Enhancing Multi-View 3D Object Detection with 2D Auxiliary Information and Feature Fusion

Mu Yu[1,†], Chaojie Zhang[1]
[1]*Department of Control Science and Engineering, Tongji University*, Shanghai, China

***Abstract*—Multi-view 3D object detection has gained increasing attention for its ability to construct a comprehensive 3D scene representation in complex scenarios. However, existing sparse query-based detectors often exhibit performance degradation in crowded scenes or long-range scenarios, primarily due to the reliance on static queries. In this paper, we propose EAFF, an enhanced multi-view 3D object detection framework that leverages 2D auxiliary supervision and adaptive feature fusion to address these limitations. Specifically, EAFF employs a multi-task 2D neural network to extract object-aware semantic representations. These representations are further leveraged to construct adaptive 3D queries, providing more informative and flexible queries than fixed sparse ones. Moreover, we design an attribute-aware feature fusion module that explicitly aligns and integrates attribute-specific features within a unified embedding space, enabling more consistent feature alignment and improved discrimination. Extensive experiments on the nuScenes benchmark demonstrate that EAFF consistently improves detection performance and achieves a 1.3% improvement in mAP compared to the state-of-the-art RayDN method, while maintaining robust detection performance in complex environments.**

***Index Terms*—Multi-view 3D Object Detection, Sparse Paradigm, Auxiliary Guidance**

I. Introduction

Multi-view image-based 3D object detection has attracted increasing interest in autonomous driving and robotic perception. Compared to LiDAR-based methods, image-based approaches offer notable advantages in terms of hardware deployment and economic efficiency.

Current approaches for multi-view 3D object detection can be broadly categorized into two paradigms: dense Bird's-Eye-View (BEV) based methods [4], [5] and sparse query-based methods [6]–[10]. Dense BEV-based approaches construct a unified representation via view transformation. However, they suffer from quadratic computational overhead as the perception range expands, making it challenging to strike a balance between computational efficiency and detection accuracy. Conversely, sparse query-based methods mitigate this complexity by employing a set of learnable queries in 3D and iteratively interacting with multi-view image features to generate corresponding results. Despite their efficiency, existing sparse methods often struggle in crowded scenes or with long-range targets. This limitation primarily stems from the sparse query initialization at predefined spatial locations, which restricts model flexibility in complex environments.

†Corresponding author.

(a) 3D Prediction BBox

(b) 2D Prediction BBox

Fig. 1. Performance comparisons on nuScenes dataset [1] between 3D detection and 2D detection. (a) and (b) demonstrate predicted boxes of RayDN [2] and YOLOX [3], respectively.

As illustrated in Fig. 1, the prediction accuracy of dense and distant objects is often lower in 3D object detection than in 2D detection. This is primarily because transforming features from the 2D perspective to the BEV representation inevitably leads to the loss of fine-grained visual details. This issue is especially pronounced for small or distant objects, resulting in a degraded detection performance. Although previous studies [8], [9], [11], [12] have leveraged 2D detection as auxiliary information, they mainly focus on geometric cues, such as spatial locations and bounding box dimensions, while the orientation information of objects is either insufficiently modeled or deferred to the later regression stage. This often leads to ambiguous heading estimation and optimization instability.

To overcome these challenges, we propose EAFF, an enhanced multi-view 3D detection framework that integrates a 2D auxiliary network with an adaptive feature fusion mechanism. Our key intuition is that 2D images inherently contain rich visual cues for orientation. By explicitly predicting these attributes in the 2D domain and incorporating them into adaptive 3D query generation, the proposed framework can recover potential targets missed by sparse global queries and achieve more accurate spatiotemporal estimation. Furthermore, we introduce a lightweight adaptive feature fusion module to

effectively integrate features across the 2D and 3D domains, yielding significant performance gains while maintaining low computational overhead.

In summary, our contributions are:

- We propose an enhanced 3D detection framework that explicitly leverages 2D auxiliary predictions to guide 3D query generation, thereby reducing uncertainty in attribute regression and improving the accuracy and stability of 3D perception.
- We introduce a lightweight adaptive feature fusion module that aligns and integrates multi-attribute features within a unified embedding space, achieving effective feature aggregation with minimal computational overhead.
- Extensive experiments on the nuScenes benchmark demonstrate that the proposed EAFF achieves state-of-the-art performance and exhibits superior robustness in crowded and long-range scenarios.

