RCQFUSION: SPARSE RADAR-CAMERA QUERY FUSION FOR ROBUST 3D DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Autonomous driving requires reliable perception under diverse and challenging environments. While multi-sensor fusion has become the standard paradigm, existing radar–camera fusion methods largely inherit LiDAR-centric dense BEV-fusion architectures, struggling with radar sparsity, information loss, cross-modal feature misalignment, and high computational cost from dense BEV representations. To address these issues, we present **RCQFusion**, a **sparse radar-camera query fusion framework for 3D detection**. Our method introduces a **density-aware transformer encoder**, a **sparse radar-camera fusion module**, a **multi-modal query generation mechanism**, and a **hybrid matching module** to offer a scalable and efficient solution for long-range, robust 3D perception in real-world autonomous-driving scenarios. Extensive experiments on nuScenes demonstrate that **RCQFusion** achieves state-of-the-art radar-camera 3D detection performance, with clear improvements for long-range, dynamic-object, and adverse-condition perception while preserving practical real-time efficiency. Additional evaluation on MAN TruckScenes further shows the generalization ability of RCQFusion to 4D radar sensing and long-range detection scenarios.

1 INTRODUCTION

Autonomous driving systems require robust perception capabilities to ensure safe operation across diverse and challenging environments. To overcome the inherent limitations of individual sensors, these systems heavily rely on multi-modal sensor fusion. Among the primary sensors (LiDAR, cameras, and radar), LiDAR delivers dense 3D point clouds but suffers from high costs and degraded performance under adverse weather conditions. In contrast, millimeter-wave radar provides cost-effective, velocity-aware detection with superior weather resilience and extended sensing range. Cameras offer rich semantic and texture information, yet they lack depth cues, perform poorly in depth estimation, and are vulnerable to extreme weather (such as rain and fog) and challenging illumination conditions. The complementary strengths and lower cost of radar and cameras make radar–camera fusion a promising alternative to LiDAR–camera systems, especially for detecting dynamic objects at long distances. However, most existing radar–camera fusion methods are adapted from LiDAR-centric frameworks, which inadequately address the inherent sparsity and noise of radar point clouds and fail to fully exploit radar's advantages for long-range, fast-moving, and occluded object detection, thus limiting their effectiveness in real-world scenarios.

BEVFusion Liu et al. (2023) aligns camera and LiDAR features in a shared Bird's-Eye-View (BEV) representation. Radar–camera frameworks including CRN Kim et al. (2023b), RCBEVDet Lin et al. (2024), HyDRa Wolters et al. (2025b), and CRT-Fusion Kim et al. (2024) also use BEV feature representations for 3D perception. Although RaCFormer Chu et al. (2025) introduces a Transformer-based paradigm with sparse queries and set-to-set prediction, it still builds query features from dense BEV representations. Consequently, the computational complexity of such BEV-based models grows approximately quadratically with BEV resolution and perception range, leading to high inference cost in large-scale or long-range scenarios. Moreover, these BEV fusion pipelines originate from dense LiDAR–camera frameworks whose voxel- or pillar-based feature extraction designs are ill-suited to the inherent sparsity and noise of millimeter-wave radar measurements. As a result, a substantial portion of voxels remain empty, GPU memory utilization is inefficient, and radar points beyond the predefined BEV range are typically discarded, thus preventing full exploitation of radar's unique

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

long-range sensing capability (as the detection range of modern 4D radar (>350 m) far exceeds the typical camera and LiDAR range (∼150 m)).

Meanwhile, models adopting the **sparse query** and **set-to-set prediction** paradigm have become widely used in pure vision-based 3D detection, including DETR3D Wang et al. (2022), PETR Liu et al. (2022), StreamPETR Wang et al. (2023), Far3D Jiang et al. (2024), and Sparse4D v3 Lin et al. (2023). These approaches employ a limited number of object queries to aggregate image features using 3D geometric information, substantially reducing computational overhead and enabling straightforward extension to multi-frame temporal fusion. However, the set-to-set matching process is highly non-linear and combinatorially complex, often leading to optimization instability and slow convergence during training. This issue becomes more pronounced in temporal sparse-query schemes, where non-homogeneous query sources such as historical queries and randomly initialized learnable queries may compete for the same ground-truth instance under strict one-to-one assignment. Although methods such as DN-DETR Li et al. (2022) attempt to mitigate these issues, training remains challenging, hindering the scalability of this paradigm to more complex multi-modal perception tasks.

To address the aforementioned challenges, we propose **RCQFusion**, a **sparse radar-camera query fusion framework**. Our framework introduces a Transformer-based **density-aware trans-encoder network** for sparse radar point cloud feature extraction and a **sparse radar-camera feature fusion module** to effectively handle radar sparsity and height uncertainty. By organizing cross-modal interaction around sparse radar points and object-centric queries, RCQFusion focuses computation on informative regions without introducing an additional dense fusion head. To overcome the limitations of conventional query initialization, which is often biased by data distribution, we exploit radar's inherent advantages in detecting dynamic, metallic, and long-range targets, along with the high-recall property of 2D vision detection, to construct a **multi-modal query generation mechanism**. Furthermore, to alleviate the optimization instability of set-to-set prediction, we propose a theoretically grounded and empirically validated **hybrid match module**, which enhances gradient stability and accelerates convergence.

Extensive experiments on the nuScenes Caesar et al. (2020) dataset demonstrate that **RCQFusion** surpasses state-of-the-art radar–camera fusion methods, with clear gains on long-range, moving, and adverse-condition objects and practical real-time efficiency under only modest additional latency over the camera-only baseline. Our **main contributions** are summarized as follows:

1. We propose a **sparse radar-camera query fusion framework** that efficiently encodes sparse radar point clouds and fuses them with visual features through sparse point- and query-level interactions, maintaining real-time-capable efficiency with limited additional latency;
2. We design a **multi-modal query generation mechanism** that jointly leverages radar priors and visual 2D detections for adaptive and efficient spatial sampling;
3. We introduce a **hybrid match module** that mitigates the convergence difficulty of set-to-set prediction through both theoretical analysis and experimental validation;
4. On the benchmark **nuScenes** dataset, **RCQFusion** outperforms multiple recent SOTA fusion methods and shows stronger robustness for long-range targets, moving objects, and adverse weather or lighting conditions while preserving practical real-time efficiency.

2 The Proposed RCQFusion Method

2.1 Overview

As illustrated in Fig. 1, the proposed **RCQFusion** is a sparse query-based framework consisting of five stages: (1) multi-view image and radar feature extraction, (2) radar point sampling and fusion, (3) multi-modal query generation, (4) query-based decoding, (5) hybrid multi-source matching. We extract multi-scale visual features via standard backbones (e.g., ResNet He et al. (2016), VoVNet Lee et al. (2019), and ViT Dosovitskiy et al. (2021)) and encode radar point clouds using a transformer-based self-attention encoder whose cost scales with the number of radar points rather than the dense BEV volume. After radar–image feature fusion through point projection and sampling enhancement, we construct heterogeneous queries from radar, radar–camera fusion, and vision modalities, together with top-K historical queries. These queries are jointly optimized under a **multi-source hybrid**

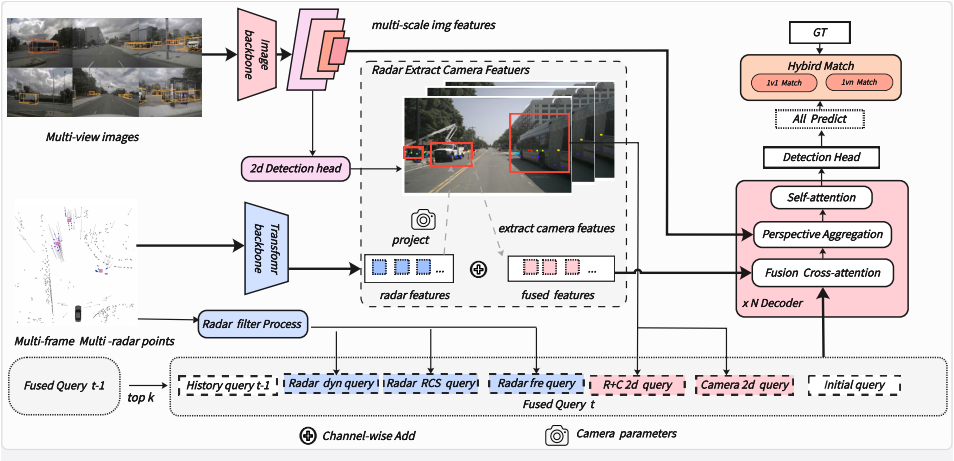


Figure 1: Overall Architecture of RCQFusion. Surround-view images and radar point clouds are processed to extract 2D and point cloud features, which are fused by projecting radar points onto the image plane. 2D boxes and radar points form multi-modal queries, combined with global and historical queries for decoding. In each decoder layer, queries interact with fused features and multi-scale image features.

matching scheme integrating one-to-one and dynamic one-to-many supervision to enhance cross-modal consistency, stabilize gradients, and accelerate convergence.

2.2 Radar Transformer Backbone with Density-Aware Attention

Unlike LiDAR, radar point clouds are extremely sparse, irregular, and often noisy. Applying full self-attention over all points causes redundant computation and feature confusion. To address this, we propose a **Density-Aware Transformer Backbone** that adaptively adjusts its attention scope based on local point density, enabling efficient and noise-robust radar encoding.

Given radar points $\mathcal{P} = \{p_i\}_{i=1}^{N}$, we first embed raw signals with a shared multilayer perceptron (MLP), which maps the radar input dimension to the feature dimension d and produces $F^{(0)}$. For each Transformer layer l, pairwise distances $D_{ij} = \|pos_i - pos_j\|_2$ are computed to construct a *density-aware mask*:

$$C_{ij}^{(l)} = [D_{ij} \leq \tau_l] \wedge (|\{k : D_{jk} \leq \tau_l\}| \geq \kappa_l), \tag{1}$$

where τ_l controls the receptive radius and κ_l determines the local density threshold. This mask limits attention to density-reachable neighbors, similar to DBSCAN clustering Ester et al. (1996), so each point interacts only within its adaptive local region rather than the entire point cloud.

The masked self-attention then becomes:

$$\text{Attn}^{(l)} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}} \odot M^{(l)}\right) V. \tag{2}$$

Thresholds follow a progressive schedule, $\tau_l = \tau_{\max} e^{-\lambda l}$ and $\kappa_l = \kappa_{\text{base}} + \Delta\kappa l$, forming hierarchical receptive fields from local geometry to global context. Finally, a residual connection fuses multi-level features as $\mathcal{R} = \mathcal{F}^{(k)} \oplus \mathcal{F}^{(0)}$, preserving both fine-grained structures and holistic scene cues. Pseudocode is provided in the **Supplementary Material** (Sec. G).

