This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

RaCFormer: Towards High-Quality 3D Object Detection via Query-based Radar-Camera Fusion

Xiaomeng Chu[1], Jiajun Deng[2*], Guoliang You[1], Yifan Duan[1], Houqiang Li[1], Yanyong Zhang[1,3*]

[1] University of Science and Technology of China, [2] The University of Adelaide,
[3] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center

Abstract

*We propose **Ra**dar-**C**amera fusion trans**former** (RaCFormer) to boost the accuracy of 3D object detection by the following insight. The Radar-Camera fusion in outdoor 3D scene perception is capped by the image-to-BEV transformation–if the depth of pixels is not accurately estimated, the naive combination of BEV features actually integrates unaligned visual content. To avoid this problem, we propose a query-based framework that enables adaptive sampling of instance-relevant features from both the bird's-eye view (BEV) and the original image view. Furthermore, we enhance system performance by two key designs: optimizing query initialization and strengthening the representational capacity of BEV. For the former, we introduce an adaptive circular distribution in polar coordinates to refine the initialization of object queries, allowing for a distance-based adjustment of query density. For the latter, we initially incorporate a radar-guided depth head to refine the transformation from image view to BEV. Subsequently, we focus on leveraging the Doppler effect of radar and introduce an implicit dynamic catcher to capture the temporal elements within the BEV. Extensive experiments on nuScenes and View-of-Delft (VoD) datasets validate the merits of our design. Remarkably, our method achieves superior results of 64.9% mAP and 70.2% NDS on nuScenes. RaCFormer also secures the state-of-the-art performance on the VoD dataset. Code is available at https://github.com/cxmomo/RaCFormer.*

1. Introduction

Precise 3D object detection plays a vital role in promoting the safety and efficiency of autonomous vehicles and intelligent robotic systems [1, 38, 39, 44]. Compared to adopting the expensive LiDAR sensor, the solution with multi-view cameras and millimeter-wave radar is more applicable due to the dramatically reduced cost, thus attracting a surge of research interests in the community [40, 50].

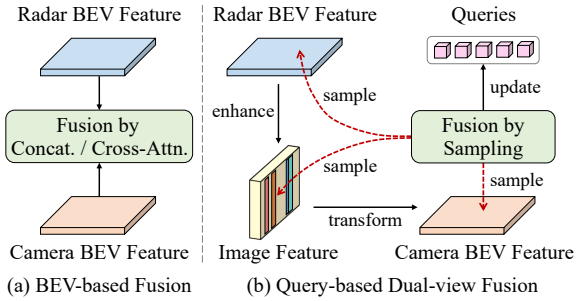


Figure 1. Motivation of RaCFormer. (a) Previous methods typically fuse BEV features from image-view transformation and radar point cloud encoding, by concatenation or cross-attention. (b) Instead, RaCFormer uses a query-based fusion framework by simultaneously sampling radar-enhanced image-view features, camera-transformed BEV features, and radar-encoded BEV features.

Despite impressive advancement, it is still a non-trivial problem in approaching the detection accuracy of LiDAR-based methods by the combination of camera and radar data. Current top-performing radar-camera fusion approaches [18, 29, 53, 54] typically adopt the BEV-based fusion framework [34] which unifies the representation in BEV space to facilitate fusion. In this paradigm, image and radar features are independently extracted, unified into a BEV representation, and subsequently fused by concatenation or cross-attention, as shown in Fig. 1 (a). However, the inherent disparity between the two modalities poses a significant challenge when relying solely on BEV features for fusion. Due to radars' hardware constraints, e.g., bandwidth and array design, their limited spatial resolution results in sparse radar BEV features. On the other hand, camera BEV features are generated from dense image features, but have the issue of feature distortion due to the inaccurate depth estimation in view transformation [43]. In contrast, the original perspective image features offer a semantically rich representation free from distortions, which has the po-

*Corresponding authors.

tential to aid camera-radar fusion. This underscores the necessity of heterogeneous feature fusion across perspectives of the front view and BEV. Inspired by this finding, this work aims to explore an effective cross-perspective fusion framework that can accommodate resolution and semantic discrepancies between the two modalities.

Motivated by the above analysis, we ask one essential question: What kind of fusion paradigm can remain unaffected by feature density while effectively utilizing information from different views? Back to the detection algorithms, the query-based approach [4] shows potential in addressing this dilemma. Specifically, the object query initialized in the 3D space can be leveraged as a medium for abstracting features from arbitrary projection views.

Formally, we propose RaCFormer, a query-based radar-camera fusion framework that improves camera-radar fusion by sampling object-relevant features from both perspective and bird's-eye views, as illustrated in the dual-view fusion paradigm in Fig. 1 (b). Our framework has three main designs: linearly increasing circular query initialization, radar-aware depth prediction, and temporal radar BEV encoding. The first optimizes the initialization distribution of queries, while the latter two refine and bolster the BEV representation. Specifically, we propose a circular query initialization strategy that places query points along concentric circles to align the projection principle of sensors. Additionally, we ensure a linear increase in the number of queries from inner to outer circles, thereby mitigating the issue of queries being much sparser at distant ranges compared to nearby areas. Furthermore, conventional automotive radars exhibit significant height estimation errors due to their limited vertical angular resolution. Therefore, we assign a default height and project radar points onto the image features to enhance depth prediction for view transformation, refining the camera BEV features. We also utilize the radar's Doppler effect to track object velocities by employing an implicit dynamic catcher with convolutional gated recurrent units, effectively capturing temporal elements on multi-frame radar BEV features.

To demonstrate the effectiveness of our proposed RaCFormer, we benchmark our method on the challenging nuScenes [3] dataset and the View-of-Delft (VoD) [41] dataset. Without bells and whistles, our approach achieves 64.9% mAP and 70.2% NDS on the nuScenes test set. Remarkably, on the VoD dataset, our method achieves a 54.4% mAP across the entire annotated area and a 78.6% mAP in the region of interest, earning a 1st-ranking performance.

