The Role of Radar in End-to-End Autonomous Driving

Anonymous ICCV submission

Paper ID 12

Abstract

End-to-end autonomous driving *systems promise stronger performance through unified optimization of perception, motion forecasting, and planning. However, vision-based approaches face fundamental limitations in adverse weather conditions, partial occlusions, and precise velocity estimation - critical challenges in safety-sensitive scenarios where accurate motion understanding and long-horizon trajectory prediction are essential for collision avoidance. To address these limitations, we propose* ***SpaRC-Drive****, a query-based end-to-end camera-radar fusion framework for planning-oriented autonomous driving. Through sparse 3D feature alignment, and doppler-based velocity estimation, we achieve strong 3D scene representations for refinement of agent anchors, map polylines and memory modelling. Our method achieves strong improvements over the state-of-the-art vision-only baselines across multiple autonomous driving tasks, including 3D detection ($+4.8\%$ mAP), multi-object tracking ($+8.3\%$ AMOTA), online mapping ($+1.8\%$ mAP), motion prediction (-4.0% mADE), and trajectory planning ($-0.1m$ L2 and -9% TPC). We achieve both spatial coherence and temporal consistency on multiple challenging benchmarks, including* ***real-world open-loop nuScenes****, long-horizon T-nuScenes, and* ***closed-loop simulator Bench2Drive****. We show the effectiveness of radar-based fusion in safety-critical scenarios where accurate motion understanding and long-horizon trajectory prediction are essential for collision avoidance. The source code of all experiments will be made available.*

1. Introduction

Autonomous driving systems have evolved from modular, multi-stage perception pipelines to unified end-to-end learning frameworks that directly map raw sensor inputs to vehicle control commands [2]. While conventional approaches decompose the driving task into independent modules for 3D object detection [9, 35, 40], multi-object tracking [41, 45, 48], and online mapping [24, 27, 44], recent end-to-end methods [8, 47] demonstrate the advantages of joint optimization across perception, prediction, and planning tasks.

The new optimization objective is to generate driving controls and trajectories for the ego vehicle, directly from sensor inputs of cameras, LiDARs, and radars [4]. Leveraging expert demonstrations through imitation learning, raw sensor signals are directly processed to output vehicle motion plans and intermediate representations optimized towards the final planning goal. Initially in Bird's Eye View (BEV) representations [28], the future trajectory of the ego vehicle is regressed from an ego-token within a transformer decoder, reducing the problem to a supervised learning setting [34].

However, state-of-the-art research has focused on vision-centric approaches, limiting their robustness in challenging scenarios such as adverse weather conditions, partial occlusions, and long-range detection.

Critical for planning safety: robust depth estimation, strong motion-forecasting, stable trajectories. Song et al. have showed, that especially in turning scenarios, models suffer from unstable trajectories, vulnerability to occlusions and temporal inconsistencies [33]. The implicit depth modeling in query-based transformers lacks geometric constraints, leading to substantial localization errors in 3D perception due to unreliable depth estimation [36]. Due to noise from highly dynamic environments and following detection errors, uncertainties arise in long-time horizon and long-range planning. Moreover, causal confusion and the reliance on temporal smoothness of the ego trajectory and past motion pose a challenge [23].

Radar sensors provide critical advantages that address fundamental limitations of vision-centric approaches in end-to-end autonomous driving. Their robust long-range detection capabilities beyond 150m, direct velocity measurements through Doppler effects, and weather-independent operation enable more reliable spatial reasoning through time-of-flight range measurements. Additionally, radar's ability to measure relative velocities enhances multi-agent intent prediction, leading to more stable and consistent trajectory planning. These complementary strengths make radar fusion particularly valuable for safety-