II. RELATED WORK

A. Sparse-based 3D Object Detection

Sparse query-based detectors have recently gained significant attention due to their superior computational efficiency. These methods can be broadly categorized based on their feature interaction mechanisms. Projection-based approaches, exemplified by DETR3D [13], define a set of sparse 3D queries and project them into the 2D image space to sample features using camera intrinsic and extrinsic parameters. However, due to the limited number of sampled points, such approaches struggle to achieve accurate global scene modeling. To alleviate this issue, position encoding-based methods, such as PETRv2 [6], employ global attention mechanisms for interaction between queries and image features and encode 3D positional information into 2D features to construct 3D-aware representations. Building upon this paradigm, StreamPETR [7] proposes an object-centric paradigm that propagates temporal information through object queries frame by frame, achieving effective spatiotemporal modeling with minimal overhead. Recent advancements have further focused on enhancing query representation quality. For example, RayDN [2] employs a ray denoising strategy to mitigate depth uncertainty, while Sparse4D [14] leverages deformable aggregation to fuse multi-timestamp information for enhanced temporal reasoning. Despite these strides, query initialization remains a critical bottleneck. Existing methods typically rely on predefined or learnable fixed queries, which often fail to effectively capture dense or long-range objects in complex driving scenarios.

B. 2D Auxiliary Tasks for 3D Perception

In recent years, leveraging 2D visual cues to enhance 3D object detection has emerged as a pivotal research direction. Several representative works have demonstrated the value of integrating 2D semantics into the 3D domain. For example, BEVFormer V2 [5] introduces a perspective head to extract perspective-aware features, which are subsequently encoded into object queries. Focal-PETR [9] and StreamPETR [7] exploit 2D detection scores to adaptively guide 3D queries toward foreground regions. Beyond implicit guidance, several methods explore explicit geometric integration between 2D and 3D tasks. SimMOD [12] generates perspective-aware candidate regions from monocular images and aggregates them following the DETR3D paradigm. MV2D [11] employs a 2D detector to extract region-of-interest (ROI) features for initializing 3D queries, while Far3D [8] jointly performs 2D detection and depth estimation to produce adaptive 3D query representations.

Despite their notable progress, existing approaches predominantly focus on positional and semantic cues, while largely neglecting object orientation cues that are readily observable in 2D images. Such omission at the query initialization stage introduces ambiguous directional priors, which complicates query-to-feature alignment and leads to unstable optimization. To alleviate this issue, our EAFF explicitly integrates both 2D attributes and 3D attributes into an adaptive query construction process. By embedding orientation-aware priors alongside spatial and semantic cues, EAFF provides a more informative and structured initialization, thereby stabilizing the convergence of 3D attribute learning.

III. METHODOLOGY

A. Overview

As illustrated in Fig. 2, we present an enhanced 3D detection pipeline that explicitly leverages 2D priors to facilitate robust query initialization. The overall architecture consists of three integral stages: feature extraction, adaptive query generation, and 3D refinement. Specifically, multi-view images are first processed by a backbone network (e.g., ResNet [15]) equipped with a feature pyramid network (FPN) [16] to extract multi-scale feature representations. To overcome the limitations of fixed query initialization, these extracted multi-scale features are fed into a lightweight 2D auxiliary branch. This module predicts pixel-wise attributes, such as object confidence, depth, and orientation, serving as dense semantic priors. Subsequently, our adaptive feature fusion module aligns these 2D attribute embeddings with geometry-aware positional encodings derived from camera projection. This process constructs a set of informative adaptive 3D queries equipped with explicit directional and spatial awareness. Finally, these initialized queries are fed into a standard Transformer decoder, which iteratively refines the predictions to generate the final 3D object detection results.

B. 2D Auxiliary Network

The 2D auxiliary network is designed to provide reliable semantic and geometric priors for adaptive 3D query construction, rather than serving as an independent detection module. Sharing the same backbone with the 3D detection network, the auxiliary network adopts a multi-task learning paradigm to jointly predict object confidence, class probabilities, bounding boxes, orientation, and depth.