In summary, our main contributions are as follows:

- We introduce RaCFormer, an innovative query-based 3D object detection method through cross-perspective radar-camera fusion. Object queries are initialized to a linearly increasing circular distribution, which aligns with camera projection principles and ensures reasonable density.
- On the image view, we refine depth estimation using the radar-aware depth head, facilitating more accurate transformations from the image plane to the BEV. Concurrently, on the BEV, we bolster the motion perception of radar BEV features with the implicit dynamic catcher.
- We perform experiments on the nuScenes and VoD datasets. Our method achieves state-of-the-art performance on both the nuScenes test set and the VoD dataset.

2. Related Work

Camera-based 3D Object Detection: Multi-camera 3D object detection methods fall mainly into BEV-based and query-based categories. Notable BEV-based approaches, such as BEVDet [15] and BEVDepth [25], apply the lift-splat-shoot method [43] to transform the image view into a top-down perspective. BEVFormer [26] uses deformable cross-attention for the construction of BEV features and integrates temporal data. FB-BEV [27] improves BEV representations with a forward-backward view transformation module, while HOP [55] employs temporal decoders to predict objects using pseudo-BEV features to capture dynamics. VideoBEV [12] stands out with its long-term recurrent fusion technique, seamlessly incorporating historical data. On the other hand, query-based methods such as DETR3D [46] and PETR [32] harness the transformer decoder to interpret image features. StreamPETR [45] extends PETR with an object-centric temporal mechanism for long-sequence modeling, using frame-by-frame object query propagation. MV2D [47] enhances detection capabilities by using 2D detectors to generate object-specific queries, while RayDN [30] improves detection precision by strategically sampling camera rays to generate depth-aware features. Sparse4D [28] refines the anchor boxes through sparse feature sampling, assigning multiple 4D key points to each 3D anchor. Lastly, SparseBEV [31] introduces a fully sparse 3D detection framework, fusing scale-adaptive attention with adaptive spatio-temporal sampling.

Radar-Camera Fusion-based 3D Object Detection: In pursuit of precise 3D object detection, various innovative sensor fusion techniques have emerged [2, 7, 24, 49, 51, 53], with radar-camera fusion gaining widespread research attention. CRN [18] generates a detailed BEV feature map by integrating camera and radar data, transforming image features into BEV, and applying multi-modal deformable attention to resolve spatial misalignment. Subsequently, HVDetFusion [22] accommodates both camera-only and radar-camera inputs, enhancing BEVDet4D [14] for camera streams and refining radar data with object priors to supplement and fuse the BEV features. CRAFT [17] proposes an early fusion strategy at the proposal level, combining spatial and contextual data from cameras and radars. Meanwhile, RADIANT [35] corrects monocular depth er-

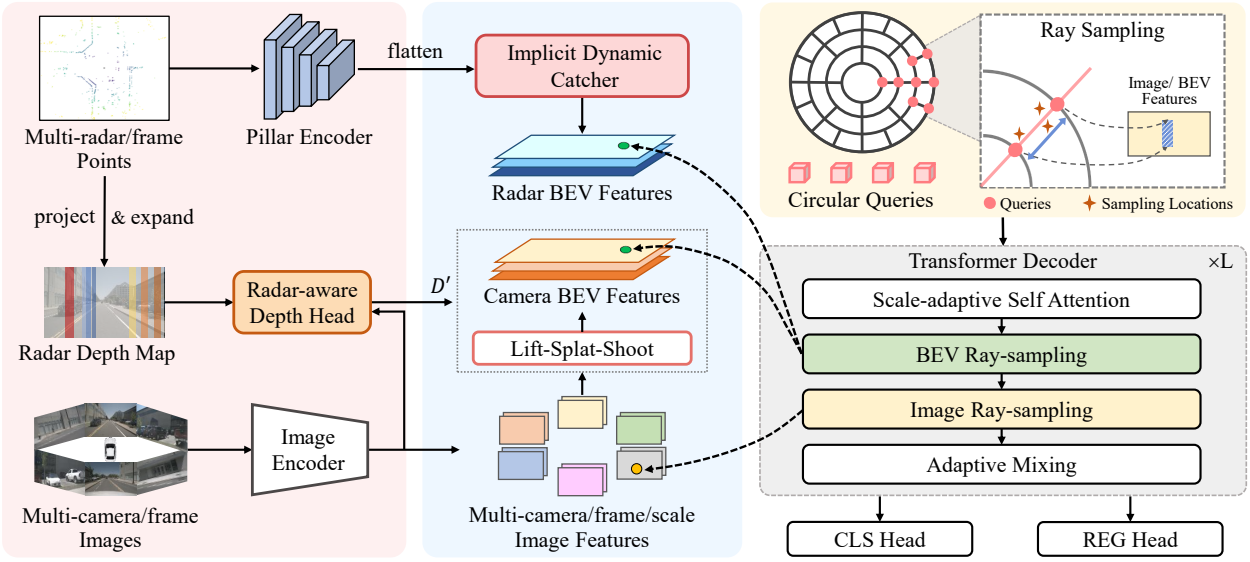


Figure 2. Overall architecture of RaCFormer. The image encoder extracts features from multiple frames of multi-camera images, while multi-frame radar points are voxelized and processed by a pillar encoder. The radar features are flattened into the BEV and enhanced by an implicit dynamic catcher. Simultaneously, radar points are re-projected onto the image plane, with their depth values extended to the full image height, and merged with image features in the depth head to refine depth prediction. The refined depth probability distribution D' and the image features are then input into the lift-splat-shoot (LSS) module to create camera BEV features. The transformer decoder initializes queries with an adjustable circular distribution. Over L layers, a ray sampling module within each layer extracts both image-view and BEV features to refine queries, enabling precise classification and regression by the subsequent heads.

rors by predicting 3D offsets between radar returns and object centers, improving accuracy without retraining existing models. RCBEVDet [29] introduces RadarBEVNet, a pioneering module for radar feature extraction in BEV, coupled with a fusion mechanism that autonomously aligns the multi-modal BEV features Lastly, HyDRa [48] employs a hybrid approach to merge camera and radar features in both perspective and BEV spaces, including a height association transformer for reliable depth estimation.