2.3 Sample Enhancement and Sparse Radar–Camera Feature Fusion

Mainstream multi-modal fusion frameworks, such as BEVFusion Liu et al. (2023), typically employ depth-prediction networks and Lift-Splat-Shoot (LSS) Philion & Fidler (2020) to lift 2D image features into 3D space for BEV fusion. However, the reliance on predicted depth often leads to projection ambiguity and geometric misalignment due to estimation errors and horizon assumption violations. Although CRN Kim et al. (2023b) utilizes radar depth to guide feature mapping, the

extreme sparsity of radar still causes uneven and inconsistent cross-modal features, as most camera pixels remain dependent on unreliable depth estimation.

To address these limitations, we adopt an inverse projection strategy **3D-to-2D feature mapping** that avoids explicit depth regression. Inspired by DETR3D Wang et al. (2022), we project radar points directly onto multi-scale camera feature maps. To compensate for radar sparsity and geometric uncertainty, we further propose a set of learnable **sampling enhancement strategies** before projection. For each radar point $p_i = (x, y, z)$, we augment its local neighborhood through three complementary operations:

- **Height Sampling** (P_h): Samples are generated along the vertical axis within typical object height ranges (0.8–5 m) to compensate for missing elevation information.
- **Scale Sampling** (P_k): An MLP predicts multiple horizontal offsets to capture local scale variation and contextual structure.
- **Density Sampling** (P_d): Within each radar cluster, the farthest d points are selected to represent spatial density and distribution.

These samples are combined into a unified enhanced set: $P_e = \{P_h, P_k, P_d\}$.

The enhanced 3D samples are projected onto 2D feature maps using the camera intrinsic and extrinsic parameters. The corresponding image features F_{2d}^i are aggregated through learned weights:

$$F_{\text{camera}} = \sum_{i=1}^{e \times N} \text{MLP}(R_i) \cdot F_{2d}^i, \tag{3}$$

We then fuse the aggregated camera feature with the radar feature through residual addition, $F_{\text{fusion}} = R_{\text{radar}} + F_{\text{camera}}$.

Overall, this sampling-enhanced 3D-to-2D projection establishes an efficient **radar–camera feature alignment pipeline** that preserves radar geometric priors, enriches feature density, and effectively mitigates cross-modal misalignment caused by radar sparsity.

2.4 Multi-Modal Query Generation

Extending the perception range in autonomous driving (e.g., from 50 m to 100 m) typically requires either enlarging the BEV coverage or increasing grid resolution, both of which significantly raise computational cost and may dilute near-range feature quality. Similarly, query-based methods often resort to a larger query set to cover distant regions, which increases attention complexity and weakens focus on dense close-range targets. Unlike prior approaches relying solely on depth-derived or learned queries Jiang et al. (2024), our design introduces physically grounded, radar-conditioned query priors that enhance geometric diversity and long-range robustness. To this end, we propose a **multi-modal query generation framework** that leverages radar's precise geometry and motion sensitivity to improve detection of distant and occluded objects.

Our method integrates three complementary query types: (1) **Radar queries**, selected from informative radar points using temporal consistency, motion cues, and high RCS; (2) **Radar–Camera association queries**, combining radar geometry with camera semantics via depth-distribution modeling; and (3) **Camera-only queries**, generated from 2D detections without radar support. This physically grounded initialization improves long-range and occluded-object detection while maintaining low query complexity.

Radar Query Encoding. Kalman-filter-based methods have been used for radar target tracking on curved roads Song et al. (2020). However, for simplicity, we select radar-query candidates using radar priors encoded with positional and semantic embeddings, while all radar returns are still preserved for radar feature encoding. These priors include radar dynamic queries generated from high-velocity radar points for moving objects, radar RCS queries generated from points with RCS above a predefined threshold for potential metallic or highly reflective targets, and radar frequent queries generated from repeatedly observed points within local regions for temporally stable object candidates.

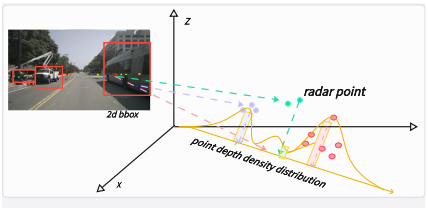


Figure 2: Radar point projection and depth-distribution modeling within 2D bounding boxes. Multiple radar points (colored by depth) are mapped to a single 2D box, and representative points are selected via depth-distribution peak sampling to ensure physically consistent multi-modal alignment.

The query is formed in one step, $Q_{\text{pos}} = \text{PosEmbed}(p_i)$, $Q_{\text{sem}} = F_{\text{fusion}}(p_i)$, and $Q = Q_{\text{pos}} + Q_{\text{sem}}$, where PosEmbed denotes sinusoidal positional encoding followed by an MLP similar to Wang et al. (2023); Jiang et al. (2024). Radar-conditioned priors inherently provide stable spatial initialization under sparse measurements, facilitating more consistent long-range generalization.

Radar–Camera Association. We project radar points into multi-view image features and associate them with 2D detections obtained from an anchor-free detector Ge et al. (2021). Since radar can perceive occluded and metallic targets, multiple radar points may correspond to a single 2D bounding box. To model such multi-depth associations and suppress unreliable radar reflections, we introduce a **depth-distribution peak sampling** strategy that explicitly models the depth histogram within each 2D detection region (Fig. 2).

Formally, for each 2D detection box b_j in image view v, let $\mathcal{R}_j = \{r_i \mid r_i \in b_j\}$ denote the set of radar points projected into this box, each with depth value d_i. We construct a normalized depth distribution:

$$P_j(d) = \frac{\exp(-\frac{(d-d_i)^2}{2\sigma^2})}{\sum_{r_i \in \mathcal{R}_j} \exp(-\frac{(d-d_i)^2}{2\sigma^2})}, \tag{4}$$

where σ controls the smoothing bandwidth. We then identify K representative depth peaks, $\mathcal{D}_j^* = \text{TopKPeaks}(P_j(d), K)$, and back-project each peak to its 3D coordinate:

$$\mathbf{c}_{rc} = K^{-1} I^{-1} [c_w d^*, c_h d^*, d^*, 1]^T, \tag{5}$$

where K and I are the camera extrinsic and intrinsic matrices, and (c_w, c_h) denotes the 2D box center. The associated query embeddings are $Q_{\text{pos}} = \text{PosEmbed}(\mathbf{c}_{rc})$, $Q_{\text{sem}} = \text{MLP}([context_{2d}, score_{2d}]) + R_{\text{radar}}$, and $Q = Q_{\text{pos}} + Q_{\text{sem}}$.

This depth-distribution modeling ensures that radar-conditioned queries are grounded on statistically stable, physically interpretable depth clusters, effectively filtering out multipath noise and improving geometric precision in occluded or distant regions.

Camera 2D Queries. For 2D detections without corresponding radar points, we back-project predicted depths to generate camera-only queries. These queries maintain high recall in visually dominant regions and serve as complementary cues where radar coverage is weak.

This hybrid multi-source query design bridges **physics-driven priors** and **data-driven learning**, balancing geometric reliability and semantic richness. It effectively reduces distribution bias and enhances generalization to challenging scenarios, an essential property for robust real-world autonomous driving perception.

2.5 Multi-Source Query Hybrid Matching

The multi-modal query generation stage (Sec. 2.4) produces heterogeneous queries from radar points, radar–camera associations, and camera-only 2D detections, together with historical memory queries and learnable queries following prior sparse-query detectors Wang et al. (2023); Jiang et al. (2024); Lin et al. (2023). While these query sources improve spatial coverage and reduce the bias of purely learned queries, multiple queries can represent the same object and compete for the single positive

assignment in DETR-style one-to-one matching Carion et al. (2020); Liu et al. (2022). In particular, clustered radar points, radar-associated visual proposals, historical queries, and learnable queries may all describe the same GT instance. Under strict 1v1 matching, these semantically equivalent but non-homogeneous queries compete for a single positive assignment, resulting in **false negatives, unstable gradients, and slow convergence**. This motivates auxiliary positive supervision for competing queries. Inspired by Li et al. (2022); Jia et al. (2023); Chen et al. (2023a), we generalize the hybrid matching principle to multi-modal query fusion and propose the **hybrid match** mechanism with the following characteristics:

- **Auxiliary 1vN branch:** A shared-decoder one-to-many (1vN) auxiliary branch is introduced alongside the main 1v1 branch.
- **Cross-modal supervision:** Queries from radar, vision, and fusion modalities that refer to the same object are treated as positives jointly.
- **Adaptive multiplicity:** Matching multiplicity is controlled by a fixed K or a dynamic-K strategy based on IoU/distance thresholds.

Unlike naive one-to-many strategies (e.g., simple global GT$\times K$ replication followed by Hungarian matching), which often produce noisy gradients and severe loss oscillations in the early training phase, our design avoids excessive supervision. We instead adopt a **dynamic-K strategy**: for each GT, we perform nearest-neighbor search over predicted queries and treat those within a geometric distance or IoU threshold as positive samples. The resulting K_{dyn} adaptively determines the number of matched queries per GT. This ensures that only high-quality candidates (geometrically close or IoU-aligned) participate in the 1vN supervision, yielding smoother gradients and more stable convergence.

For stability, the **1v1 main branch** uses 1v1 Hungarian matching to do full classification and regression supervision, while the **1vN auxiliary branch** emphasizes classification (improving positive recall) or applies reduced regression weights to alleviate positional conflicts. During **inference**, only the main 1v1 branch is used for prediction, preserving the end-to-end, NMS-free property and avoiding redundant outputs.

For loss computation, we follow standard 3D detection practice Wang et al. (2023); Liu et al. (2022), employing focal loss for classification and L1 loss for bounding-box regression. The overall objective is defined as:

$$\mathcal{L}_{\text{total}} = (1 - \lambda_{\text{aux}})\mathcal{L}_{\text{one2one}} + \lambda_{\text{aux}}\mathcal{L}_{\text{one2many}}, \tag{6}$$

where λ_{aux} balances the main and auxiliary losses. This hybrid design balances convergence stability and query diversity, achieving faster training and stronger radar-camera consistency while maintaining the simplicity of DETR-based end-to-end detection.

3 Experimental Results

3.1 Datasets and Metrics

nuScenes Caesar et al. (2020) is one of the most widely used multimodal perception benchmarks for autonomous driving, providing synchronized multi-view camera images and multi-angle radar measurements. It is also the **only large-scale public dataset** that offers both multi-camera and radar modalities, as other major benchmarks such as Waymo Sun et al. (2020) and Argoverse 2 Wilson et al. (2021) do not include millimeter-wave radar.