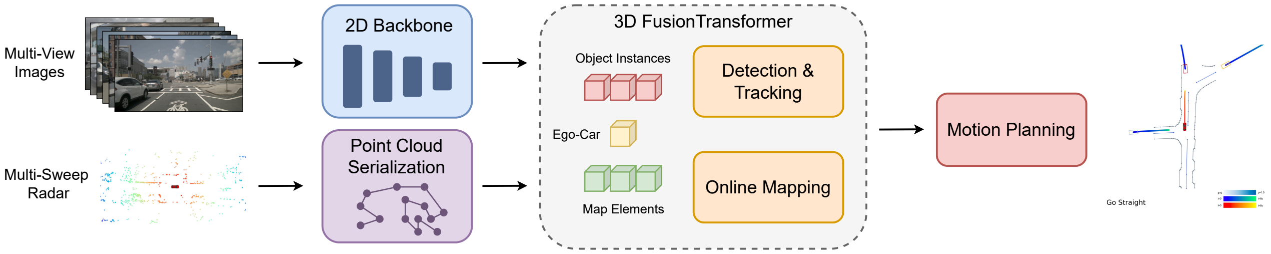

Figure 1. **Overview of SpaRC-Drive**. We propose a query-based end-to-end camera-radar fusion framework for autonomous driving that jointly optimizes perception, prediction and planning.

critical autonomous driving applications.

While multi-modal fusion with cameras and LiDAR has shown benefits [4], and radar fusion has proven effective for modular perception [5, 26, 37], the integration of radar into end-to-end autonomous driving remains unexplored. We investigate the impact and potential of including radar into the end2end optimization and how to leverage the additional motion cues reflected from the environment. Due to sparsity of the radar representation and precise spatial-temporal calibration, we propose a query-based approach that iteratively refines the motion and positional charactersitcs of map and traffic agent representations.

In this work, we address the critical gap in radar-based end-to-end autonomous driving by proposing SpaRC-Drive, extending the sparse representation paradigm of radar points and scene instances in a coherent end-to-end framework, and creating synergies between radar data characteristics and planning requirements. Our approach iteratively refines motion and positional characteristics of both map and agent representations by leveraging spatial proximity of reflected radar points as strong inductive biases.

Our main contributions are:

- First radar-based end-to-end autonomous driving baseline on key benchmarks.
- Extension of sparse fusion design for simultaneous detection, tracking, and planning queries.
- Holistic radar-based fusion improves 3D detection (+4.8% mAP), multi-object tracking (+8.3% AMOTA), online mapping (+1.8% mAP), and motion forecasting (-4.0% mADE), optimizing trajectory prediction consistency (-9.0% TPC) and simulation success rates (+10.0%).
- Extensive evaluation on multiple benchmarks of open-loop nuScenes [1] and closed-loop simulation of Bench2Drive [12].
- We provide additional qualitative analysis demonstrating superior performance through enhanced perception range, more accurate motion modeling, and increased robustness under challenging environmental conditions.

2. Related Work

2.1. Planning Oriented Autonomous Driving

A new paradigm has emerged in autonomous driving research, moving from multi-stage frameworks [11, 20, 21] to end-to-end autonomous driving [2]. This evolution addresses the fundamental limitations of modular approaches: information loss and error accumulation across subsequent, which constrain optimal system performance. The goal is to strengthen generalization to complex driving scenarios in a data-driven manner.

Typically the state-of-the-art methods follow an encoder-decoder principle, first encoding the sensor data into a latent representation, then decoding the intermediate representation into a driving policy [4, 7]. The pioneering works of UniAD [8] and VAD [3, 14] have recently shown that all tasks are communicated within unified query interfaces, enabling goal-oriented optimization through vectorized scene representations. VADv2 [3] extends the planner to probabilistic planning, while Hydra-MDP [22] integrates additional supervision from rule-based planning modules. SparseDrive [34] explores sparse scene representations for efficient scene modelling, discarding Birds-Eye-View (BEV) representations.

2.2. Camera-Radar 3D Perception

In 3D object detection, radar-camera-based approaches have emerged as low-cost and robust alternative to lidar-based perception. Initial works fused in the perspective view [17, 29–31], associating the sparsely projeceted radar-points to the dense encoded image features.