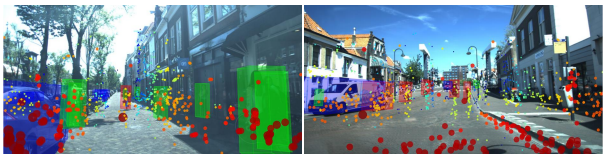
3. Method

3.1. Overall Framework

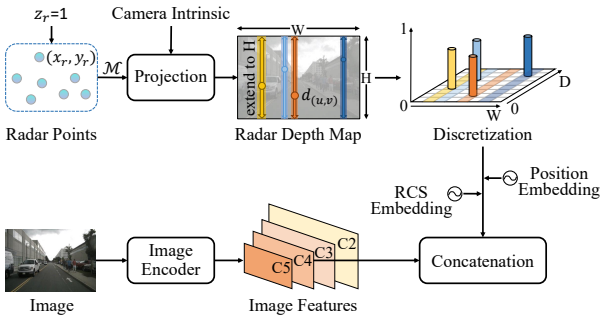
RaCFormer, as shown in Fig. 2, offers a query-based 3D object detection framework that integrates radar and camera inputs. The core modules of the framework include an image encoder, a pillar encoder, a radar-aware depth head, an LSS view transformation module, an implicit dynamic catcher, and a transformer decoder. The image encoder extracts features from camera frames, while the pillar encoder processes radar points and flattens the features to BEV. Subsequently, the radar BEV features are refined by an implicit dynamic catcher to capture moving elements. Radar points are also projected into the image plane and combined with visual features in the radar-aware depth head to form a depth probability distribution D'. The enhanced depth distribution is merged with image features in the LSS module to generate camera BEV features. Queries serve as a medium for cross-perspective and cross-modality feature fusion, initialized in an adjustable circular distribution and refined by the transformer decoder. The transformer decoder, comprising L layers, includes a scale-adaptive self-attention module [31] for dynamically adjusting receptive fields, two ray-sampling modules for extracting BEV and image-view features, and an adaptive mixer [9] for feature aggregation. Finally, the classification and regression heads interpret the refined queries for accurate object detection.

3.2. Camera-transformed BEV Feature Generation with Radar-aware Depth Prediction

Radar-aware Depth Prediction: We propose to enhance the image features using radar data for better depth estimation. However, conventional automotive radars provide distance and velocity measurements within their field of view, but their limited vertical angular resolution leads to significant height estimation errors. As depicted in Fig. 3(a), many radar points are projected onto the image with their vertical coordinates falling outside the objects' 2D bounding boxes, due to inaccuracies in the z-coordinates of the raw radar points. Therefore, we design a pre-processing step before the depth head, shown in Fig. 3(b). To maximize the number of radar points falling onto the image's field of view, we initially set $z_r = 1$ for all points, denoted as (x_r, y_r, z_r), and then project them onto the image plane using the camera's intrinsic parameters. The specific transformation for-



(a) Radar points projected onto the image



(b) The pre-processing of the radar-aware depth head

Figure 3. The visualization of (a) radar points with raw z-coordinates projected onto the image and the flowchart of (b) pre-processing input data for the radar-aware depth head.

mula is as follows:

$$
\begin{aligned}
(x_c, y_c, d) &= \mathcal{M} \cdot (x_r, y_r, z_r), \quad z_r = 1, \\
u &= \frac{f_x \cdot x_c}{d} + c_x, \quad v = \frac{f_y \cdot y_c}{d} + c_y,
\end{aligned} \tag{1}
$$

where $\mathcal{M}$ is the transformation matrix mapping radar coordinates to camera coordinates. The focal lengths f_x and f_y correspond to the camera's x and y axes, respectively, while c_x and c_y specify the image's principal point location.

Next, we extend the vertical coordinate of each projected point to the full height H of the image and assign its depth value, creating a coarse radar depth map. We then use a spacing-increasing discretization strategy [8] to discretize these depths within the range $[0, \mathrm{D}]$. Furthermore, the Radar Cross Section (RCS) attribute indicates an object's detectability. We embed RCS and the pixel position of the projected radar points into the discretized depth to generate comprehensive radar-aware features, which are concatenated with the $16\times$ downsampled image features $C4$ and input into the depth head.

Camera-transformed BEV Feature Generation: We follow the methods established by BEV-based 3D object detection works [15, 25] and employ the lift-splat-shoot [43] approach for transformation from image view to BEV. The process begins with lifting the 2D image features into a 3D space using discretized depth. The lifted features are then splatted or distributed onto the BEV plane according to their 3D positions. The shooting step involves rendering the BEV features for subsequent perception tasks.

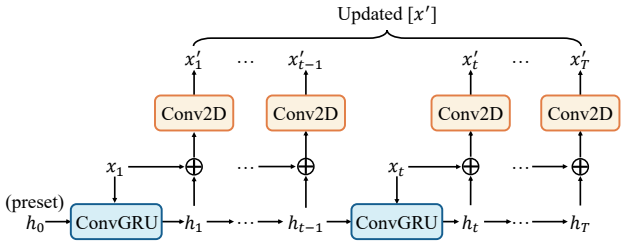


Figure 4. The structure of our implicit dynamic catcher. h_t represents the hidden state at time t, with h_0 being a preset value of zeros. x_t denotes the BEV features output by the pillar encoder at time t, while x'_t indicates the updated BEV features from x_t.