The dataset contains 1000 driving scenes (700 for training, 150 for validation, 150 for testing), each lasting 20 seconds with 2 Hz annotated keyframes. Following the official protocol, we report mean Average Precision (mAP) and nuScenes Detection Score (NDS), as well as the standard error metrics: Average Translation Error (ATE), Average Scale Error (ASE), Average Orientation Error (AOE), Average Velocity Error (AVE), and Average Attribute Error (AAE). Due to the restriction that the official test server is required for test-set evaluation and is not always accessible, we follow common practice by conducting comparisons on the validation set and reporting our best-performing configuration on the test set. Due to the sparsity of radar point clouds in nuScenes, we follow Kim et al. (2023b); Lin et al. (2024); Kim et al. (2024) to apply a multi-sweep fusion strategy for fair comparison.

Table 1: Performance comparison on the nuScenes val set. "C": camera, "R": radar, "L": LiDAR.

Method	Modality	Resolution	Backbone	mAP ↑	NDS ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓
StreamPETR Wang et al. (2023)	C	704×256	Res50	0.449	0.546	0.618	0.274	0.420	0.267	0.208
CRAFT† Kim et al. (2023a)	R+C	800×448	DLA34	0.411	0.517	0.494	0.276	0.454	0.486	**0.176**
CRN Kim et al. (2023b)	R+C	704×256	Res50	0.490	0.560	0.487	0.277	0.542	0.344	0.197
RCBEVDet Lin et al. (2024)	R+C	704×256	Res50	0.453	0.568	0.486	0.285	**0.404**	0.220	0.192
HyDRa Wolters et al. (2025b)	R+C	704×256	Res50	0.494	0.585	**0.463**	0.268	0.478	0.227	0.182
CRT-Fusion Kim et al. (2024)	R+C	704×256	Res50	0.500	0.572	0.499	0.277	0.531	0.261	0.192
RaCFormer-L Chu et al. (2025)	R+C	704×256	Res50	0.510	0.588	-	-	-	-	-
RaCFormer Chu et al. (2025)	R+C	704×256	Res50	0.541	0.613	0.478	0.261	0.446	**0.208**	0.180
RCQFusion-L (Ours)	R+C	704×256	Res50	0.525	0.600	0.501	0.263	0.442	0.241	0.181
RCQFusion (Ours)	R+C	704×256	Res50	**0.550**	**0.613**	0.510	**0.258**	0.421	0.248	0.196
Far3D Jiang et al. (2024)	C	800×320	VoV-99	0.507	0.587	0.567	0.257	0.400	0.255	0.191
CMT-R+C Yan et al. (2023)	R+C	800×320	VoV-99	0.409	0.463	0.772	0.270	0.488	0.746	0.205
RCQFusion (Ours)	R+C	800×320	VoV-99	**0.556**	**0.621**	**0.500**	**0.257**	**0.387**	**0.238**	**0.191**
Far3D Jiang et al. (2024)	C	1408×512	Res101	0.510	0.594	0.551	0.258	0.372	0.238	0.195
CRN Kim et al. (2023b)	R+C	1408×512	Res101	0.525	0.592	0.460	0.273	0.443	0.352	0.180
CRT-Fusion Kim et al. (2024)	R+C	1408×512	Res101	0.554	0.621	0.452	0.264	0.433	0.237	0.193
RaCFormer Chu et al. (2025)	R+C	1408×512	Res101	0.573	0.630	0.476	0.261	0.428	**0.213**	0.183
RCQFusion (Ours)	R+C	1408×512	Res101	**0.574**	**0.637**	**0.482**	**0.258**	**0.360**	0.221	**0.178**
Far3D Jiang et al. (2024)	C	800×320	ViT	0.537	0.617	0.514	0.252	0.300	0.247	**0.198**
RCQFusion (Ours)	R+C	800×320	ViT	**0.587**	**0.650**	**0.467**	**0.251**	**0.287**	**0.232**	0.201

Note: All results are reported on the nuScenes validation set. Best results under each backbone/resolution setting are highlighted in bold. †: trained with CBGS. ‡: uses test-time augmentation. Our RCQFusion-L uses fewer queries.

3.2 Main Results

Main Results (ResNet-50 backbone). Under the widely adopted ResNet-50 setting in Table 1, **RCQFusion** achieves 0.550 mAP and 0.613 NDS at 704×256, outperforming CRN Kim et al. (2023b) by +0.060 mAP and +0.053 NDS while using 12.9M rather than 61.4M parameters. Under a matched ResNet-50 inference setting (704×256, FP16 autocast), it also improves mAP over RaCFormer Chu et al. (2025) (0.550 vs. 0.541, with the same 0.613 NDS) while reducing latency from 216.0 ms to 56.1 ms and peak memory from 4.55 GB to 2.76 GB. Further protocol details and the module-wise breakdown are provided in the **Supplementary Material** (Sec. D, Tab. 10). It also surpasses recent dense-BEV fusion baselines Kim et al. (2024); Lin et al. (2024); Wolters et al. (2025b), demonstrating a favorable accuracy–efficiency trade-off.

Scaling Behavior. We evaluate our framework with VoVNet-99 at 800×320, where **RCQFusion** achieves improvements of +0.05 mAP / +0.034 NDS compared with Far3D. With the ViT backbone at 800×320 and the ResNet-101 backbone at 1408×512, our method achieves similar gains. Across all evaluated backbones and resolutions, RCQFusion consistently yields monotonic improvements (+0.05 mAP/+0.03 NDS) over its camera-only baseline Far3D Jiang et al. (2024), indicating robust scalability without performance saturation or collapse.

SOTA in test set. Tab. 2 reports the full nuScenes test-set comparison. With VoV-99 at 1600×640, **RCQFusion** achieves 0.612 mAP and 0.663 NDS, surpassing HyDRa and RaCFormer under the same setting. With the EVA-02 ViT backbone Fang et al. (2024) at 1600×800, it attains **0.640 mAP** and **0.686 NDS** without relying on future-frame inputs or test-time augmentation.

Generalization to 4D radar. As shown in the MAN TruckScenes results in the **Supplementary Material** (Tab. 11 and Tab. 12), RCQFusion maintains strong cross-dataset generalization on 4D radar long-range detection. Notably, using only radar-camera inputs, RCQFusion even surpasses the LiDAR-only CenterPoint-V baseline on this 4D radar dataset (23.9%/36.0% vs. 22.6%/35.3% mAP/NDS).

3.3 Ablation Studies

We conduct main ablations on the nuScenes validation set using Far3D Jiang et al. (2024) (VoVNet-99, 800×320) as our camera-only baseline. Unless otherwise stated, all components (backbone, FPN, 2D detector, perspective-aware aggregation) follow the original configuration in Jiang et al. (2024).

Table 2: Performance comparison on the nuScenes test set. "C": camera, "R": radar, "L": LiDAR.

Method	Modality	Resolution	Backbone	mAP ↑	NDS ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓
CenterFusion Nabati & Qi (2021)	R+C	1408×512	DLA34	0.326	0.449	0.631	0.261	0.516	0.614	0.115
MVFusion Wu et al. (2023)	R+C	1408×512	VoV-99	0.453	0.517	0.569	0.246	0.379	0.781	0.128
CRAFT Kim et al. (2023a)	R+C	1408×512	DLA34	0.411	0.523	0.467	0.268	0.456	0.519	0.114
RCBEVDet Lin et al. (2024)	R+C	1408×512	V2-99	0.550	0.639	0.390	**0.234**	0.362	0.259	**0.113**
HyDRa Wolters et al. (2025b)	R+C	1600×640	VoV-99	0.574	0.642	0.398	0.251	0.423	0.249	0.122
CRN‡ Kim et al. (2023b)	R+C	1600×640	ConvNeXt-B	0.575	0.624	0.416	0.264	0.456	0.365	0.130
CRT-Fusion Kim et al. (2024)	R+C	1408×512	ConvNeXt-B	0.583	0.649	**0.365**	0.261	0.405	0.262	0.132
RaCFormer Chu et al. (2025)	R+C	1600×640	VoV-99	0.592	0.659	0.407	0.244	0.345	**0.238**	0.132
RCQFusion (Ours)	R+C	1600×640	VoV-99	**0.612**	**0.663**	0.468	0.252	**0.271**	0.245	0.192
RICCARDO(+6) Long et al. (2025)	R+C	1600×640	VoV-99	0.630	0.695	0.363	0.240	0.311	0.167	0.118
RCQFusion (Ours)	R+C	1600×800	ViT	**0.640**	**0.686**	0.404	0.242	**0.306**	0.256	0.130

Note: All results are from the nuScenes test set. Best-performing settings are bold. ‡ denotes test-time augmentation. A "+t" suffix indicates that t future frames are used.

Table 3: Core module ablations on the nuScenes validation set using VoV-99 at 800×320.

Variant	mAP↑	NDS↑
Camera only	0.507	0.587
+ radar queries	0.517	0.592
+ cross fusion feature	0.523	0.595
+ sample enhancement	0.531	0.601
+ camera–radar 2D queries	0.539	0.604
+ multi-source queries + hybrid matching	0.556	0.621

Module-wise Ablation Analysis. Tab. 3 isolates the roles of the main components. Radar queries add geometric 3D priors (+0.010 mAP, +0.005 NDS); cross-fusion feature interaction refines sparse radar features (+0.005 mAP, +0.003 NDS); sample enhancement expands radar-anchored image sampling to address elevation ambiguity and sparse spatial support. Its progressive gain is small (+0.008 mAP/+0.006 NDS) after complementary radar-fusion stages. Camera–radar 2D queries add visual semantics (+0.009 mAP), while multi-source query generation combines radar, radar–camera association, camera, temporal, and learnable queries.

Additional ablations compare matching strategies and radar encoders. Dynamic-K 1vN reaches 0.556/0.621 versus 0.551/0.612 for fixed 1vN, while density-aware radar attention improves over distance-aware attention in the controlled encoder study. The grouped matching variant adds training-time query groups; the first three variants share the same base queries and decoder. Radar-encoder results are reported in Table 6; full matching results and optimization curves are in the **Supplementary Material** (Tab. 8, Sec. I.2).

Weather-aware Analysis. We follow the scene-wise split protocol and evaluate nuScenes validation scenes under Day (111), Rain (24), and Night (15) conditions. Tab. 4 reports mAP/NDS; CRN Kim et al. (2023b) only provides weather-wise mAP in its original paper.

