% NDS. The LrFA module enhances the ability of the model to capture contextual relationships between distant pixels, which significantly boosts the accuracy of detecting objects at long distances. This demonstrates the importance of context awareness in handling long-range objects and improving feature representation.

ReDN Ablation. Next, we assess the effectiveness of ReDN by comparing the baseline method (without ReDN) to our full model that incorporates ReDN. As shown in Row #1 and #3 of Table 5, ReDN brings a substantial performance boost of 0.9% mAP and 1.1% NDS. The ReDN module addresses the challenge of depth ambiguity and resolution degradation in long-range object detection by introducing a remote detection denoising mechanism. This mechanism enhances depth estimation and reduces fake positives, particularly in long-range scenarios. The results clearly highlight the crucial role of ReDN in improving detection accuracy and robustness in complex environments.

Combined Effect of LrFA and ReDN. Finally, we analyze the combined effect of both LrFA and ReDN in Table 5. The joint use of these two modules results in a significant overall improvement of 1.8% mAP and 1.5% NDS, demonstrating that LrFA and ReDN complement each other effectively. LrFA boosts feature attention for long-range objects, while ReDN improves the precision of depth estimation and reduces noise in long-range detections. Together, they provide a robust solution for enhancing detection accuracy in autonomous driving, particularly in complex, large-scale environments.

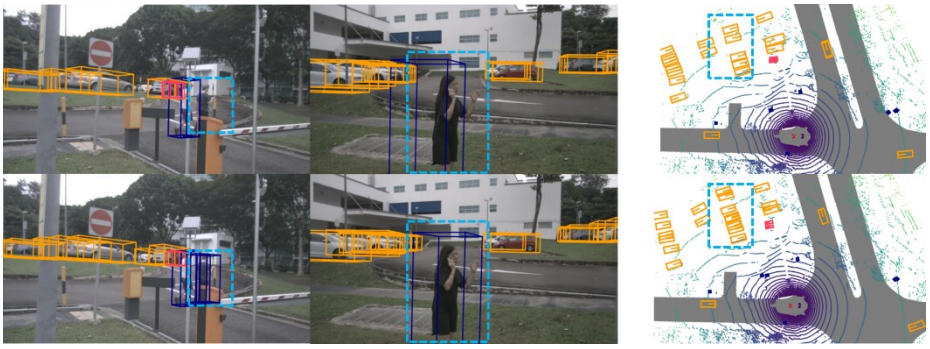


Fig. 3. Visualization of the images and BEV spatial detection results for comparison with StreamPETR, each set of images shows the detection results of StreamPETR on the top and those of FarDetect3D on the bottom.

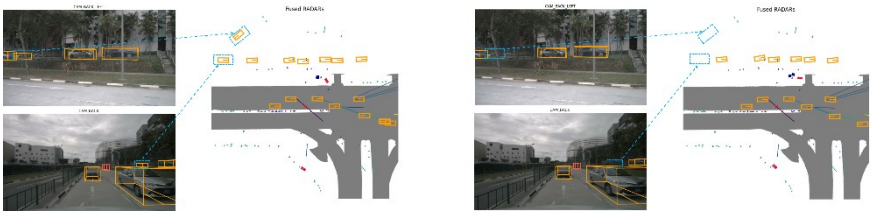


Fig. 4. Visualization of the images and Fused RADARs spatial detection results for comparison with StreamPETR, each set of images shows the detection results of StreamPETR on the top and those of our proposed model on the bottom.

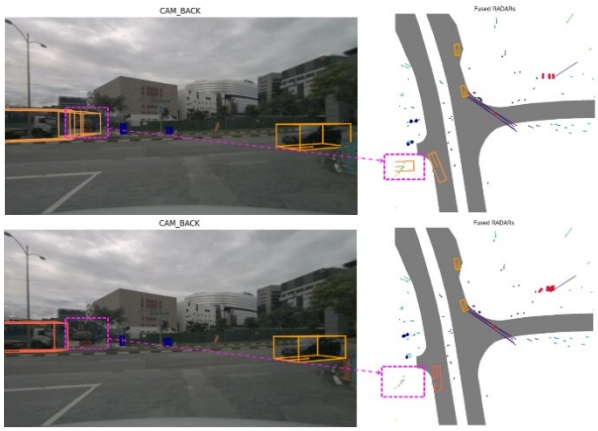


Fig. 5. Visualization of images and Fused RADARs spatial detections for comparison with StreamPETR, two sets of images showing the detections of StreamPETR on the left side and the detections of our proposed model on the right side.