3.3. Radar-encoded BEV Feature Generation with Implicit Dynamic Catcher

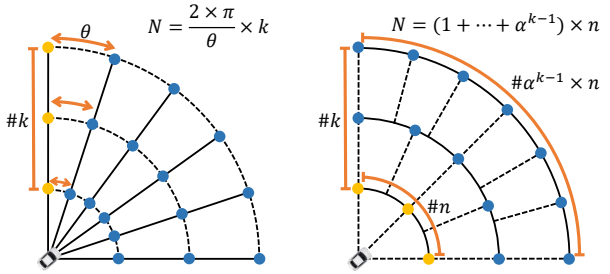
Implicit Dynamic Catching: Millimeter-wave radars leverage the Doppler effect for velocity measurement of moving objects. To harness this, we introduce an implicit dynamic catcher module designed to capture the temporal elements from multi-frame radar-derived BEV features. The ConvGRU, an extension of the GRU that integrates convolutional layers, excels at processing sequential data while discerning spatial hierarchies. This makes it an ideal core component for our implicit dynamic catcher, as depicted in Fig. 4. Specifically, the dynamic catcher involves accumulating hidden states across consecutive frames $0 \sim T$. For instance, the BEV feature of the t-th frame, x_t, along with the previous frame's hidden state, h_{t-1}, are fed into the ConvGRU. This process yields the current frame's hidden state, h_t. Subsequently, h_t is combined with x_t, and goes through a 2D convolutional layer to produce the refined BEV feature x'_t. The process is expressed as follows:

$$
\begin{aligned}
h_t &= \text{ConvGRU}(x_t, h_{t-1}), \\
x'_t &= \text{Conv2D}(h_t \oplus x_t).
\end{aligned} \tag{2}
$$

Radar-encoded BEV Feature Generation: RaCFormer processes raw radar data by encoding it in a manner analogous to LiDAR point clouds, utilizing a pillar-based method [19]. We begin by setting the z-coordinates of radar points to zero and then project them onto the BEV plane using their (x, y) coordinates. The BEV perception range is then segmented into small square pillars, each corresponding to a specific local area. Within each pillar, we apply a pillar feature network to process the enclosed point cloud data to generate local features. Finally, we construct a BEV feature map by performing max pooling across these pillars.

3.4. Query Initialization and Ray Sampling

Linearly Increasing Circular Query Distribution: The radial query initialization from RayFormer [6] mimics camera rays, reducing the overlap of different queries projecting



(a) Radial (b) Circular

Figure 5. Comparison of query initialization methods: (a) Radial distribution: Queries are evenly spaced along each ray, with a constant angle θ separating adjacent rays. (b) Linearly increasing circular distribution: The parameter n denotes the query count in the innermost circle, and the linear growth factor of each outer circle is α. The parameter k indicates the query count per ray in (a) and the number of concentric circles in (b).

on single objects. However, it results in dense queries near the camera and sparser coverage at greater distances, affecting far-object detection. To address this, we introduce a circular query initialization that linearly increases query density according to distance, with adjustable coefficients.

As depicted in Fig. 5(a), the radial distribution emits rays from the BEV center at uniform angles θ and places k queries per ray. The total query count N is given by:

$$N = \frac{2 \times \pi}{\theta} \times k. \tag{3}$$

Our proposed circular initialization method, as illustrated in Fig. 5(b), organizes queries in k concentric circles. Starting with n queries in the innermost circle, each subsequent circle contains α times the query count of the adjacent inner one, up to $\alpha^{k-1} \times n$ queries in the outermost circle. The total query count N is calculated as follows:

$$\begin{aligned} N &= (1 + \alpha + ... + \alpha^{k-1}) \times n \\ &= \begin{cases} k \times n, & \alpha = 1, \\ \frac{\alpha^k - 1}{\alpha - 1} \times n, & \alpha \neq 1. \end{cases} \end{aligned} \tag{4}$$

When $\alpha = 1$, all circles have equal query counts, making the method equivalent to the radial one in this specific case.

Ray Sampling Across Perspectives and Modalities: We employ the ray sampling method following RayFormer [6], which takes ray segments as units, reflecting the natural optical properties of cameras. In this approach, each query defines a segment whose length corresponds to the interval between adjacent circles. Within this segment, several adaptive sampling points are selected to gather features from the image view and the BEV. For BEV ray sampling, we integrate historical BEV features into the ego coordinate system and apply deformable attention. Image ray sampling involves projecting sampling points onto multi-camera images from multiple timestamps to extract pixel features. Finally, the adaptive mixing process [9, 31] aggregates the spatio-temporal features across channels and points.

4. Experiments

4.1. Datasets and Metrics

NuScenes: NuScenes [3], renowned for its extensive perception challenges, is meticulously divided into 1,000 instances, with 700 for training, 150 for validation, and 150 for testing. Each scene, annotated at 2Hz, provides a 20-second duration. The nuScenes evaluation metrics, including average translation error (ATE), average scale error (ASE), average orientation error (AOE), average velocity error (AVE), and average attribute error (AAE), assess the precision of object detection in terms of position, size, orientation, motion, and attributes. The mean Average Precision (mAP) and the nuScenes detection score (NDS) further measure the overall effectiveness of detection systems.

View-of-Delft (VoD): VoD [41] comprises 8693 frames, each containing synchronized and calibrated 64-layer LiDAR-, (stereo) camera-, and 3+1D radar-data, all captured in complex urban traffic scenarios. It features 123,106 3D bounding box annotations for a variety of moving and stationary objects, encompassing pedestrian, cyclist, and car. The evaluation criteria follow the KITTI [10, 11] dataset, which assesses the detection performance using the mean Average Precision metric of 3D bounding boxes.

4.2. Implementation Details

For the nuScenes dataset, our BEV perception area, encompassing a 65-meter radius circle, is segmented into $k = 6$ concentric circles. We initiate with $n = 80$ queries in the innermost circle, expanding outwards by a factor of $\alpha \approx 1.25$ per subsequent circle, culminating in 900 queries overall. Due to the use of monocular images, the perception area of VoD is a 55-meter radius sector spanning $3/4\pi$ radians, divided into $k = 8$ concentric arcs. We allocate $n = 30$ queries to the inner arc, maintaining the $\alpha \approx 1.25$ and totaling 600 queries. Our transformer decoder comprises 6 layers with shared weights for efficiency. Additionally, we adopt the query denoising strategy derived from PETRv2 [33] to accelerate convergence.