Specifically, we adopt the YOLOX framework to generate 2D bounding boxes and object confidence scores, and further introduce lightweight convolutional branches for orientation

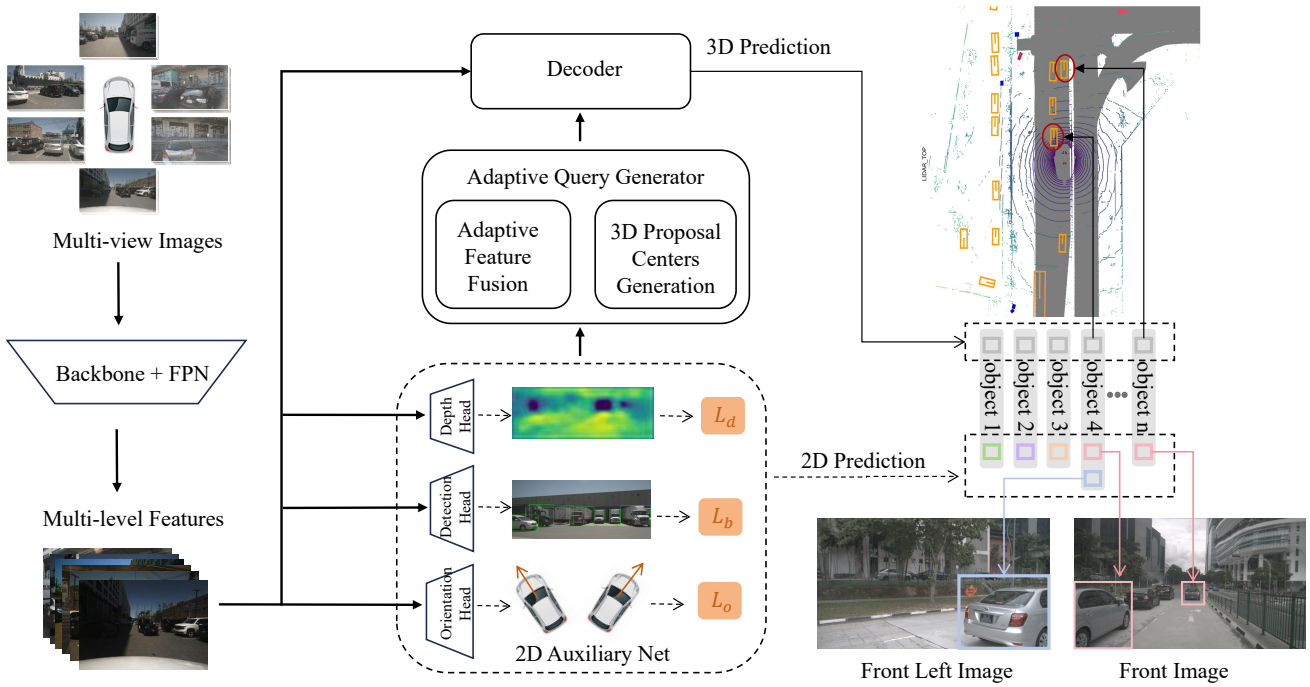

Fig. 2. The overview of our proposed EAFF. Feeding multi-view images into the backbone and FPN neck, we obtain multi-scale image features. Subsequently, these features are fed into a 2D auxiliary detector to obtain object-related representations. This information is leveraged by our adaptive query generation module to construct corresponding 3D adaptive queries. These generated queries, along with the initial 3D global queries, are iteratively refined by decoder layers to predict the final 3D attributes.

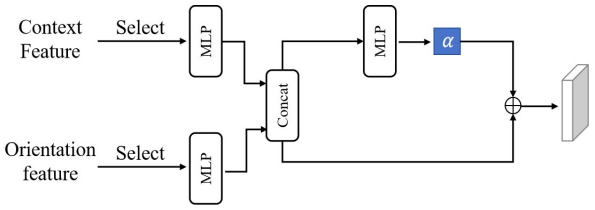

Fig. 3. Illustration of the Adaptive Feature Fusion

and depth estimation. However, due to viewpoint variations across different cameras, the same object may exhibit significantly different visual orientations, making it ineffective to directly regress a global heading angle.