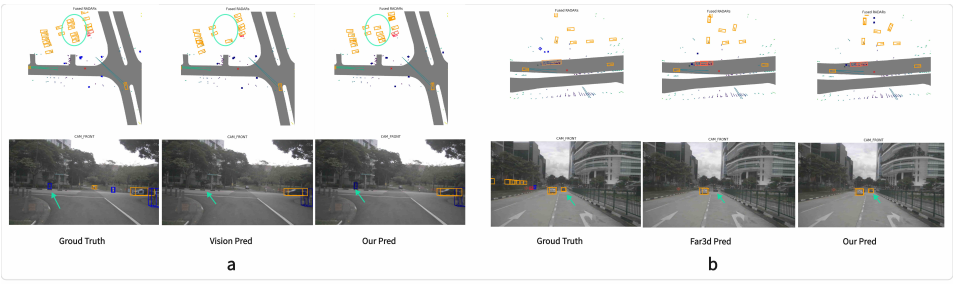


Figure 3: Visualization results on nuScenes. Our method surpasses the camera-only baseline in detecting **blurred moving objects** (green arrow in a), **occluded vehicles** (green circle in a), and **distant metallic targets** (green arrow in b) under challenging conditions.

Table 4: Weather-aware mAP/NDS on the nuScenes validation set using the ResNet-101 backbone

Method	Day	Rain	Night
Far3D Jiang et al. (2024)	0.497 / 0.577	0.572 / 0.646	0.302 / 0.362
CRN Kim et al. (2023b)	0.551 / -	0.570 / -	0.304 / -
RCQFusion	**0.551 / 0.617**	**0.626 / 0.681**	**0.355 / 0.402**

Table 5: Distance-aware and Velocity-aware mAP/NDS on the nuScenes validation set.

Method	Distance			Velocity	
	0–20m	20–40m	40–60m	Moving	Static
Far3D Jiang et al. (2024)	0.573/0.633	0.347/0.470	0.084/0.200	0.250/0.367	0.491/0.590
Ours	**0.603/0.658**	**0.442/0.533**	**0.160/0.252**	**0.337/0.435**	**0.530/0.615**

RCQFusion consistently improves over the camera-only Far3D baseline across all conditions, with larger gains in Rain and Night. Compared with CRN, RCQFusion also improves mAP from 0.570 to 0.626 under Rain and from 0.304 to 0.355 under Night, indicating stronger robustness from radar-guided query fusion. Due to the different split sizes, we use these results as supporting robustness evidence rather than standalone rankings.

Distance-aware Analysis. We further evaluate RCQFusion by grouping nuScenes validation objects according to distance and motion state. For distance-aware evaluation, objects are divided based on their longitudinal distance to the ego vehicle. As shown in Tab. 5, RCQFusion improves mAP/NDS by +0.030/+0.025, +0.095/+0.063, and +0.076/+0.052 at 0–20 m, 20–40 m, and 40–60 m, respectively. Notably, at 40–60 m, RCQFusion nearly doubles the mAP of Far3D (0.160 vs. 0.084), indicating its advantage in long-range perception. This gain is not merely from accurate geometric information of radar features, but also from radar-depth-guided camera-radar 2D queries, which alleviate the inaccurate depth initialization of conventional camera queries.

Velocity-aware Analysis. For velocity-aware evaluation, objects with velocity larger than 0.5 m/s are treated as moving objects, and the remaining ones as static objects. RCQFusion achieves larger gains on moving objects (+0.087/+0.068) than on static objects (+0.039/+0.025). This suggests that RCQFusion is particularly effective for dynamic targets, benefiting from both radar motion cues and the radar prior query design that preferentially selects moving radar responses as object queries.

Qualitative Visualization. We visualize detection results of **RCQFusion** and its camera-only baseline Jiang et al. (2024) on nuScenes. As shown in Fig. 3, our method achieves superior performance in challenging cases, including **blurred moving objects**, **occluded vehicles**, and **distant stationary targets**. Radar cues provide reliable geometric priors when visual evidence is weak, allowing robust localization and consistent identity detection even under severe motion blur or occlusion. More visualizations and analyses are included in the **Supplementary Material** (Sec. H).

4 Conclusion

We present RCQFusion, a sparse radar-camera fusion framework that combines radar geometry and camera semantics through multi-source queries and hybrid matching, without dense BEV fusion. Experiments on nuScenes and MAN TruckScenes demonstrate improved detection, particularly for long-range and moving targets and under adverse conditions, with practical inference efficiency. Performance remains sensitive to radar sparsity, noise, calibration, and synchronization, especially for small, static, or weakly reflective targets. The query and matching designs are tailored to sparse-query decoders and may not transfer directly to dense-head frameworks.

AI USE STATEMENT

AI tools were used to assist with writing parts of the manuscript. We also referred to AI-generated code during experimental implementation.

REPRODUCIBILITY STATEMENT

The model architecture and training objectives are described in the main paper. Dataset splits, evaluation metrics, implementation settings, data preprocessing, additional ablations, and qualitative results are documented in the supplementary material to support reproduction of the reported results. Code and implementation details for the core modules of our method are provided in the code accompanying the supplementary material.

REFERENCES

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1090–1099, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Bai_TransFusion_Robust_LiDAR-Camera_Fusion_for_3D_Object_Detection_With_Transformers_CVPR_2022_paper.html.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11621–11631, 2020. doi: 10.1109/CVPR42600.2020.01164. URL https://www.nuscenes.org/.

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*, pp. 213–229. Springer, 2020. URL https://arxiv.org/abs/2005.12872.

Qiang Chen, Xiaokang Chen, Jian Wang, Shan Zhang, Kun Yao, Haocheng Feng, Junyu Han, Errui Ding, Gang Zeng, and Jingdong Wang. Group DETR: Fast DETR Training with Group-wise One-to-Many Assignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6633–6642, 2023a. URL https://openaccess.thecvf.com/content/ICCV2023/html/Chen_Group_DETR_Fast_DETR_Training_with_Group-Wise_One-to-Many_Assignment_ICCV_2023_paper.html.

Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. VoxelNeXt: Fully Sparse VoxelNet for 3D Object Detection and Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21674–21683, June 2023b. URL https://openaccess.thecvf.com/content/CVPR2023/html/Chen_VoxelNeXt_Fully_Sparse_VoxelNet_for_3D_Object_Detection_and_Tracking_CVPR_2023_paper.html.

Xiaomeng Chu, Jiajun Deng, Guoliang You, Yifan Duan, Houqiang Li, and Yanyong Zhang. RaCFormer: Towards High-Quality 3D Object Detection via Query-based Radar-Camera Fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17081–17091, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Chu_RaCFormer_Towards_High-Quality_3D_Object_Detection_via_Query-based_Radar-Camera_Fusion_CVPR_2025_paper.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Martin Ester, Hans-Peter Kriegel, J"org Sander, and Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second*

International Conference on Knowledge Discovery and Data Mining (KDD), pp. 226–231, 1996. URL https://cdn.aaai.org/KDD/1996/KDD96-037.pdf.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA-02: A Visual Representation for Neon Genesis. *Image and Vision Computing*, 149:105171, 2024. doi: 10.1016/j.imavis.2024.105171. URL https://arxiv.org/abs/2303.11331.

Felix Fent, Fabian Kuttenreich, Florian Ruch, Farija Rizwin, Stefan Juergens, Lorenz Lechermann, Christian Nissler, Andrea Perl, Ulrich Voll, Min Yan, and Markus Lienkamp. MAN TruckScenes: A multimodal dataset for autonomous trucking in diverse conditions. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 62062–62082. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/71ac06f0f8450e7d49063c7bfb3257c2-Paper-Datasets_and_Benchmarks_Track.pdf.

Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: Exceeding YOLO Series in 2021, 2021. URL https://arxiv.org/abs/2107.08430.

Adam W Harley, Zhaoyuan Fang, Jie Li, Rares Ambrus, and Katerina Fragkiadaki. Simple-BEV: What really matters for multi-sensor BEV perception? In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2759–2765. IEEE, 2023. URL https://simple-bev.github.io/.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. DETRs with Hybrid Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19702–19712, 2023. doi: 10.1109/CVPR52729.2023.01887. URL https://openaccess.thecvf.com/content/CVPR2023/html/Jia_DETRs_With_Hybrid_Matching_CVPR_2023_paper.html.

Xiaohui Jiang, Shuailin Li, Yingfei Liu, Shihao Wang, Fan Jia, Tiancai Wang, Lijin Han, and Xiangyu Zhang. Far3D: Expanding the Horizon for Surround-View 3D Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 2561–2569, 2024. doi: 10.1609/aaai.v38i3.28033. URL https://ojs.aaai.org/index.php/AAAI/article/view/28033.

Jisong Kim, Minjae Seong, and Jun Won Choi. CRT-Fusion: Camera, Radar, Temporal Fusion Using Motion Information for 3D Object Detection. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c49a28241640407b23bba8f2495f4bc9-Abstract-Conference.html.

Youngseok Kim, Jun Won Choi, and Dongsuk Kum. GRIF Net: Gated Region of Interest Fusion Network for Robust 3D Object Detection from Radar Point Cloud and Monocular Image. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10857–10864, 2020. doi: 10.1109/IROS45743.2020.9341177. URL https://doi.org/10.1109/IROS45743.2020.9341177.

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*, volume 37, pp. 1160–1168, 2023a. URL https://ojs.aaai.org/index.php/AAAI/article/view/25198.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17615–17626, 2023b. URL https://openaccess.thecvf.com/content/ICCV2023/html/Kim_CRN_Camera_Radar_Net_for_Accurate_Robust_Efficient_3D_Perception_ICCV_2023_paper.html.

Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. PointPillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12697–12705, 2019. URL https://arxiv.org/abs/1812.05784.

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, 2019. URL https://arxiv.org/abs/1904.09730.

Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M. Ni, and Lei Zhang. DN-DETR: Accelerate DETR Training by Introducing Query DeNoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13619–13627, 2022. doi: 10.1109/CVPR52688.2022.01325. URL https://openaccess.thecvf.com/content/CVPR2022/html/Li_DN-DETR_Accelerate_DETR_Training_by_Introducing_Query_DeNoising_CVPR_2022_paper.html.

Xuewu Lin, Zixiang Pei, Tianwei Lin, Lichao Huang, and Zhizhong Su. Sparse4D v3: Advancing End-to-End 3D Detection and Tracking. *arXiv preprint arXiv:2311.11722*, 2023. URL https://arxiv.org/abs/2311.11722.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14928–14937, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Lin_RCBEVDet_Radar-camera_Fusion_in_Birds_Eye_View_for_3D_Object_CVPR_2024_paper.html.

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*, pp. 531–548. Springer, 2022. URL https://arxiv.org/abs/2203.05625.

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. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023. URL https://arxiv.org/abs/2205.13542.

Yunfei Long, Abhinav Kumar, Xiaoming Liu, and Daniel Morris. RICCARDO: Radar Hit Prediction and Convolution for Camera-Radar 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22276–22285, 2025. URL https://arxiv.org/abs/2504.09086.

Ramin Nabati and Hairong Qi. CenterFusion: Center-based radar and camera fusion for 3D object detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1527–1536, 2021. URL https://arxiv.org/abs/2011.04841.