Grid-rendering approaches have adapted the BEVFusion [28] paradigm to the characteristics of radar sensors [15, 16, 18, 26, 37] have been proposed. Encoded by PointPillar [19] or VoxelNet [50], dense paramatrized, but sparse in information density, feature maps are combined in BEV space. CRN [18], HyDRa [37], and RCBEVDet [26] tackle the spatial misalignment between radar and camera sensors, surpassing vision-based approaches in stronger velocity prediction, depth estimation and robustness in adverse weather conditions.

While RaCFormer [5] still utilizes BEV-encoded radar features but decodes the features via sampling in a transformer, SpaRC [36] proposes a new state-of-the-art in 3D object detection via fully sparse encoding and fusion of radar points. Through point cloud serialization in the backbone, it enables a direct point-to-object interaction, dynamically weighted, with strong priors for the subsequent perspective aggregation and hierarchical query optimization. We will leverage these principles to design a query-based fusion of radar points and scene instances of the full surrounding environment, relevant for the planning task. reducing the spatial and temporal uncertainty

3. Architecture

3.1. Framework Overview.

SpaRC-Drive extends the sparse-centric transformer of SparseDrive [34] by integrating the adaptive radar fusion strategies from SpaRC [36] into a unified end-to-end autonomous driving framework. Our approach addresses the fundamental challenge of fusing radar representations with dense visual features in a planning-oriented optimization pipeline.

The overall architecture consists of three main components: (1) multi-modal sparse feature encoding that processes camera and radar inputs into compatible representations, (2) unified sparse fusion that leverages query-based interactions between modalities, and (3) parallel motion planning that jointly optimizes the strengthened spatial scene representations for perception, prediction, and trajectory generation. This design enables direct end-to-end fusion and optimization without leveraging inefficient grid-based representations.

Our framework processes 360-degree surround-view images through a 2D convolutional neureal network backbone with a feature pyramid neck, generating multi-view multi-scale feature maps. Simultaneously, multi-sweep radar point clouds (spatial coordinates, RCS intensity, and Doppler velocity) are encoded into sparse feature representations through point-wise encoding and serialization using Point Transformer [38], producing a set of 3D embeddedradar features.

3.2. Query Design

Detection queries represent surrounding traffic agents as anchor boxes with eleven parameters: $x, y, z, \ln w, \ln h, \ln l, \sin\theta, \cos\theta, v_x, v_y, v_z,$ where spatial coordinates, dimensions, orientation, and velocity are jointly predicted and optimized. These anchors $\mathbf{B}_d \in \mathbb{R}^{N_d \times 11}$ are paired with instance features $\mathbf{F}_d \in \mathbb{R}^{N_d \times C}$ obtained through K-means clustering on the training set.

Map element queries model static road infrastructure as polylines with N_p waypoints: $x_0, y_0, x_1, y_1, \ldots, x_{N_p-1}, y_{N_p-1}$. Map instances are represented by features $\mathbf{F}_m \in \mathbb{R}^{N_m \times C}$ and anchor polylines $\mathbf{L}_m \in \mathbb{R}^{N_m \times N_p \times 2}$, with each element containing up to 20 waypoints.

3.3. Sparse Fusion

Following SpaRC's design [36], we implement range-adaptive aggregation that dynamically weights radar features based on their spatial proximity to query locations. We aggregate nearby radar features for each query instance using distance-weighted attention that dynamically adjusts feature importance based on spatial proximity:

$$\text{Attn}(\mathbf{q}, \mathbf{k}, \mathbf{v}) = \text{softmax}\left(\frac{\mathbf{q}\mathbf{k}^T}{\sqrt{d}} - \alpha \frac{\|\mathbf{p}_q - \mathbf{p}_k\|_2}{r_{\text{max}}}\right)\mathbf{v} \quad (1)$$

where $\mathbf{q} \in \mathbb{R}^{N_q \times d}$ queries attend to radar key-value pairs $\mathbf{k}, \mathbf{v} \in \mathbb{R}^{N_k \times d}$ via scaled dot-product attention with a distance-based penalty term. The 3D positions $\mathbf{p}_q \in \mathbb{R}^{N_q \times 3}$ and $\mathbf{p}_k \in \mathbb{R}^{N_k \times 3}$ are normalized by r_{max}.