To address this issue, we explicitly model a localized observation angle for each camera view. Given the global heading yaw in the LiDAR coordinate system, we first construct the corresponding yaw rotation matrix Y_{Lidar} as:

$$Y_{\text{Lidar}} = \begin{bmatrix} \cos(\text{yaw}) & -\sin(\text{yaw}) & 0 \\ \sin(\text{yaw}) & \cos(\text{yaw}) & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{1}$$

This global orientation is then transformed into the local coordinate system of the i-th camera via the extrinsic rotation matrix R_i:

$$Y_{\text{cam}_i} = R_i \cdot Y_{\text{Lidar}}, \tag{2}$$

In addition to the object's heading, the viewing direction of the camera also influences the perceived orientation. Therefore, we compute the ray angle between the object center and the camera optical axis. Let C_{Lidar} and C_{cam} represent the object center coordinates in the LiDAR and camera systems, respectively, where $C_{\text{cam}} = R_i \cdot C_{\text{Lidar}}$. The ray angle β is defined as:

$$\beta = \arctan 2(x, z), \tag{3}$$

where x and z denote the horizontal and depth coordinates in the camera coordinate system. The final observation angle α, which characterizes the relationship between the object's heading direction and the camera's viewing ray on the BEV plane, is defined as:

$$\alpha = \beta + \text{yaw}_{\text{cam}_i}, \tag{4}$$

where $\text{yaw}_{\text{cam}_i}$ represents the local heading angle derived from the rotation matrix Y_{cam_i}. Moreover, to handle the periodicity of angular representation, we regress the sine and cosine components of the orientation angle instead of directly predicting the angle value:

$$\sin(\theta), \cos(\theta) = \text{Conv}(p_{uv}), \tag{5}$$

where p_{uv} denotes the pixel-wise feature.

For depth estimation, we discretize the depth range into K predefined bins $\{d_1, d_2, \ldots, d_K\}$ and formulate depth prediction as a classification task. The probability distribution over

TABLE I
PERFORMANCE COMPARISON ON THE NUSCENES VAL SPLIT

Method	Epochs	NDS ↑	mAP ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓
PETRv2 [6]	60	45.6	34.9	0.700	0.275	0.580	0.437	0.187
BEVStereo [17]	90	50.0	37.2	0.598	0.270	0.438	0.367	0.190
BEVPoolv2 [18]	90	52.6	40.6	0.572	0.275	0.463	0.275	0.188
Sparse4Dv2 [19]	100	53.9	43.9	0.598	0.267	0.475	0.282	**0.179**
StreamPETR [7]	60	55.0	45.0	0.613	0.267	0.413	0.265	0.196
SparseBEV [10]	36	55.8	44.8	0.581	0.271	0.373	**0.247**	0.190
RayDN [2]	24	54.2	45.0	0.606	0.269	0.480	0.271	0.202
EAFF (Ours)	24	55.7	46.3	0.586	0.266	0.408	0.280	0.202
EAFF (Ours)	36	**57.0**	**47.2**	**0.569**	**0.265**	**0.362**	0.261	0.200

depth bins is obtained via a lightweight convolutional network followed by a softmax operation:

$$d_{uv} = \text{Argmax}\left(\text{Softmax}(\text{Conv}(p_{uv}))\right). \tag{6}$$

This design provides stable depth priors for subsequent 3D query initialization.

C. Adaptive Feature Fusion

A naive fusion of semantic and orientation features often leads to feature interference due to their inconsistent attribute distributions, which may degrade the effectiveness of query representations.

To effectively integrate different types of 2D attributes, we propose a lightweight adaptive feature fusion module that aligns semantic and orientation features into a unified embedding space.

We first compute a pixel-wise importance score by combining object confidence and class probability:

$$\mathbf{S}_{uv} = \mathbf{P}_{obj} \odot \max(\mathbf{P}_{cls}, \text{dim} = 1), \tag{7}$$

where $\odot$ denotes the Hadamard product. To suppress redundant responses, we apply a 3×3 max-pooling operation as a local non-maximum suppression. A set of valid indices $\mathbf{I}_v$ is then obtained by selecting pixels whose confidence scores exceed a predefined threshold τ:

$$\mathbf{I}_v = \text{MaxPool}(\mathbf{S}_{uv}) > \tau. \tag{8}$$

Using these indices, semantic features p_{uv} and orientation features o_{uv} are extracted and projected into a shared embedding space via MLPs:

$$\mathbf{e}_p = \text{MLP}(\mathbf{I}_v(p_{uv})), \tag{9}$$

$$\mathbf{e}_o = \text{MLP}(\mathbf{I}_v(o_{uv})). \tag{10}$$

The resulting embeddings are concatenated and further refined by a fusion MLP with a residual connection:

$$\mathbf{Q}_{sem} = \text{LN}\left(\lambda \cdot \text{MLP}([\mathbf{e}_p, \mathbf{e}_o]) + [\mathbf{e}_p, \mathbf{e}_o]\right), \tag{11}$$

where λ is a learnable scalar that adaptively balances the contribution of different attributes. This design enables effective feature alignment while introducing negligible computational overhead.