4.5 Visualization

For better showing the result comparison between FarDetect3D and the existing methods, we visualize the detection results in different forms. Fig. 3 visualizes the detection results of StreamPETR and FarDetect3D. The prediction results of our model show better localization and higher-quality bounding boxes, especially for long-range objects. Fig. 4 illustrates the detection results from different viewpoints within the same scene. While StreamPETR fails to detect some long-distance targets, our proposed model demonstrates improved detection performance for these targets. Fig. 5 illustrates how StreamPETR incorrectly detects a target while our proposed model correctly identifies the target.

5 Conclusion

In this paper, we present a novel multi-view 3D object detection approach named FarDetect3D to boost the performance of long-range object detection. In FarDetect3D, we introduce two modules: Remote Detection Denoising (ReDN) enhances the accuracy of long-range object detection by utilizing fake depth information to generate 3D queries; Long-range Feature Attention (LrFA), which focuses on strengthening central features and capturing contextual relationships between distant pixels. Our experimental results demonstrate that FarDetect3D outperforms existing methods, particularly in camera-based systems. This approach offers a robust solution for detecting long-range objects, ensuring safer autonomous driving in complex real-world environments. The promising performance of our method underscores its potential for practical deployment in autonomous vehicles.

Acknowledgements

This work was supported by the Open Project of YNU Resilience and Excellence Platform for Developing Children's Character Traits [Grant No. K207003250005], and Yunnan Provincial Research Foundation for Basic Research, China [Grant No. 202401AT070427 and 202401AT070442].

References

1. Huang, J., Huang, G.: BEVDet4D: Exploit Temporal Cues in Multi-camera 3D Object Detection. arXiv preprint arXiv:2203.17054 (2022).
2. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: BEV-Former: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. In: ECCV (2022).
3. Yang, C., Chen, Y., Tian, H., Tao, C., Zhu, X., Zhang, Z., Huang, G., Li, H., Qiao,Y., Lu, L.: BEVFormer v2: Adapting Modern Image Backbones to Bird's-Eye-View Recognition via Perspective Supervision. In: CVPR (2023).
4. Liu, Y., Yan, J., Jia, F., Li, S., Gao, A., Wang, T., Zhang, X.: PETRv2: A Unified Framework for 3D Perception from Multi-Camera Images. In: ICCV (2023).
5. Wang, S., Liu, Y., Wang, T., Li, Y., Zhang, X.: Exploring Object-Centric Temporal Modeling for Efficient Multi-View 3D Object Detection. In: ICCV (2023).
6. Philion, J., Fidler, S.: Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D. In: ECCV (2020).
7. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-End Object Detection with Transformers. In: ECCV (2020).
8. Chen, H., Wang, P., Wang, F., Tian, W., Xiong, L., Li, H.: EPro-PnP: Gener-alized End-to-End Probabilistic Perspective-n-Points for Monocular Object PoseEstimation. arXiv preprint arXiv:2103.00390 (2021).
9. Wang, T., Zhu, X., Pang, J., Lin, D.: Probabilistic and Geometric Depth: Detecting Objects in Perspective. arXiv preprint arXiv:2107.04957 (2021).