Su Pang, Daniel Morris, and Hayder Radha. TransCAR: Transformer-based camera-and-radar fusion for 3D object detection. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10902–10909. IEEE, 2023. URL https://arxiv.org/abs/2305.00397.

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*, 2020. URL https://arxiv.org/abs/2008.05711.

Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3D object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1991–1999, 2019. URL https://arxiv.org/abs/1905.12365.

Shiping Song, Jian Wu, Sumin Zhang, Yunhang Liu, and Shun Yang. Research on Target Tracking Algorithm Using Millimeter-Wave Radar on Curved Road. *Mathematical Problems in Engineering*, 2020:3749759, 2020. doi: 10.1155/2020/3749759. URL https://doi.org/10.1155/2020/3749759.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Ben Caine, et al. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2443–2451, 2020. doi: 10.1109/CVPR42600.2020.00252. URL https://waymo.com/open/.

Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. PointPainting: Sequential fusion for 3D object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4604–4612, 2020. URL https://arxiv.org/abs/1911.10150.

Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring Object-Centric Temporal Modeling for Efficient Multi-View 3D Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3621–3631, October 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Wang_Exploring_Object-Centric_Temporal_Modeling_for_Efficient_Multi-View_3D_Object_Detection_ICCV_2023_paper.html.

Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. FCOS3D: Fully convolutional one-stage monocular 3D object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 913–922, 2021. URL https://arxiv.org/abs/2104.10956.

Yue Wang, Vitor Campagnolo 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*, pp. 180–191. PMLR, 2022. URL https://proceedings.mlr.press/v164/wang22b.html.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next Generation Datasets for Self-Driving Perception and Forecasting. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021)*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/4734ba6f3de83d861c3176a6273cac6d-Abstract-round2.html.

Philipp Wolters, Johannes Gilg, Torben Teepe, Fabian Herzog, Felix Fent, and Gerhard Rigoll. SpaRC: Sparse Radar-Camera Fusion for 3D Object Detection, 2025a. URL https://arxiv.org/abs/2411.19860.

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. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7467–7474, 2025b. doi: 10.1109/ICRA55743.2025.11127412. URL https://arxiv.org/abs/2403.07746.

Zizhang Wu, Guilian Chen, Yuanzhu Gan, Lei Wang, and Jian Pu. MVFusion: Multi-View 3D Object Detection with Semantic-aligned Radar and Camera Fusion. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2766–2773, May 2023. doi: 10.1109/ICRA48891.2023.10161329. URL https://arxiv.org/abs/2302.10511. Available as arXiv:2302.10511.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18268–18278, 2023. URL https://arxiv.org/abs/2301.01283.

Yan Yan, Yuxing Mao, and Bo Li. SECOND: Sparsely Embedded Convolutional Detection. *Sensors*, 18(10):3337, 2018. URL https://www.mdpi.com/1424-8220/18/10/3337.

Bin Yang, Runsheng Guo, Ming Liang, Sergio Casas, and Raquel Urtasun. RadarNet: Exploiting radar for robust perception of dynamic objects. In *European conference on computer vision*, pp. 496–512. Springer, 2020. URL https://arxiv.org/abs/2007.14366.

Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3D object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11784–11793, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Yin_Center-Based_3D_Object_Detection_and_Tracking_CVPR_2021_paper.html.

SUPPLEMENTARY MATERIAL

A OVERVIEW

This supplementary material provides more related work (Sec. B), additional ablation studies (Sec. C), runtime analysis (Sec. D), and evaluation on 4D radar MAN TruckScenes (Sec. E) supporting the RCQFusion framework. We further provide implementation details (Sec. F), the detailed density-aware radar encoder algorithm (Sec. G), qualitative visualizations (Sec. H), and matching-strategy analysis (Sec. I).

The **additional ablations** examine the radar encoder, radar sampling enhancement, matching strategies, and supplementary multi-modal queries. The **runtime analysis** compares inference latency and memory consumption with RaCFormer, while the MAN TruckScenes experiments evaluate detection performance across weather conditions and distance ranges. The **qualitative visualizations** illustrate night, rain, and long-range cases, showing how multi-modal queries recover missed detections from camera-only predictions. Finally, the **matching-strategy analysis** presents the motivation for multi-source query hybrid matching and compares the loss and gradient dynamics of 1v1, naive 1vN, and Dynamic-K 1vN. Together, these results provide further evidence on component contributions, efficiency, and robustness.

B RELATED WORK

B.1 3D OBJECT DETECTION HEAD FRAMEWORKS

Modern 3D object detection frameworks employ a variety of detection head designs to predict bounding boxes and class labels from sensor data. These approaches can generally be categorized into anchor-based, anchor-free, and set prediction-based paradigms, each offering unique advantages and facing specific limitations.

Anchor-based detection heads regress object parameters relative to predefined anchors. SECOND Yan et al. (2018) uses sparse 3D convolutions for voxel-based LiDAR detection, while PointPillars Lang et al. (2019) encodes vertical pillars and applies a 2D convolutional detection network. For monocular detection, MonoDIS Simonelli et al. (2019) introduces disentangled 2D and 3D detection losses and a learned confidence score for 3D boxes.

Anchor-free detectors avoid predefined anchor boxes. CenterPoint Yin et al. (2021) detects object centers in a BEV heatmap and regresses 3D attributes. FCOS3D Wang et al. (2021) adapts fully convolutional one-stage detection to monocular 3D prediction, assigning targets using projected 3D centers and feature-pyramid levels.

DETR Carion et al. (2020) formulates detection as set prediction with bipartite matching. DETR3D Wang et al. (2022) uses 3D queries to sample multi-view image features, while PETR Liu et al. (2022) encodes 3D coordinates into image features for query-based detection. These camera-only methods provide relevant foundations for sparse multi-modal detectors.

B.2 MULTI-MODAL FUSION STRATEGIES

Fusion methods differ in their sensor inputs and in the representations used for cross-modal interaction.

Proposal-guided radar–camera fusion includes CenterFusion Nabati & Qi (2021), which associates radar returns with image detections through frustums and uses radar feature maps to refine object properties. CRAFT Kim et al. (2023a) associates camera proposals with radar points in polar coordinates and exchanges features through cross-attention. GRIF Net Kim et al. (2020) uses gated region-of-interest feature fusion between radar and monocular images.

Point-level augmentation includes PointPainting Vora et al. (2020), which appends image semantic-segmentation scores to LiDAR points. RadarNet Yang et al. (2020) instead combines LiDAR and radar through voxel-based early fusion and attention-based late fusion to exploit geometry and radial velocity. These methods use different sensor inputs from radar–camera-only detection.

BEV feature fusion aligns modalities on a common spatial grid. Simple-BEV Harley et al. (2023) studies camera and radar representations for BEV perception, including the effects of input resolution and batch size. BEVFusion Liu et al. (2023) unifies camera and LiDAR features in BEV for multiple perception tasks. CRN Kim et al. (2023b) uses radar-assisted view transformation, RCBEVDet Lin et al. (2024) combines radar encoding with cross-attention-based BEV fusion, and HyDRa Wolters et al. (2025b) combines hybrid fusion with radar-informed depth estimation and refinement.

Query-based fusion offers another form of cross-modal interaction. TransFusion Bai et al. (2022) initializes proposals from LiDAR BEV features and fuses image features through a transformer decoder; it therefore still uses a BEV representation. TransCAR Pang et al. (2023) first updates queries with camera features and then attends to radar features using a distance-based attention mask. Our approach builds on sparse interactions while introducing radar-conditioned multi-modal queries and hybrid supervision.

Comparison with SpaRC. SpaRC Wolters et al. (2025a) combines sparse frustum fusion, range-adaptive radar aggregation, and local self-attention for point-based radar–camera fusion without dense BEV grids, whereas RCQFusion focuses on radar-conditioned multi-modal queries and hybrid matching. We discuss SpaRC qualitatively; a controlled quantitative comparison under our training and evaluation settings is not included in this study.

C Additional Ablation Studies

Repeated-Seed Evaluation. We repeat two key ablation configurations—sampling-enhanced sparse radar–camera fusion and multi-source queries with hybrid matching—with three random seeds (0, 1, and 2) on the same machine, consistently evaluating the final-epoch checkpoints. The sample standard deviations are 0.0008–0.0012 for mAP and 0.0010–0.0012 for NDS on the 0–1 scale. These fluctuations are smaller than the corresponding reported ablation gains, and the three-run means remain close to the original single-run results, supporting the stability of the evaluated configurations.

Radar-Encoder Ablation We evaluate the radar encoder in a simplified radar-only setting to isolate the effect of its attention design. The distance-aware baseline uses a fixed Euclidean-distance mask, analogous to the fixed-radius query–radar attention mask in TransCAR Pang et al. (2023). Our density-aware attention achieves 21.21% mAP and 29.54% NDS, improving over this baseline by 0.90 and 2.31 percentage points, respectively, and also outperforming the vanilla Transformer and VoxelNeXt baselines. This comparison indicates that density-aware attention provides a more suitable receptive field than a fixed distance threshold for sparse, non-uniform radar points, because it adapts neighborhood construction to local point density.

Table 6: Comparison of radar encoders (controlled setting).

Variant	mAP (%)	NDS (%)
VoxelNeXt Chen et al. (2023b)	19.97	27.08
Vanilla Transformer	20.24	27.21
Distance-aware attention	20.31	27.23
Density-aware attention (ours)	**21.21**	**29.54**

Radar Sampling Enhancement. To isolate the effect of our radar sampling strategies, we evaluate height, scale, and density sampling under a controlled radar-augmented setting (Table 7). Scale-based sampling achieves the best performance, as it learns adaptive offsets on the BEV plane, capturing object extent more effectively than fixed vertical sampling.

Table 7: Comparison of sampling strategies (controlled setting).

Strategy	mAP (%)	NDS (%)
Height Sampling	14.24	28.62
Density Sampling	15.85	31.12
Scale Sampling	**15.98**	**32.10**

Note: Results are reported under a simplified configuration to analyze sampling behavior;

Table 8: Matching-strategy ablations on the nuScenes validation set.

Variant	mAP (%)	NDS (%)
Matching strategies		
1v1 Hungarian	53.9	60.4
Naive 1vN ($N = 4$)	55.1	61.2
Dynamic-K 1vN	55.6	62.1
N-group Dynamic-K ($N = 4$, training groups)	56.0	62.7

Matching-strategy ablations. Table 8 reports the corresponding nuScenes validation results. Naive 1vN improves over 1v1 by **1.2** and **0.8** percentage points in mAP and NDS, respectively, but its optimization curves show unstable regression. Dynamic-K improves over naive 1vN by a further **0.5** and **0.9** percentage points in mAP and NDS, respectively, consistent with its more stable regression behavior. The N-group Dynamic-K variant obtains the highest scores (**56.0% mAP** and **62.7% NDS**), but uses additional training query groups; our same-query Dynamic-K variant retains most of this gain without introducing those groups. Under matched VoVNet-99 and 24-epoch training settings, our Group-DETR-style implementation with additional query groups increases training time by approximately **35%–45%** relative to our default formulation without extra query groups, depending on the number of groups and available GPU memory. Our default Dynamic-K 1vN instead applies auxiliary matching directly to the original query set, avoiding this additional group-wise decoding cost during training.