D. 3D Adaptive Query Generation

Introducing informative priors at the query initialization stage can significantly constrain the solution space of subsequent regression tasks, thereby stabilizing optimization and improving convergence.

To explicitly incorporate such priors into the detection process, each adaptive 3D query consists of two components: positional embedding and semantic embedding. Using the valid indices, we extract the pixel-wise 2D bounding box center (c_u, c_v) and depth d_{uv}, which are combined to form a 3D proposal center:

$$\mathbf{c}_{3d} = R_i^{-1} K_i^{-1} \begin{bmatrix} c_u \cdot d_{uv} \\ c_v \cdot d_{uv} \\ d_{uv} \\ 1 \end{bmatrix}, \tag{12}$$

where R_i and K_i denote the camera extrinsic and intrinsic matrices, respectively.

The 3D proposal centers are encoded into positional embeddings as follows:

$$\mathbf{Q}_{pos} = \text{PosEmbed}(\mathbf{c}_{3d}), \tag{13}$$

where PosEmbed$(\cdot)$ consists of a sinusoidal transformation and an MLP. Finally, the adaptive 3D query is constructed by combining positional and semantic embeddings:

$$\mathbf{Q} = \mathbf{Q}_{pos} + \mathbf{Q}_{sem}. \tag{14}$$

By generating an additional set of sparse queries at potential object locations and initializing them with extracted prior features, the proposed adaptive queries provide strong prior information. This design facilitates stable training in subsequent 3D feature regression and detection tasks, leading to improved overall performance.

IV. EXPERIMENTS

A. Dataset and Metrics

To validate the effectiveness and generalization of the proposed framework, we evaluate our method using the nuScenes dataset. This dataset comprises 1000 driving scenes, each spanning 20 seconds and annotated at a frequency of 2Hz. The sensor rig captures the full 360° field of view (FOV), providing a rich context for 3D perception. The dataset contains a total of

1.4M annotated 3D bounding boxes for 10 classes. Consistent with official splits, the scenes are divided into training, validation, and testing sets. Following prior work, we adopt the nuScenes Detection Score (NDS) and mean Average Precision (mAP) as the primary metrics for detection. Additionally, we report five True Positive (TP) metrics to measure specific error types: 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).

B. Implementation Details

For implementation, we build our proposed EAFF method based on the RayDN framework and employ ResNet-50 as the backbone. Consistent with prior works, the backbone is initialized with weights pre-trained on nuImages [1], and the performance was primarily evaluated on the nuScenes val set.

During the training phase, the models are optimized using the AdamW [20] optimizer with a total batch size of 8 and a weight decay of 0.01. We set the initial learning rate to 2×10^{-4} and employ a cosine annealing policy for learning rate scheduling. Regarding the training schedules, the models in the ablation study are trained for 24 epochs for efficiency, while the models are trained for 36 epochs when compared with other state-of-the-art methods to ensure convergence. No test-time augmentation (TTA) or CBGS strategy is used in our experiments.

C. Comparisons with State-of-the-Art Methods

We compare our proposed EAFF with state-of-the-art multi-view 3D object detection methods on the nuScenes validation split. As detailed in Tab. I, EAFF demonstrates superior performance on mAP, NDS, mATE, mASE, and mAOE. Using an input resolution of 704×256 with a ResNet50 backbone, EAFF outperforms the baseline model, RayDN, by 1.5% in NDS and 1.3% in mAP. Furthermore, when trained for 36 epochs, compared with SparseBEV, which utilizes 8 frames for temporal information, our method achieves a 2.4% improvement in mAP and 1.2% improvement in NDS. These results validate the effectiveness of our proposed framework.

D. Ablation Study

Efficient Feature Fusion Module. To analyze the effect of adaptive feature fusion on 3D query construction, we conduct an ablation study comparing several representative fusion strategies, as summarized in Tab. II. The Add strategy fuses orientation and content features via element-wise addition while the Concat strategy concatenates features along the channel dimension followed by an MLP. As shown in the results, our proposed fusion module consistently outperforms both Add and Concat in terms of mAP and NDS, while introducing only a marginal increase in parameters, demonstrating its effectiveness and efficiency for adaptive query construction.