10. Wang, T., Zhu, X., Pang, J., Lin, D.: FCOS3D: Fully Convolutional One-Stage Monocular 3D Object Detection. In: ICCV (2021).
11. Wang, Y., Guizilini, V., Zhang, T., Wang, Y., Zhao, H., Solomon, J.: DETR3D:3D Object Detection from Multi-view Images via 3D-to-2D Queries. arXiv preprint arXiv:2110.03946 (2021).
12. Liu, Y., Wang, T., Zhang, X., Sun, J.: PETR: Position Embedding Transformation for Multi-View 3D Object Detection. arXiv preprint arXiv:2104.08063 (2021).
13. 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).
14. Wang, Y., Zhang, X., Yang, T., Sun, J.: Anchor DETR: Query Design for Transformer-Based Detector. In: AAAI (2022).
15. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable Transformers for End-to-End Object Detection. arXiv preprint arXiv:2010.04159 (2020).
16. Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L.: DN-DETR: Accelerate DETR Training by Introducing Query DeNoising. In: CVPR (2022).
17. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L., Shum, H.Y.: DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. In: ICLR (2022).
18. Liu, H., Teng, Y., Lu, T., Wang, H., Wang, L.: SparseBEV: High-Performance Sparse 3D Object Detection from Multi-Camera Videos. In: ICCV (2023).
19. 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: CVPR (2020).
20. Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B.,Kumar, R., Hartnett, A., Pontes, J.K., et al.: Argoverse 2: Next generation datasets for self-driving perception and forecasting. In: NeurIPS (2021).
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: CVPR (2016).
22. Lee, Y., Hwang, J.W., Lee, S., Bae, Y., Park, J.: An Energy and GPU-Computationally Efficient Backbone Network for Real-Time Object Detection. In: CVPR (2019).
23. Park, D., Ambrus, R., Guizilini, V., Li, J., Gaidon, A.: Is Pseudo-Lidar Needed for Monocular 3D Object Detection? In: ICCV (2021).
24. Zhu, B., Jiang, Z., Zhou, X., Li, Z., Yu, G.: Class-balanced Grouping and Sampling for Point Cloud 3D Object Detection. arXiv preprint arXiv:1908.09492 (2019).
25. Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., Li, Z.: BEVDepth: Acquisition of Reliable Depth for Multi-view 3D Object Detection. In: AAAI (2023).
26. Li, Y., Bao, H., Ge, Z., Yang, J., Sun, J., Li, Z.: BEVStereo: Enhancing Depth Estimation in Multi-view 3D Object Detection with Dynamic Temporal Stereo. In: AAAI (2023).
27. Park, J., Xu, C., Yang, S., Keutzer, K., Kitani, K.M., Tomizuka, M., Zhan, W.: Time Will Tell: New Outlooks and a Baseline for Temporal Multi-View 3D Object Detection. In: ICLR (2022).
28. Wang, J., Li, Z., Sun, K., et al.: DVPE: Divided View Position Embedding for Multi-view 3D Object Detection. arXiv preprint arXiv:2407.16955 (2024).
29. Lin, X., Lin, T., Pei, Z., Huang, L., Su, Z.: Sparse4D: Multi-view 3D Object Detection with Sparse Spatial-Temporal Fusion. arXiv preprint arXiv:2211.10581 (2022).
30. Wang, Z., Huang, Z., Fu, J., Wang, N., Liu, S.: Object as Query: Lifting any 2D Object Detector to 3D Detection. In: ICCV (2023).
31. Jiang, Y., Zhang, L., Miao, Z., Zhu, X., Gao, W., Hu, W., Jiang, Y.G.: Polarformer:Multi-camera 3D Object Detection with Polar Transformer. In: AAAI (2023).

32. Zong, Z., Jiang, D., Song, G., Xue, Z., Su, J., Li, H., Liu, Y.: Temporal Enhanced Training of Multi-view 3D Object Detector via Historical Object Prediction. In: ICCV (2023).
33. Li, P., Shen, W., Huang, Q., et al.: DualBEV: A multi-camera 3D object detection approach. arXiv preprint arXiv:2305.14018 (2023).
34. Lin, X., Lin, T., Pei, Z., Huang, L., Su, Z.: Sparse4D v2: Recurrent temporal fusion with sparse model. arXiv preprint arXiv:2305.14018 (2023).