Effectiveness of Multi-Modal Queries. We compare Far3D's camera 2D queries Jiang et al. (2024) with our radar and radar–camera associated queries, using matched hyperparameters and training schedules for shared components. After removing randomly initialized and historical memory queries (900 in total in this setting), we evaluate predictions from the remaining queries.

As shown in Tab. 9, our multi-modal queries achieve **50.25% mAP** and **59.58% NDS**, compared with **0.2% mAP** and **2.1% NDS** from Far3D's camera 2D queries. They also achieve comparable mAP and higher NDS than Far3D's full query set (50.7% mAP and 58.7% NDS). These results support the effectiveness of radar-informed queries for detection, with qualitative comparisons in Fig. 4–Fig. 7.

Table 9: Supplementary Query Ablation Results

	mAP (%)	NDS (%)
Camera 2D queries in Jiang et al. (2024)	0.2	2.1
All queries in Jiang et al. (2024)	50.7	58.7
Supplementary multi-modal queries (ours)	50.3	59.6

,

D Runtime Analysis

Matched inference comparison with RaCFormer. We also compare RCQFusion with the radar–camera query detector RaCFormer under a matched ResNet-50, 704×256 inference setting with autocast FP16. Table 10 reports the latency breakdown. RCQFusion takes 56.1 ms (17.8 FPS), compared with 216.0 ms (4.6 FPS) for RaCFormer, and uses 2.76 GB peak inference memory instead

of 4.55 GB. Thus, in this setting RCQFusion is about 3.9× faster and uses about 39% less peak memory. RaCFormer's decoder accounts for most of its measured latency.

Table 10: Matched inference latency (ms) of RCQFusion and RaCFormer with ResNet-50 at 704×256 using autocast FP16.

Method	Camera	Radar	LSS	Fusion / Query	Decoder	Total
RCQFusion	22.4	9.6	–	9.0	15.1	**56.1**
RaCFormer	21.3	11.2	10.7	–	172.8	216.0

Peak memory was measured using a unified inference protocol. In the reported expanded-range setting (up to $[-51.2, 102.4]$ m), RaCFormer latency increases by about 56 ms and peak memory by 0.4 GB, whereas RCQFusion shows no measurable increase. This is consistent with the cost of dense BEV sampling growing with the spatial grid, while RCQFusion processes sparse radar points and queries.

E Evaluation on MAN TruckScenes

To examine whether the proposed design generalizes beyond nuScenes' sparse automotive radar, we further evaluate RCQFusion on MAN TruckScenes Fent et al. (2024), a large-scale 4D radar dataset for long-range perception and challenging driving scenarios. The dataset contains 6 LiDARs, 4 cameras, and 6 4D radars, with 523 training scenes, 75 validation scenes, and 23,902 samples in total. Its data format and evaluation protocol are similar to nuScenes, but the 4D radar setting provides denser and longer-range geometric observations.

For this evaluation, we adopt a VoVNet-99 visual pipeline similar to Far3D, train for 60 epochs with a training recipe close to StreamPETR, and use a detection range of 152.4 m longitudinally and 51.2 m laterally. Both the historical query set and learnable query set are kept at 900. Tab. 11 reports the validation results.

Table 11: Validation results on MAN TruckScenes.

Method	Input	Backbone	mAP (%) ↑	NDS (%) ↑
CenterPoint-V Fent et al. (2024)	LiDAR	Voxel	22.6%	35.3%
Far3D Jiang et al. (2024)	Camera	VoV-99	10.2%	18.5%
RCQFusion	Radar+Camera	VoV-99	**23.9%**	**36.0%**

Table 12: Weather-wise and distance-wise mAP (%) on the MAN TruckScenes validation set.

Method	Weather			Distance			
	Day	Rain	Night	0–25 m	0–50 m	0–100 m	0–150 m
Far3D Jiang et al. (2024)	10.4%	6.2%	5.4%	29.8%	23.0%	13.5%	10.2%
RCQFusion	**24.2%**	**15.2%**	**18.0%**	**41.2%**	**35.2%**	**28.0%**	**23.9%**

We also apply the same weather-wise and distance-wise protocols used in our nuScenes analysis. Tab. 12 shows that RCQFusion consistently outperforms the camera-only baseline across adverse conditions and distance ranges.

Compared with Far3D, RCQFusion improves mAP by +13.8%, +9.0%, and +12.6% under Day, Rain, and Night conditions, respectively. Across distance ranges, the gains are +11.4% at 0–25 m, +12.2% at 0–50 m, +14.5% at 0–100 m, and +13.7% at 0–150 m. The gains remain large beyond 50 m and under rain or night conditions, indicating that the proposed radar-camera query design transfers well to 4D radar and remains effective in long-range and adverse-condition perception.

F IMPLEMENTATION DETAILS

Training Settings. We adopt Far3D Jiang et al. (2024) and StreamPETR Wang et al. (2023) as our primary baselines to ensure a fair and comprehensive evaluation. Our network architecture closely follows that of Far3D, utilizing an image backbone, a Feature Pyramid Network (FPN) neck, and a 2D detection head for image feature extraction and object localization.

Our model is optimized using AdamW, with a global batch size of 8 distributed across available GPUs. The learning rate is initialized at 2×10^{-5} for the image backbone and 2×10^{-4} for all other network parameters. We employ a cosine annealing schedule for learning rate decay throughout training. Unless otherwise stated, we do not use advanced data augmentation techniques such as Class-Balanced Grouping and Sampling (CBGS), test-time augmentation, or future-frame aggregation, in order to provide a clear and fair comparison with prior works.

Backbone Settings and Fair Comparison Protocol. For image feature encoding, we experiment with both ResNet-50 He et al. (2016) and VoVNet-99 Lee et al. (2019) to ensure fair comparison under commonly adopted CNN backbones. ResNet-50 is pre-trained on nuImages, while VoVNet-99 follows the FCOS3D Wang et al. (2021) pre-training pipeline to provide strong 3D detection priors.

To guarantee fairness, all validation-set comparisons are conducted under **identical backbones and identical input resolutions** (ResNet-50 and VoVNet-99 at 704×256 and 800×320), matching CRN Kim et al. (2023b), RCBEVDet Lin et al. (2024), RaCFormer Chu et al. (2025), as well as recent fusion baselines Kim et al. (2024); Wolters et al. (2025b); Kim et al. (2023a). These settings allow strict apples-to-apples evaluation on the validation set, where all fairness-critical analysis is performed.

High-resolution ViT Backbone on the Test Set. The nuScenes test server supports very limited submission attempts per year, and occasionally becomes inaccessible for long periods. Consequently, as is standard practice in nuScenes 3D detection, test-set submissions typically use the **best-performing configuration**, rather than re-running all backbone–resolution combinations. Following StreamPETR and Far3D Wang et al. (2023); Jiang et al. (2024), we therefore adopt a pure transformer-based **ViT backbone with high input resolution** for the test-set submission.

For comparison with RaCFormer Chu et al. (2025), a strong recent radar–camera baseline, we additionally evaluate RCQFusion with VoVNet-99 at 1600×640, matching its image backbone and input resolution. To reflect realistic autonomous-driving deployment, none of our reported test-set results use future-frame inputs.

Radar Modality. For each sample, we aggregate the current radar frame with up to 5 historical radar sweeps to obtain denser point clouds, yielding at most 2000 radar points per sample. Multi-sweep radar aggregation is also used in prior camera–radar fusion frameworks Kim et al. (2023b); Lin et al. (2024); Kim et al. (2024); the sweep count here specifies our own input setting. Unlike newer fifth-generation radars such as the Continental ARS540, the Continental ARS408 radar used in nuScenes does not provide cluster-level attributes or pre-processed object lists. As a result, point density and neighborhood statistics must be computed explicitly. Although this introduces a small additional CPU overhead, the computation is lightweight (on the order of only a few milliseconds for 2000 points) and does not affect the real-time behavior of the overall perception pipeline. Following CRNKim et al. (2023b) and other prior work, these computations are performed offline during preprocessing, and thus impose no latency on the online inference stage.

It is also important to note that we apply *no* attribute-based filtering or denoising to the radar point clouds. All radar returns are preserved, ensuring that our evaluation reflects realistic and challenging operating conditions associated with automotive-grade radars.

Radar Encoder Inference Acceleration. To reduce online latency, the multi-sweep Transformer encoder could be trained with intra-sweep density masks and causal cross-sweep attention, preventing earlier sweeps from accessing later sweeps at every encoder layer. With consistent coordinate/time encodings and memory-window policies, per-layer keys and values from the preceding n sweeps could be cached, so only newly arriving points require encoder updates. This would avoid repeatedly

encoding historical points, offering a way to accelerate radar perception for autonomous driving at the cost of additional cache memory.

This experimental protocol is designed to ensure that all results are directly comparable and that improvements can be attributed to our proposed method rather than differences in training settings or data preprocessing.

G Density-Aware Masked Attention

G.1 Details

The **Density-Aware Masked Attention** module enhances point cloud feature learning by dynamically constructing attention masks based on local density. Given a point cloud $\mathcal{P}$ with associated coordinates $\mathcal{L}$, the algorithm iteratively computes pairwise distances to identify local neighborhoods within a threshold τ_l for each layer. Points with sufficient neighbors (at least κ_l) are marked as core points, and clusters are expanded using a DBSCAN-like procedure Ester et al. (1996). A binary mask $M^{(l)}$ is then generated, indicating whether pairs of points belong to the same local cluster. This mask constrains the multi-head attention in each layer, allowing information aggregation only within density-aware regions. In the final stage, global attention is applied to capture long-range dependencies. The output is a set of point features that are adaptively aggregated according to the underlying spatial density structure. The pseudocode is given in Alg. 1.