Efficient Adaptive Query Components. As shown in Tab. III, introducing position-aware adaptive queries (Row II) over the baseline setting (Row I) yields a 0.5% improvement in mAP and a 0.2% gain in NDS, while mAOE increases by 2.6%. This suggests that positional priors help improve object recall but provide limited semantic support for accurate orientation estimation. On top of this, incorporating content features (Row III) leads to an additional 0.5% increase in mAP and a 3.7% improvement in mAOE, demonstrating the benefit of enhanced semantic representation. Further adding explicit orientation supervision (Row IV) brings an additional 0.3% gain in mAP and a substantial 6.1% improvement in mAOE, confirming its effectiveness in reducing orientation error. Overall, adaptive query initialization improves recall in challenging scenarios, while explicit orientation-aware supervision significantly enhances orientation accuracy, leading to consistent gains in both mAP and NDS.

TABLE II
ABLATION STUDY OF FUSION METHODS FOR 3D ADAPTIVE QUERY CONSTRUCTION

#	Fusion Method			mAP↑	NDS↑	Params↓
	Add	Concat	Fusion			
I	✓			0.459	0.555	**0.76M**
II		✓		0.459	0.556	1.51M
III			✓	**0.463**	**0.557**	1.51M

TABLE III
ABLATION STUDY OF FUSION COMPONENTS FOR 3D ADAPTIVE QUERY CONSTRUCTION

#	3D Adaptive Query			mAP↑	NDS↑	mAOE↓
	Pos	Context	Orien			
I				0.450	0.542	0.480
II	✓			0.455	0.544	0.506
III	✓	✓		0.460	0.547	0.469
IV	✓	✓	✓	**0.463**	**0.557**	**0.408**

Visualization of Detection Results. We present a qualitative comparison between the baseline and our EAFF in Fig. 4. In the BEV plots, red bounding boxes represent ground truth, while blue boxes denote model predictions. As evident in the zoomed-in regions, the baseline model tends to suffer from missed detections in challenging scenarios, such as long-range targets or severe occlusions. In contrast, the proposed EAFF effectively recovers these missed objects. This improvement is attributed to the 2D auxiliary network, which generates adaptive queries that serve as informative priors during initialization. By compensating for the lack of explicit geometric cues in difficult conditions, our method enables robust detection even where the baseline fails.

V. CONCLUSION

In this paper, we propose EAFF, an effective framework that enhances multi-view 3D object detection by incorporating 2D auxiliary cues into adaptive query modeling. The proposed approach leverages a lightweight 2D auxiliary network to extract object-related attributes and employs an adaptive feature fusion mechanism to construct informative 3D queries. By injecting orientation-aware priors at the query initialization