Our models are trained using the AdamW [37] optimizer with a global batch size of 8. We initiate the learning process with a learning rate of 2e-5 for the backbone and 2e-4 for other parameters, applying a cosine annealing [36] policy for rate adjustment. For image feature encoding, we adopt the standard networks ResNet [13] and VoVNet-99 (V2-99) [20]. In line with established practices [28, 31, 45], the ResNet parameters are pre-trained on nuImages [3], and the V2-99 parameters are pre-trained on DD3D [42] with

Methods	Input	Image Size	Backbone	Epochs	**mAP**↑	**NDS**↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
StreamPETR [45]	C	256×704	ResNet50	60	45.0	55.0	0.613	0.267	0.413	0.265	0.196
RayFormer [6]	C	256×704	ResNet50	36	45.9	55.8	0.568	0.273	0.425	0.261	0.189
HVDetFusion [22]	C+R	256×704	ResNet50	24	45.1	55.7	0.557	0.527	0.270	0.473	0.212
RCBEVDet [29]	C+R	256×704	ResNet50	12	45.3	56.8	0.486	0.285	0.404	0.220	0.192
CRN [18]	C+R	256×704	ResNet50	24	49.0	56.0	0.487	0.277	0.542	0.344	0.197
HyDRa [48]	C+R	256×704	ResNet50	20	49.4	58.5	0.463	0.268	0.478	0.227	0.182
RaCFormer (Ours)	C+R	256×704	ResNet50	36	**54.1**	**61.3**	0.478	0.261	0.449	0.208	0.180
StreamPETR [45]	C	512×1408	ResNet101	60	50.4	59.2	0.569	0.262	0.315	0.257	0.199
RayFormer [6]	C	512×1408	ResNet101	24	51.1	59.4	0.565	0.265	0.331	0.255	0.200
CRN [18]	C+R	512×1408	ResNet101	24	52.5	59.2	0.460	0.273	0.443	0.352	0.180
HyDRa [48]	C+R	512×1408	ResNet101	20	53.6	61.7	0.416	0.264	0.407	0.231	0.186
RaCFormer (Ours)	C+R	512×1408	ResNet101	24	**57.3**	**63.0**	0.476	0.261	0.428	0.213	0.183

Table 1. Comparison of different methods on the nuScenes `val` set. 'C' and 'R' represent camera and radar, respectively.

Methods	Input	Image Size	Backbone	Epochs	**mAP**↑	**NDS**↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
CenterPoint [52]	L	-	VoxelNet	20	60.3	67.3	0.262	0.239	0.361	0.288	0.136
VoxelNeXt [5]	L	-	Sparse CNNs	20	64.5	70.0	0.268	0.238	0.377	0.219	0.127
UVTR [23]	C	900×1600	V2-99	24	47.2	55.1	0.577	0.253	0.391	0.508	0.123
PolarFormer [16]	C	640×1600	V2-99	24	49.3	57.2	0.556	0.256	0.364	0.439	0.127
RayFormer [6]	C	640×1600	V2-99	24	55.5	63.3	0.507	0.245	0.326	0.247	0.123
RCBEVDet [29]	C+R	640×1600	V2-99	12	55.0	63.9	0.390	0.234	0.362	0.259	0.113
CRN [18]	C+R	640×1600	ConvNeXt-B	24	57.5	62.4	0.416	0.264	0.456	0.365	0.130
HyDRa [48]	C+R	640×1600	V2-99	20	57.4	64.2	0.398	0.251	0.423	0.249	0.122
RaCFormer (Ours)	C+R	640×1600	V2-99	24	**59.2**	**65.9**	0.407	0.244	0.345	0.238	0.132
HVDetFusion [22] (+8)	C+R	640×1600	InternImage-B	20	60.9	67.4	0.379	0.243	0.382	0.172	0.132
RaCFormer (+6)	C+R	640×1600	V2-99	24	**64.9**	**70.2**	0.358	0.240	0.329	0.179	0.119

Table 2. Comparison on the nuScenes `test` set. The VoVNet-99 (V2-99) [21] is pre-trained from DD3D [42] with extra data. 'L', 'C', and 'R' represent LiDAR, camera, and radar, respectively. "(+t)" indicates using future and historical frames, each by t frames.

additional datasets. Unless specifically indicated, training is conducted for a standard 24 epochs for all models.

4.3. Main Results

NuScenes Results: In Tab. 1 and 2, we compare our method with existing state-of-the-art 3D detection methods on the nuScenes validation and test sets. We include both camera-only and radar-camera fusion algorithms for a thorough comparison. We default to an 8-frame sequence with 0.5-second intervals for comparable analysis. On the validation set, RaCFormer equipped with a ResNet-50 backbone at a resolution of 256×704 surpasses HyDRa [48] by 4.7% in mAP and 2.8% in NDS, achieving an mAP of 54.1% and an NDS of 61.3%. When employing ResNet-101 with input dimensions of 512×1408, RaCFormer achieves an mAP of 57.3% and an NDS of 63.0%, representing a 3.7% increase in mAP and a 1.3% increase in NDS compared to HyDRa. On the test set, with Vovnet-99 as the backbone and 7 historical frames, our method reaches 59.2% in mAP and 65.9% in NDS, marking the corresponding enhancements of 1.8% and 1.7% over HyDRa. Furthermore, we enhance performance by using 6 past and 6 future frames. This results in a 64.9% mAP and 70.2% NDS, outperforming HVDetFusion with more input frames by 4.0% and 2.8%, respectively. Additionally, RaCFormer outperforms representative LiDAR-based methods like VoxelNeXt [5] and CenterPoint [52], partially bridging the modality gap.

VoD Results: We evaluate our method by calculating the 3D AP for cars, pedestrians, and cyclists across two regions: the entire annotated area and the region of interest. The results, as detailed in Tab. 3, show that RaCFormer notably enhances AP across most categories. Specifically, across the entire annotated area, RaCFormer achieves a 4.45% higher mAP compared to RCBEVDet. In the region of interest, RaCFormer leads with an mAP of 78.57%, marking an 8.77% improvement over RCBEVDet, thus demonstrating the state-of-the-art performance.

Visualization Results: Fig. 6 presents the qualitative detection outcomes of RaCFormer in both image view and BEV, showcasing its robustness across various challenging environments. These include adverse conditions such as rain and darkness, as well as object-filled scenarios.