Algorithm 1 Density-Aware Masked Attention

Require: Point cloud $\mathcal{P} \in \mathbb{R}^{N\times d}$, positional coordinates $\mathcal{L} \in \mathbb{R}^{N\times 3}$, thresholds $\{\tau_l, \kappa_l\}_{l=1}^{k-1}$

$D \leftarrow$ pairwise_distance($\mathcal{L}$)
for $l = 1$ **to** $k-1$ **do**
  $M^{(l)} \leftarrow \mathbf{0}_{N\times N}$
  for $i = 1$ **to** N **do**
    $\mathcal{N}_i \leftarrow \{j : D_{ij} \le \tau_l\}$ ▷ Find neighbors within distance threshold
    if $|\mathcal{N}_i| \ge \kappa_l$ **then**
      $C_{i,j}^{(l)} \leftarrow 1 \; \forall j \in \mathcal{N}_i$ ▷ Core point validation
    end if
  end for
  ClusterLabels $\leftarrow$ DBSCAN_Expand($\mathcal{L}, \mathrm{C}^{(l)}, \tau_l$) ▷ Expand clusters from core points (assign labels)
  $M^{(l)} \leftarrow$ (ClusterLabels $\times$ ClusterLabels$^\top$) > 0 ▷ Generate density mask (same cluster = 1)
  $\mathcal{F}^{(l)} \leftarrow$ MaskedMultiHeadAttention($\mathcal{F}^{(l-1)}, M^{(l)}$)
end for
$\mathcal{F}^{(k)} \leftarrow$ GlobalMultiHeadAttention($\mathcal{F}^{(k-1)}$) ▷ Final layer uses full attention
return $\mathcal{F}^{(k)}$

H Visualization in Adverse Weather and Lighting

We further compare our approach with Far3D through qualitative visualizations under extreme weather and poor lighting conditions. Beyond the missed detections in single-visual-modality scenarios illustrated in Fig. 4 and Fig. 5, Fig. 6 and Fig. 7 demonstrate that the purely visual solution of Far3D (as shown in its corresponding columns) suffers from issues such as **inaccurate depth estimation for distant targets** and **redundant predictions for small targets like VRUs** (Vulnerable Road Users) along the ray direction.

The root cause of these problems lies in the absence of precise depth information within visual features. Traditional methods project visual features onto the ray directions of targets and vehicles, while query-based methods also introduce redundancies due to limitations in query sampling. In contrast, our proposed RCQFusion addresses these challenges from two perspectives: performing sparse radar-camera fusion and generating multi-modal queries. By effectively leveraging the depth information inherent in radar features and the accurate 3D positions provided by radar queries, our method resolves these two prevalent issues.

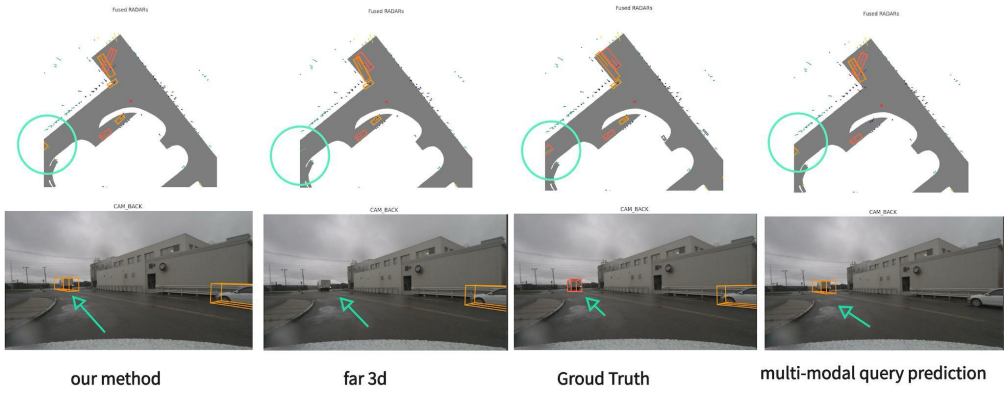


Figure 4: Visualization on the nuScenes validation set. The bounding boxes indicate our detection results, with the upper section presenting the radar perspective: the origin represents the sensor position, and the length of each line corresponds to the speed magnitude. As illustrated, compared with the pure vision-based Far3D Jiang et al. (2024) solution, which fails to detect distant vehicle targets (highlighted by the green arrow in the camera view and the green circle in the radar perspective), our proposed multi-modal query approach leverages **radar's sensitivity to moving targets and metallic objects**. This integration significantly improves detection performance for vehicle targets under challenging conditions.

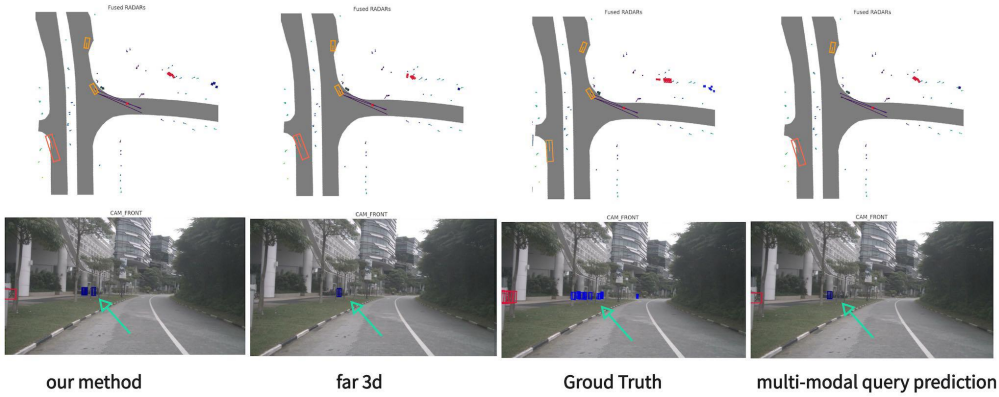


Figure 5: Visualization on the nuScenes validation set. The figure presents a top-down view with radar data alongside the corresponding camera perspective. Compared to the pure vision-based Far3D solution, which detects only a single pedestrian target (indicated by the green arrow in the camera view and the green circle in the radar perspective), our proposed multi-modal query approach leverages **precise 3D positioning from radar** to substantially enhance detection performance for VRU (Vulnerable Road User) targets.

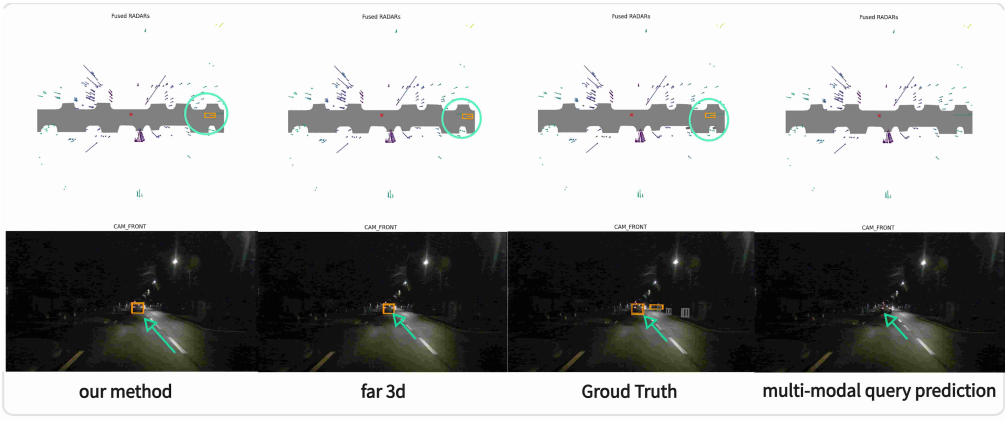


Figure 6: Visualization on the nuScenes night scene. As shown in the figure, under low-light conditions, the vision-based Far3D Jiang et al. (2024) approach (indicated by green arrows and circles from different perspectives) exhibits inaccuracies in depth estimation for front-target detection. In contrast, our multi-sensor fusion solution leverages accurate positional information from radar to significantly improve the **precision of target depth estimation**.

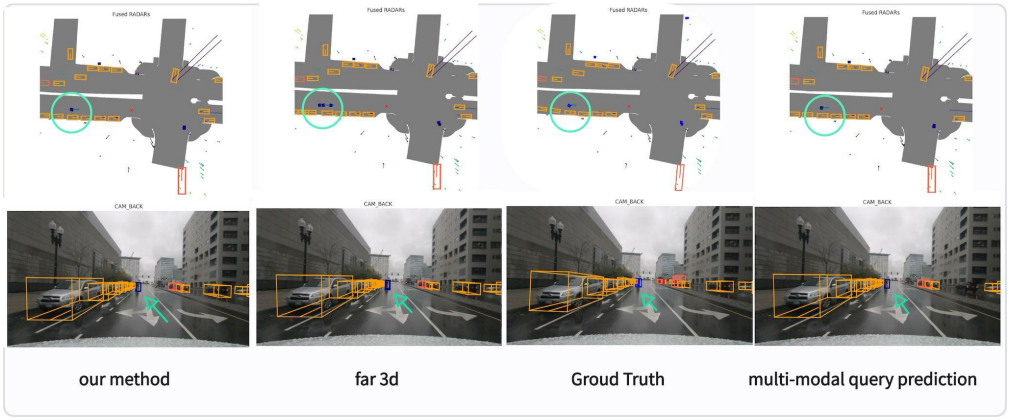


Figure 7: Visualization on the nuScenes rainy scene. As shown in the figure, the vision-only Far3D scheme produces **redundant predictions along the ray direction for small targets** such as VRUs (Vulnerable Road Users), as indicated by green arrows and circles from different perspectives. In contrast, our multi-sensor fusion approach effectively leverages radar information to mitigate this issue.

I MATCHING-STRATEGY COMPARISON: 1V1, NAIVE 1VN, AND DYNAMIC-K 1VN

I.1 THEORETICAL STANDPOINT FOR MULTI-SOURCE QUERY HYBRID MATCHING

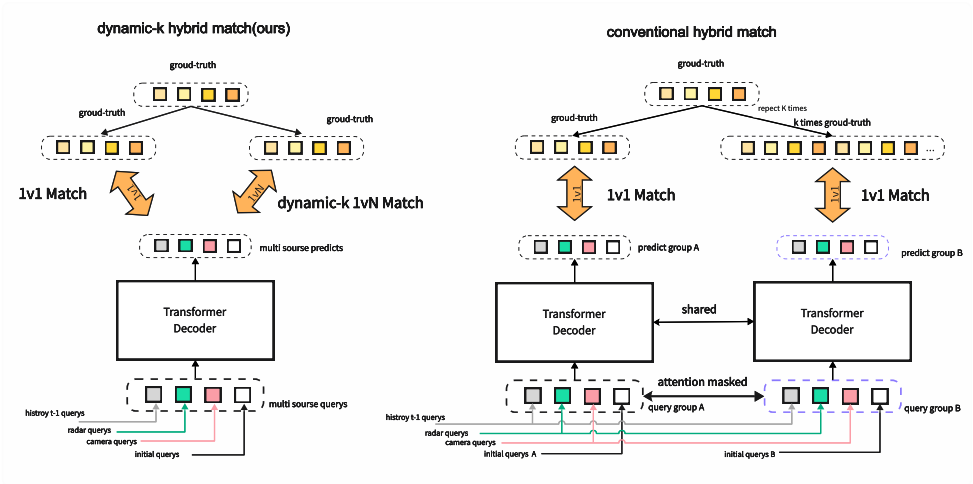


Figure 8: Comparison of our dynamic-K hybrid matching and conventional hybrid matching.