For map elements, we compute the minimum distance between a radar point and polyline segments:

$$d_{\text{min}} = \min_{i=1}^{N_p-1} \|\mathbf{p}_r - (\mathbf{p}_i + t \cdot (\mathbf{p}_{i+1} - \mathbf{p}_i))\|_2 \quad (2)$$

where $\mathbf{p}_r$ is the radar point position, $\mathbf{p}_i$ and $\mathbf{p}_{i+1}$ are consecutive polyline points, and t is the projection parameter clamped between 0 and 1. The projection parameter t is computed as:

$$t = \text{clamp}\left(\frac{(\mathbf{p}_r - \mathbf{p}_i) \cdot (\mathbf{p}_{i+1} - \mathbf{p}_i)}{\|\mathbf{p}_{i+1} - \mathbf{p}_i\|_2^2}, 0, 1\right) \quad (3)$$

This distance metric enables effective attention between radar points and map elements by considering the closest line segment of each polyline. After adar-based set-to-set aggregation, the decoder module encompases iterative blocks of deformable perspective aggregation, self-attention and feedforward networks. While the deformable aggregations uses learnable keypoints around the anchor boxes, radar module aggregates dynamically the closest radar features in the vicinity of the anchor boxes and polyline.

3.4. Multi-modal perspective feature maps

To align multi-modal features across perspective and 3D representations, we additionally employ sparse frustum fusion that projects radar points into camera frustums and performs cross-attention between radar features and image regions. Thus, the ego-vehicle instance benefits directly from the radar-enriched representation, when Avergage Pooling the feature representation into a single query intialization.

Method	**Input**	**Backbone**	**L2 (m) ↓**				**Col. Rate (%) ↓**				**TPC (m) ↓**			
			1s	2s	3s	Avg.	1s	2s	3s	Avg.	1s	2s	3s	Avg.
UniAD [8]	C	R101	0.45	0.70	1.04	0.73	0.62	0.58	0.63	0.61	0.41	0.68	0.97	0.68
VAD [14]	C	R50	0.41	0.70	1.05	0.72	0.07	0.17	0.41	0.22	0.36	0.66	0.91	0.64
GenAD [49]	C	R50	0.28	0.49	0.78	0.52	0.08	0.14	0.34	0.19	-	-	-	-
MomAD [33]	C	R50	0.31	0.57	0.91	0.60	0.01	0.05	0.22	0.09	0.30	0.53	0.78	0.54
BridgeAD [46]	C	R50	0.29	0.57	0.92	0.59	0.01	0.05	0.22	0.09	-	-	-	-
DiffusionDrive [25]	C	R50	0.27	0.54	0.90	0.57	0.03	0.05	0.16	**0.08**	-	-	-	-
SparseDrive [34]	C	R50	0.29	0.58	0.96	0.61	0.01	0.05	0.18	**0.08**	0.30	0.57	0.85	0.57
SpaRC-Drive (Ours)	C+R	R50	**0.24**	**0.47**	**0.79**	**0.50**	0.01	0.06	0.20	0.09	**0.27**	**0.47**	**0.70**	**0.48**

Table 1. Comparison on **nuScenes** dataset with **open-loop** metrics. Metric calculation follows VAD [14] and MomAD [33]. C and R denote Camera and Radar. Similar to SparseDrive [34] and MomAD [33], we deactivate the ego status information for a fair comparison (preventing ego status leakage as analyzed in[23]).

This provides the ego instance with rich semantic and geometric information essential for planning-oriented optimization, incorporating both visual context and radar-derived motion cues.

3.5. Probabilitic Trajectory Modeling

On top of the fusion representation, we leverage agent-level interactions via cross-attention, fusing history information of the agents and map elements. Each query, including the ego-instance predicts multi-modal trajectories following the three driving commands: turn left, turn right, and go straight. Each trajectory gets rescored, based on the proximity to other agent's trajectories.