4.4. Ablation Studies

Unless specified, we perform ablation studies using single-frame inputs with an image resolution of 256×704 and a ResNet-50 backbone. For a comprehensive comparison, we benchmark our model against two single-modality detec-



Figure 6. Qualitative analysis across varied scenarios—rainy, nighttime, and object-filled. Images (left) exhibit 3D bounding boxes in diverse colors for different categories, while the BEV radar point clouds (right) depict ground truth in green and predicted boxes in blue.

Methods	Input	AP in the Entire Annotated Area (%)				AP in the Region of Interest (%)			
		Car	Pedestrian	Cyclist	mAP	Car	Pedestrian	Cyclist	mAP
PointPillars [19]	R	37.06	35.04	63.44	45.18	70.15	47.22	85.07	67.48
RadarPillarNet [53]	R	39.30	35.10	63.63	46.01	71.65	42.80	83.14	65.86
RCFusion [53]	C+R	41.70	38.95	68.31	49.65	71.87	47.50	88.33	69.23
RCBEVDet [29]	C+R	40.63	38.86	**70.48**	49.99	72.48	49.89	87.01	69.80
RaCFormer (Ours)	C+R	**47.30**	**46.21**	69.80	**54.44**	**89.26**	**56.78**	**89.67**	**78.57**

Table 3. Comparison of 3D object detection results on VoD `val` set. The region of interest is the driving corridor located close to the ego-vehicle. The IoU thresholds for AP are set to 0.5 for cars, 0.25 for pedestrians, and 0.25 for cyclists.

tors: RayFormer [6] for camera-based detection and CenterPoint [52] for point-based detection.

Feature Decoding and Fusion: Tab. 4 illustrates the impact of feature decoding methods and perspective selection. The first two rows present the results of employing the standard BEVFusion paradigm [34] using the concatenation operation or deformable cross-attention for BEV feature fusion, combined with a center-head decoder. Switching to a transformer decoder with queries, focusing solely on the sampling of BEV features, yields a 1.9% enhancement in mAP and a 2.5% improvement in NDS, along with a significant reduction in mAVE but an increase in mATE. Further sampling of image-view features improves 3.1% at mAP and 4.1% at NDS, respectively.

Radar-aware Depth Prediction: In Tab. 5, we examine the impact of embedding radar points' depth and RCS value. The model that only incorporates radar depth embedding achieves an improvement of 0.7% in mAP and 1.2% in NDS. Similarly, the embedding of RCS enhances 0.5% mAP and 0.8% NDS. When both embeddings are used, the performance is further improved, with overall gains of 1.1% in mAP and 2.1% in NDS.

Implicit Dynamic Catching: In Tab. 6, we evaluate the performance of our implicit dynamic catcher (IDC) by evaluating mAP and mAVE for moving objects in the nuScenes validation set, divided by objects' velocity: static (0 m/s), slow (<5 m/s), and fast (>5 m/s) objects. Adding radar data without IDC substantially increases the relative mAP of slow and fast objects by 49.2% and 48.1%, while decreasing relative mAVE by 30.2% and 11.8%. With IDC, RaCFormer further improves relative mAP by 4.3% and 3.5% and relatively reduces mAVE by 2.9% and 2.0%. The integration of radar data and the IDC module reveals subtle temporal feature changes, leading to more pronounced optimization on slow-moving objects.

Fusion	Views	mAP↑	NDS↑	mATE↓	mASE↓	mAVE↓
Concat.	B	35.9	42.7	0.625	0.289	0.613
Def. Attn.	B	38.6	45.1	0.576	0.280	0.601
Queries	B	40.5	47.6	0.655	0.281	0.365
Queries	B+I	43.6	51.7	0.577	0.274	0.341

Table 4. Ablation study of the feature decoding. 'B' and 'I' correspond to the BEV and image view, respectively.

Linearly Increasing Circular Query Initialization: In Tab. 7, we evaluate the impact of hyper-parameters of circular distribution while keeping the total query count around 900. We first initialize the queries using the conventional grid distribution as a baseline (ID=A), then vary the linear

Depth	RCS	mAP↑	NDS↑	mATE↓	mASE↓	mAVE↓
		43.6	51.7	0.577	0.274	0.341
✓		44.3	52.9	0.560	0.268	0.319
	✓	44.1	52.5	0.564	0.269	0.332
✓	✓	44.7	53.8	0.558	0.270	0.313

Table 5. Ablation study about the depth and RCS embedding in the pre-processing step of radar-aware depth prediction.

Input	IDC	0 m/s	(0, 5] m/s		>5 m/s	
		mAP	mAP	mAVE	mAP	mAVE
C	✗	45.5	12.6	0.731	29.3	0.798
C+R	✗	49.4	18.8	0.510	43.4	0.704
C+R	✓	50.1	19.6	0.495	44.9	0.690

Table 6. Analysis of the implicit dynamic catcher (IDC) in detecting moving objects with 8-frame inputs.

ID	α	k	n	mAP↑	NDS↑	mATE↓	mAVE↓
A	-	-	-	42.4	52.4	0.582	0.329
B	1	6	150	43.9	53.2	0.557	0.313
C	2		15	43.5	52.4	0.576	0.329
D	1.25		80	44.7	**53.8**	0.558	0.313
E		4	155	44.1	52.6	0.569	0.337
F		8	45	**44.8**	53.6	0.549	0.339

Table 7. Ablation study on circular query initialization, varying the linear growth factor α, the number of concentric circles k, and the query count n in the innermost circle.

growth factor α and the number of circles k to study their effects. Setting α to 1 and k to 6 (ID=B) matches the radial initialization of RayFormer [6], improving 1.5% mAP compared to the grid distribution. Both $\alpha = 2$ (ID=C) and $k = 4$ (ID=E) lead to a performance dip, due to unreasonable initial query distribution of over-density in outer circles and low-density in the depth direction, respectively. Optimal performance is found with α at 1.25 (ID=D&F). We adopt $k = 6$ as our standard in this paper.