In the **multi-modal query generation** stage, we introduce multiple heterogeneous queries originating from non-homogeneous sources. Specifically, we construct queries from millimeter-wave radar point clouds, from radar–camera fusion detections, and from independent 2D visual detections. Following Wang et al. (2023); Jiang et al. (2024); Lin et al. (2023), we additionally incorporate top-K historical-memory queries and a small set of randomly initialized learnable queries. This heterogeneous design enriches BEV sampling density and reduces the modality-induced bias observed in purely self-learned DETR queries.

Limitations of 1-to-1 Matching Under Multi-modal Query Fusion. DETR-style detectors Carion et al. (2020); Liu et al. (2022) use one-to-one (1v1) bipartite matching to supervise set prediction without NMS. This assignment does not require homogeneous or independent queries. In our multi-modal setting, however, multiple queries may correspond to the same physical object. Radar clusters naturally yield several queries for the same target; radar and visual queries frequently refer to the same instance; self-learned queries may overlap with either modality. Under a strict 1v1 constraint, these semantically equivalent queries are forced to *compete* for a single ground truth (GT), leading to **false negatives, unstable gradients, and slow early-stage convergence**.

Formally, DETR-style matching is a nested optimization:

$$\min_{\theta} \mathbb{E}_x \left[\min_{\pi \in \mathcal{S}_N} \sum_i \mathcal{L}(y_i, \hat{y}_{\pi(i)}; \theta) \right]. \tag{7}$$

In multi-modal scenarios, the optimal permutation $\pi^\star$ becomes *non-unique* due to multiple queries being equally valid matches for the same GT. This induces discontinuities in the matching landscape and amplifies the variance of the training gradients.

Hybrid Match for Multi-modal Alignment. Motivated by Li et al. (2022); Jia et al. (2023); Chen et al. (2023a), we generalize hybrid matching to multi-modal supervision and propose a **hybrid match** mechanism with three key components:

- **Auxiliary 1vN supervision:** A shared-decoder one-to-many (1vN) branch complements the main 1v1 branch and supplies additional positive samples.

- **Cross-modal positive consistency:** Radar-, vision-, and fusion-based queries corresponding to the same instance are jointly treated as positives.
- **Adaptive multiplicity:** The number of positives is controlled by either a fixed K or a *dynamic-K* strategy based on IoU or geometric proximity.

Comparison to Conventional Hybrid Matching. As illustrated in Fig. 8, the conventional hybrid matching strategy Jia et al. (2023) introduces an additional learnable query group at initialization. Two groups pass through the decoder in parallel with shared weights, using an attention mask to prevent cross-group interference. The first group uses standard 1v1 matching, whereas the second replicates each GT K times to construct a 1-to-K branch.

In contrast, our framework does *not* create additional learnable query groups. The proposed dynamic-K 1vN assignment is applied directly to decoder predictions, preserving a **single, unified query set** without modifying decoder architecture.

This avoids the additional memory usage, compute overhead, and training-time increase caused by multi-group query decoding. Despite the simpler design, dynamic-K still provides effective positive enrichment and stabilizes optimization.

The conventional 1vN hybrid-matching baseline in Fig. 8 introduces additional query groups as a confounding factor: performance gains may come from either the matching scheme or the enlarged query set. We conduct controlled matching experiments to distinguish these effects.

Dynamic-K Matching. Rather than globally replicating GTs, we adopt a **dynamic-K strategy**. For each GT, we perform nearest-neighbor search over predicted queries in the shared BEV embedding space and select those within a distance or IoU threshold as positives. The resulting K_{dyn} varies adaptively per instance. This ensures that only high-quality, geometrically aligned candidates contribute to the 1vN supervision, reducing false positives and preventing over-regularization.

I.2 Training Curve Comparison

Formally, dynamic-K can be viewed as a *soft relaxation* of Hungarian matching:

$$\pi^\star(i) = \arg\min_{j \in \mathcal{P}(i)} \mathcal{L}(y_i, \hat{y}_j), \tag{8}$$

where $\mathcal{P}(i)$ is the dynamically selected geometric neighborhood. This relaxation expands the feasible matching space, reduces gradient variance $\mathrm{Var}(\nabla_\theta L)$, and empirically mitigates optimization oscillation in early training.

To thoroughly evaluate the effect of different label assignment strategies in multi-modal DETR-style detectors, we compare three variants under identical training conditions:

- **1v1 Matching (Baseline)**: the standard DETR Hungarian assignment.
- **Naive 1vN Hybrid Matching**: a hybrid-matching variant in which each GT is replicated K times and 1v1 Hungarian matching is adapted for an auxiliary 1vN branch.
- **Dynamic-K 1vN Hybrid Match (Ours)**: an adaptive multi-positive assignment using IoU matching for an auxiliary 1vN branch.

Hybrid Match Always Includes a 1v1 Branch. Following the main one-to-one and auxiliary one-to-many supervision design of hybrid matching Jia et al. (2023), all hybrid-matching variants in our experiments retain a full 1v1 branch. The additional 1vN supervision is always implemented as an *auxiliary* branch sharing the same decoder and query set, without altering the network architecture.

This ensures that the end-to-end, NMS-free nature of DETR is preserved during inference, and that any performance differences stem solely from the auxiliary assignment mechanism.

Experimental Configuration. All experiments use the same ResNet-50 backbone and optimizer, with 20k training iterations. Unlike prior hybrid matching approaches that introduce additional learnable query groups Jia et al. (2023); Chen et al. (2023a), our Dynamic-K 1vN method reuses the *same* original queries to avoid confounding factors and extra compute.

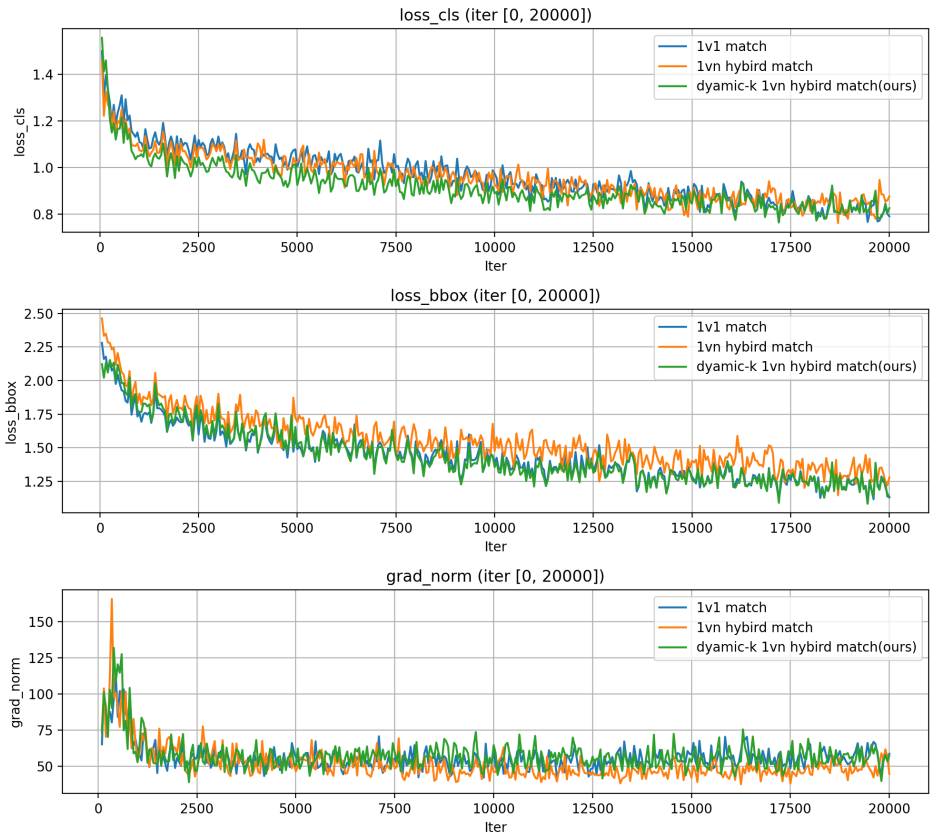


Figure 9: Comparison of loss and gradient curves. **1v1 matching**: the standard DETR Hungarian assignment. **Naive 1vN hybrid matching**: a hybrid-matching variant in which each GT is replicated K times for an auxiliary 1vN branch. **Dynamic-K 1vN hybrid matching (ours)**: an adaptive multi-positive assignment using IoU matching for an auxiliary 1vN branch.

To address difficulties with IoU matching during early training, we adopt an area-aware IoU threshold: for large objects (area > 4), we set the IoU threshold to 0.2; for small objects, we relax it to 0.01 to ensure sufficient positives. During later training, the IoU threshold can be tightened to promote more accurate regression.

The training loss is defined as:

$$\mathcal{L}_{\text{total}} = (1 - \lambda_{\text{aux}})\mathcal{L}_{\text{one2one}} + \lambda_{\text{aux}}\mathcal{L}_{\text{one2many}}, \tag{9}$$

where λ_{aux} balances the main and auxiliary losses.

Observed Optimization Behaviors. Fig. 9 compares loss and gradient dynamics across the three methods:

- **1v1 Matching** provides stable optimization but $loss_{cls}$ suffers from slow early convergence because the number of positives is limited.
- **Naive 1vN Hybrid Matching** achieves faster early classification due to more positives, but produces numerous geometrically inconsistent matches caused by GT replication and naive 1v1 Hungarian matching. This results in persistently high regression loss and a sharp decline in gradient norm in later stages—indicating training collapse.
- **Dynamic-K 1vN Hybrid Match (Ours)** achieves fast early classification *and* stable late-stage regression. Gradient norms remain high throughout training, and regression loss aligns closely with the 1v1 baseline while significantly outperforming naive 1vN.

Underlying Cause. Under focal loss, gradients from low-quality positives quickly vanish. Naive 1vN introduces many such false positives, degrading learning signals and creating optimization stagnation.

Dynamic-K selects positives based on IoU or geometric proximity, ensuring only reliable matches are used as auxiliary supervision. This reduces gradient variance, stabilizes optimization, and avoids the regression degradation inherent in naive GT replication.

Conclusion. Across the matching comparisons, Dynamic-K 1vN offers a favorable trade-off: it converges faster early than 1v1 and provides more stable regression than naive 1vN. It improves validation performance from **55.1%/61.2%** to **55.6%/62.1%** mAP/NDS while reusing the same query set. N-group Dynamic-K achieves the highest scores (**56.0%/62.7%**), but uses additional training query groups. Thus, Dynamic-K improves the matching strategy without relying on an expanded query set.