3.6. Loss Design

The final loss function is the average displacement error (ADE) between output and ground truth trajectories of the planned ego vehicle and the forecasted surrounding traffic agents. Focal loss handles the classification of the trajectory modes (lowest ADE corresponds to the positive sample, others as negative samples) and L1 loss supervises the actual trajectory. The queries are regularized by detection and mapping loss through hungarian matching and box/point regression losses. A depth head in the perspective view guides with an additional L1 loss.

The unified architecture enables joint optimization of radar fusion and planning objectives, resulting in improved spatial coherence, temporal consistency, and collision avoidance compared to vision-only baselines.

4. Experiments

4.1. Experimental Setup

For comprehensive evaluation, we evaluate our approach on real-world open-loop benchmarks as well as a closed-loop simulation environment.

nuScenes Open-Loop [1] We evaluate on the standard nuScenes dataset containing 1000 driving scenes of 20 seconds each at 2Hz, captured by six surround-view cameras, one LiDAR and 5 radars, collecting point clouds including RCS and Doppler velocity measurements.

Long-Horizon Turning-nuScenes [33] To better assess the temporal consistency of predicted trajectories, Song et al. introduced a new validation benchmark based on the most challengingturning scenarios within nuScenes validation set.

Bench2Drive [12] The NeurIPS 2024 benchmark is a reactive simulation environment for autonomous driving following a closed-loop evaluation protocol under CARLA Leaderboard 2.0 [10]. We use the official base configuration of 1000 simulated driving scenes, captured by six surround-view cameras and 5 radar sensors collecting sparse point clouds with velocity measurements. The sensor setup closely resembles the vehicle configuration of nuScenes. the dev10 protocol [13], an officially curated subset of of varying weather conditions, locations and traffic densities selected to cover a wide range of difficult driving scenarios with low variance.

Evaluation Metrics We follow the established evaluation protocols for comprehensive assessment across all autonomous driving tasks: 3D Object Detection: Average precision (mAP) and nuScenes Detection Score (NDS), which comprises the weighted sum of mAP and five True Positive metrics: Translation (mATE), Scale (mASE), Orientation (mAOE), Velocity (mAVE), and Attribute Error (mAAE). Multi-Object Tracking: Average Multi-Object Tracking Accuracy (AMOTA) and Average Multi-Object Tracking Precision (AMOTP). Online Mapping: Map segmentation accuracy using mean Average Precision (mAP) for different map elements including pedestrian crossings (AP_{ped}), lane dividers (AP_d), and lane boundaries (AP_b). Motion Prediction:Minimum Average Displacement Error (minADE), minimum Final Displacement Error (minFDE), Miss Rate (MR), and End-to-end Prediction Accuracy (EPA) c[8]. Planning: L2 Displacement Error (L2), Collision Rate and Trajectory Prediction Consistency (TPC) [33]. For all plan-

Method	3D Object Detection							Multi-Object Tracking			Online Mapping				Motion Prediction			
	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓	AMOTA↑	AMOTP↓	Recall↑	mAP↑	$\mathrm{AP}_{\mathrm{ped}}$ ↑	AP_d ↑	AP_b ↑	mADE↓	mFDE↓	MR↓	EPA↑
UniAD [8]	38.0	49.8	0.684	0.277	0.383	0.381	0.192	0.359	1.320	0.467	-	-	-	-	0.71	1.02	0.151	0.456
VAD [14]	31.2	43.5	0.610	0.288	0.541	0.534	0.228	-	-	-	47.6	40.6	51.5	50.6	-	-	-	-
MomAD [33]	42.3	53.1	0.561	0.269	0.549	0.258	0.188	0.391	1.243	0.509	55.9	50.7	**58.1**	58.9	0.61	0.98	0.137	0.499
SparseDrive [34]	41.8	52.5	0.566	0.275	0.552	0.261	0.190	0.386	1.254	0.499	55.1	49.9	57.0	58.4	0.62	0.99	0.136	0.482
SpaRC-Drive (Ours)	**46.6**	**57.0**	**0.512**	**0.271**	**0.494**	**0.173**	**0.177**	**0.469**	**1.129**	**0.553**	**56.9**	**53.7**	55.4	**61.7**	**0.58**	**0.93**	**0.121**	**0.53**