Weather and Light Conditions: To verify the impact of radar fusion, we categorize the nuScenes validation set into four scenarios based on weather and lighting: sunny, rainy, daytime, and nighttime. We compared our method with two benchmarks–CenterPoint utilizing LiDAR- and RayFormer utilizing camera-input. As shown in Tab. 8, RaCFormer surpasses both baselines in all scenarios. Specifically, our method achieves an 8.5% higher mAP in sunny conditions and a 6.6% increase in rainy conditions compared to RayFormer. Similarly, it outperforms RayFormer by 8.4% during the day and by 4.3% at night. RaCFormer reaches 94% and 89% of the mAP achieved by LiDAR-based CenterPoint under rain and darkness, respectively, partially compensating for the limitations of image-based perception.

Robustness: To evaluate the robustness gains from radar-camera fusion in our system, we test its performance under sensor failure scenarios, as detailed in Tab. 9. We systematically exclude either image or radar inputs, recording the AP for detecting cars. Our method shows a 2.7% improvement over the CRN under full sensor data conditions and maintains better performance when camera or radar data is missing. Specifically, without any camera data, RaCFormer still achieves a car AP of 27.2%, outperforming CRN by 14.4%. In the absence of radar data, RaCFormer's car AP is 52.8%, which is a 9.0% enhancement over CRN.

Methods	Input	Sunny	Rainy	Day	Night
CenterPoint [52]	L	62.9	59.2	62.8	35.4
RayFormer [6]	C	44.7	49.1	45.7	27.3
RaCFormer	C+R	**53.2**	**55.7**	**54.1**	**31.6**

Table 8. Analysis of the performance under various weather and lighting scenarios using the mAP metric.

Methods	Input	Drop	# of view drops			
			0	1	3	All
CenterPoint [52]	R	R	30.6	25.3	14.9	0
RayFormer [6]	C	C	52.0	44.4	24.0	0
CRN [18]	C+R	C	68.8	62.4	48.9	12.8
		R		64.3	57.0	43.8
RaCFormer	C+R	C	**71.5**	**65.6**	**53.7**	**27.2**
		R		**69.3**	**62.5**	**52.8**

Table 9. Analysis of robustness using Car class AP. "All" denotes that the single modality is entirely off.

Inference Time: To enable real-time detection, we develop a lightweight version of our model: it uses 4 historical frames, 450 queries, and a BEV grid of 64×64 with a resolution of 1.6 meters per voxel. These adjustments reduce the computational load compared to our standard model. Despite these reductions, our model still delivers state-of-the-art mAP and NDS of 51.0% and 58.8%, respectively, outperforming HyDRa [48] by 1.6% in mAP and 2.8% in NDS. Operating on a single RTX 3090 GPU, it achieves a frame rate of 12 FPS, satisfying real-time requirements.

5. Conclusion

In this paper, we present RaCFormer, a novel query-based 3D object detection method that fuses radar and camera data by cross-perspective feature sampling. In particular, by enhancing depth estimation with a radar-guided preprocessing, designing a circular query initialization with the linearly increasing strategy, and leveraging the radar's Doppler effect for BEV temporal encoding, RaCFormer enables both modalities to capitalize their respective strengths and complement each other effectively. Our method achieves superior results on the nuScenes and VoD datasets, marking a significant leap forward in high-performance and robust 3D perception for autonomous driving.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62332016) and the Key Research Program of Frontier Sciences, CAS (No. ZDBS-LY-JSC001).

References

[1] Eduardo Arnold, Omar Y. Al-Jarrah, Mehrdad Dianati, Saber Fallah, David Oxtoby, and Alex Mouzakitis. A survey on 3d object detection methods for autonomous driving applications. *IEEE Trans. Intell. Transp. Syst.*, 20(10):3782–3795, 2019. 1

[2] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[3] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 5

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[5] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 21674–21683, 2023. 6

[6] Xiaomeng Chu, Jiajun Deng, Guoliang You, Yifan Duan, Yao Li, and Yanyong Zhang. Rayformer: Improving query-based multi-camera 3d object detection via ray-centric strategies. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pages 4620–4629, 2024. 4, 5, 6, 7, 8

[7] Jiajun Deng, Sha Zhang, Feras Dayoub, Wanli Ouyang, Yanyong Zhang, and Ian Reid. Poifusion: Multi-modal 3d object detection via fusion at points of interest. *CoRR*, abs/2403.09212, 2024. 2

[8] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 4

[9] Ziteng Gao, Limin Wang, Bing Han, and Sheng Guo. Adamixer: A fast-converging query-based object detector. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5354–5363, 2022. 3, 5

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 3354–3361, 2012. 5

[11] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *Int. J. Robotics Res.*, 32(11):1231–1237, 2013. 5

[12] Chunrui Han, Jinrong Yang, Jianjian Sun, Zheng Ge, Runpei Dong, Hongyu Zhou, Weixin Mao, Yuang Peng, and Xiangyu Zhang. Exploring recurrent long-term temporal fusion for multi-view 3d perception. *IEEE Robotics Autom. Lett.*, 9 (7):6544–6551, 2024. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[14] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *CoRR*, 2022. 2

[15] Junjie Huang, Guan Huang, Zheng Zhu, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *CoRR*, 2021. 2, 4

[16] Yanqin Jiang, Li Zhang, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, and Yu-Gang Jiang. Polarformer: Multi-camera 3d object detection with polar transformers. *CoRR*, 2022. 6

[17] Youngseok Kim, Sanmin Kim, Jun Won Choi, and Dongsuk Kum. Craft: Camera-radar 3d object detection with spatio-contextual fusion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1160–1168, 2023. 2

[18] Youngseok Kim, Juyeb Shin, Sanmin Kim, In-Jae Lee, Jun Won Choi, and Dongsuk Kum. CRN: camera radar net for accurate, robust, efficient 3d perception. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 17569–17580, 2023. 1, 2, 6, 8

[19] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 12697–12705, 2019. 4, 7

[20] Youngwan Lee, Joong-Won Hwang, Sangrok Lee, Yuseok Bae, and Jongyoul Park. An energy and gpu-computation efficient backbone network for real-time object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, 2019. 5