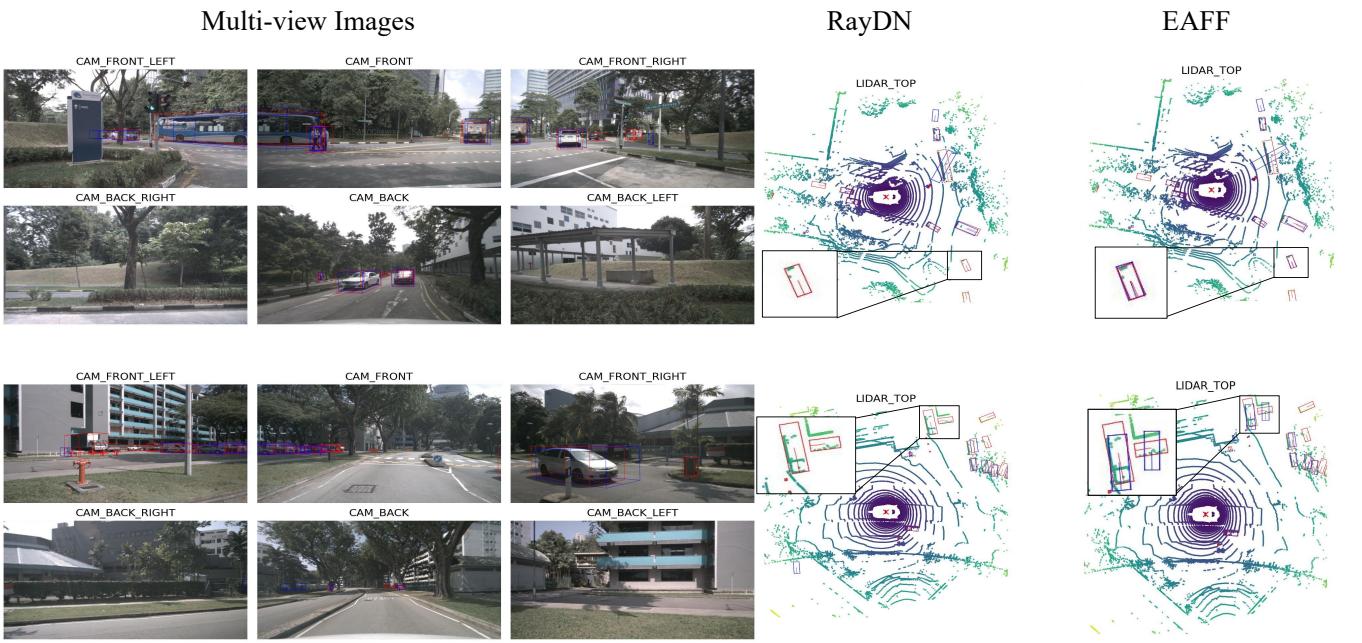

Fig. 4. Visualization results of EAFF and baseline. On the BEV plane (right), the ground truth and predictions are drawn in red and blue rectangles respectively. Observations reveal that the baseline model tends to miss predictions for distant and densely packed objects. In contrast, our proposed EAFF method is capable of accurately detecting objects in such challenging scenarios

stage, EAFF provides more structured and reliable representations for subsequent 3D reasoning. Extensive experiments on the nuScenes benchmark demonstrate the effectiveness and robustness of the proposed approach.

REFERENCES

[1] H. Caesar, V. Bankiti, A. H. Lang, S. Vora *et al.*, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[2] F. Liu, T. Huang, Q. Zhang, H. Yao, C. Zhang, F. Wan, Q. Ye, and Y. Zhou, "Ray denoising: Depth-aware hard negative sampling for multi-view 3d object detection," in *European Conference on Computer Vision*. Springer, 2024, pp. 200–217.

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

[4] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[5] C. Yang, Y. Chen, H. Tian *et al.*, "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*, 2023, pp. 17 830–17 839.

[6] Y. Liu, J. Yan, F. Jia, S. Li, A. Gao, T. Wang, and X. Zhang, "Petrv2: A unified framework for 3d perception from multi-camera images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 3262–3272.

[7] S. Wang, Y. Liu, T. Wang, Y. Li, and X. Zhang, "Exploring object-centric temporal modeling for efficient multi-view 3d object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 3621–3631.

[8] X. Jiang, S. Li, Y. Liu, S. Wang, F. Jia, T. Wang, L. Han, and X. Zhang, "Far3d: Expanding the horizon for surround-view 3d object detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 3, 2024, pp. 2561–2569.

[9] S. Wang, X. Jiang, and Y. Li, "Focal-petr: Embracing foreground for efficient multi-camera 3d object detection," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 1481–1489, 2023.

[10] H. Liu, Y. Teng, T. Lu, H. Wang, and L. Wang, "Sparsebev: High-performance sparse 3d object detection from multi-camera videos," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 18 580–18 590.

[11] Z. Wang, Z. Huang, J. Fu, N. Wang, and S. Liu, "Object as query: Lifting any 2d object detector to 3d detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3791–3800.

[12] Y. Zhang, W. Zheng, Z. Zhu, G. Huang, J. Lu, and J. Zhou, "A simple baseline for multi-camera 3d object detection," in *Proceedings of the AAAI Conference on artificial intelligence*, vol. 37, no. 3, 2023, pp. 3507–3515.

[13] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "Detr3d: 3d object detection from multi-view images via 3d-to-2d queries," in *Conference on robot learning*. PMLR, 2022, pp. 180–191.

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

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

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

[17] Y. Li, H. Bao, Z. Ge *et al.*, "Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 1486–1494.

[18] J. Huang and G. Huang, "Bevpoolv2: A cutting-edge implementation of bevdet toward deployment," *arXiv preprint arXiv:2211.17111*, 2022.

[19] X. Lin, T. Lin, Z. Pei, L. Huang, and Z. Su, "Sparse4d v2: Recurrent temporal fusion with sparse model," *arXiv preprint arXiv:2305.14018*, 2023.

[20] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.