[21] Youngwan Lee, Joong-Won Hwang, Sangrok Lee, Yuseok Bae, and Jongyoul Park. An energy and gpu-computation efficient backbone network for real-time object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, 2019. 6

[22] Kai Lei, Zhan Chen, Shuman Jia, and Xiaoteng Zhang. Hvdetfusion: A simple and robust camera-radar fusion framework. *CoRR*, abs/2307.11323, 2023. 2, 6

[23] Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun, and Jiaya Jia. Unifying voxel-based representation with transformer for 3d object detection. *CoRR*, abs/2206.00630, 2022. 6

[24] Yao Li, Jiajun Deng, Yu Zhang, Jianmin Ji, Houqiang Li, and Yanyong Zhang. Ezfusion: A close look at the integration of lidar, millimeter-wave radar, and camera for accurate 3d object detection and tracking. *IEEE Robotics Autom. Lett.*, 7 (4):11182–11189, 2022. 2

[25] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1477–1485, 2023. 2, 4

[26] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[27] Zhiqi Li, Zhiding Yu, Wenhai Wang, Anima Anandkumar, Tong Lu, and José M. Álvarez. FB-BEV: BEV representation from forward-backward view transformations. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 6896–6905, 2023. 2

[28] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. *CoRR*, abs/2211.10581, 2022. 2, 5

[29] Zhiwei Lin, Zhe Liu, Zhongyu Xia, Xinhao Wang, Yongtao Wang, Shengxiang Qi, Yang Dong, Nan Dong, Le Zhang, and Ce Zhu. Rcbevdet: Radar-camera fusion in bird's eye view for 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 14928–14937, 2024. 1, 3, 6, 7

[30] Feng Liu, Tengteng Huang, Qianjing Zhang, Haotian Yao, Chi Zhang, Fang Wan, Qixiang Ye, and Yanzhao Zhou. Ray denoising: Depth-aware hard negative sampling for multi-view 3d object detection. *CoRR*, abs/2402.03634, 2024. 2

[31] Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, and Limin Wang. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 18534–18544, 2023. 2, 3, 5

[32] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. PETR: position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[33] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. Petrv2: A unified framework for 3d perception from multi-camera images. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3239–3249, 2023. 5

[34] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. *CoRR*, 2022. 1, 7

[35] Yunfei Long, Abhinav Kumar, Daniel Morris, Xiaoming Liu, Marcos Castro, and Punarjay Chakravarty. Radiant: Radar-image association network for 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1808–1816, 2023. 2

[36] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017. 5

[37] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. 5

[38] Xinzhu Ma, Wanli Ouyang, Andrea Simonelli, and Elisa Ricci. 3d object detection from images for autonomous driving: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46 (5):3537–3556, 2024. 1

[39] Jiageng Mao, Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. 3d object detection for autonomous driving: A comprehensive survey. *Int. J. Comput. Vis.*, 131(8):1909–1963, 2023. 1

[40] Ramin Nabati and Hairong Qi. Centerfusion: Center-based radar and camera fusion for 3d object detection. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 1526–1535, 2021. 1

[41] Andras Palffy, Ewoud A. I. Pool, Srimannarayana Baratam, Julian F. P. Kooij, and Dariu M. Gavrila. Multi-class road user detection with 3+1d radar in the view-of-delft dataset. *IEEE Robotics Autom. Lett.*, 7(2):4961–4968, 2022. 2, 5

[42] Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. Is pseudo-lidar needed for monocular 3d object detection? In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 5, 6

[43] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 4

[44] Rui Qian, Xin Lai, and Xirong Li. 3d object detection for autonomous driving: A survey. *Pattern Recognit.*, 130:108796, 2022. 1

[45] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3598–3608, 2023. 2, 5, 6

[46] Yue Wang, Vitor Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. DETR3D: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning (CoRL)*, 2021. 2

[47] Zitian Wang, Zehao Huang, Jiahui Fu, Naiyan Wang, and Si Liu. Object as query: Lifting any 2d object detector to 3d detection. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3768–3777, 2023. 2

[48] Philipp Wolters, Johannes Gilg, Torben Teepe, Fabian Herzog, Anouar Laouichi, Martin Hofmann, and Gerhard Rigoll. Unleashing hydra: Hybrid fusion, depth consistency and radar for unified 3d perception. *CoRR*, abs/2403.07746, 2024. 3, 6, 8

[49] Junjie Yan, Yingfei Liu, Jianjian Sun, Fan Jia, Shuailin Li, Tiancai Wang, and Xiangyu Zhang. Cross modal transformer: Towards fast and robust 3d object detection. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 18222–18232, 2023. 2

[50] Bin Yang, Runsheng Guo, Ming Liang, Sergio Casas, and Raquel Urtasun. Radarnet: Exploiting radar for robust perception of dynamic objects. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVIII*, pages 496–512, 2020. 1

[51] Junbo Yin, Jianbing Shen, Runnan Chen, Wei Li, Ruigang Yang, Pascal Frossard, and Wenguan Wang. Is-fusion: Instance-scene collaborative fusion for multimodal 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 14905–14915, 2024. 2

[52] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3d object detection and tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 6, 7, 8

[53] Lianqing Zheng, Sen Li, Bin Tan, Long Yang, Sihan Chen, Libo Huang, Jie Bai, Xichan Zhu, and Zhixiong Ma. Rcfusion: Fusing 4d radar and camera with bird's-eye view features for 3d object detection. *IEEE Transactions on Instrumentation and Measurement*, 72, 2023. 1, 2, 7

[54] Taohua Zhou, Junjie Chen, Yining Shi, Kun Jiang, Mengmeng Yang, and Diange Yang. Bridging the view disparity between radar and camera features for multi-modal fusion 3d object detection. *IEEE Trans. Intell. Veh.*, 8(2):1523–1535, 2023. 1

[55] Zhuofan Zong, Dongzhi Jiang, Guanglu Song, Zeyue Xue, Jingyong Su, Hongsheng Li, and Yu Liu. Temporal enhanced training of multi-view 3d object detector via historical object prediction. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3758–3767, 2023. 2