Table 2. Perception and motion results on the **nuScenes** validation dataset. † indicates the results are reproduced with the official checkpoint. AP_d denotes $\mathrm{AP}_{\mathrm{divider}}$. AP_b denotes $\mathrm{AP}_{\mathrm{boundary}}$. mADE denotes minADE. mFDE denotes minFDE.

ning metrics, we are following [33, 34] which follow the official settings introduced by VAD [14]. During reactive closed-loop evaluation, we additionally evaluate the Bench2Drive driving score and the success rate of the planned trajectories.

4.2. Implementation Details

We follow the multi-stage training pipeline of [34]. In the first stage, we train the multi-modal sparse feature encoder and the detection head. Each modality backbone is trained from scratch (ResNet initialized from an ImageNet checkpoint).

Sparc-Drive uses a single configuration of 900 anchors for detection, 100 polylines for mapping, and 6 decoder layers. We employ the AdamW optimizer and Cosine Annealing learning rate scheduler for 100 epochs (similar to [34] and [33]) in stage one and 10 epochs in stage two. Further hyper-parameters will be provided in the accompanying code repository.

The perception range is set to 50m, with an instance memory queue of three key frames, training in a streaming manner [35]. The motion forecasting horizon is set to 12s and the planning prediction to 6s. The vison backbone encompases a ResNet-50 with an input-size of 256x704 on nuScenes and 384x704 on Bench2Drive (same as all compared model configurations). Our models are trained with a batch size of 48.

We deactivate ego status information following SparseDrive conventions [33, 34] to prevent ego status leakage as analyzed in [23], ensuring fair comparison across all methods.

4.3. Main Results

4.3.1. Perception and Motion Forecasting Results

As shown in Tab. 2, SpaRC-Drive achieves significant improvements across all perception tasks compared to the SparseDrive baseline. Our radar fusion framework demonstrates a 4.8% mAP improvement and 4.5 NDS enhancement on the nuScenes validation set. The improvements are particularly pronounced in velocity estimation (mAVE: 0.173 vs 0.261), highlighting radar's effective contribution through Doppler measurements.

Moreover, SpaRC-Drive achieves state-of-the-art tracking performance with 8.3% AMOTA improvement over vision-centric SparseDrive. The enhanced velocity estimation from radar Doppler directly benefits object-level motion modeling, leading to more stable tracking trajectories. Combined with improved precision, our approach demonstrates superior capability in maintaining object identity across frames, critical for planning-oriented autonomous driving systems.

The radar fusion provides also 1.8% mAP improvement in online mapping, with particularly strong gains in lane boundary detection. Finally, SpaRC-Drive achieves a 4.0% reduction in mADE, demonstrating improved motion forecasting accuracy. The integration of radar-derived velocity information enhances multi-agent intent prediction, leading to more accurate trajectory forecasts.

4.3.2. Open-Loop Planning Results

Tab. 1 evaluates the performance of SpaRC-Drive in open-loop planning settings, with the lowest average L2 error (0.50m) compared to SparseDrive (0.61m), UniAD (0.73m), and MomAD (0.60m). Most significantly, we achieve a 9% improvement in Trajectory Prediction Consistency (TPC) compared to SparseDrive, indicating more consistent trajectory prediction.

In summary, SpaRC-Drive achieves state-of-the-art performance on the nuScenes open-loop benchmark, demonstrating the effectiveness of radar fusion in improving perception, tracking, and motion forecasting capabilities. The raw strength in feature representation also outperforms more sophisticated planner like MomAD [33] or DiffusionDrive [25].

4.3.3. Turning Scenarios

When focusing the evaluation on the most complex and challenging scenarios (*cf*. Tab. 3), the difference to vision-based models increases. We are able to significantly improve the L2 (-0.26m) and TPC metrics (-0.15), while mainting the overall low collision rate (-31%) of 0.09, in contrast to SparseDrive. This safety-critical scenario analysis shows the effectiveness of our radar-based approach and emphasizes the importance of multi-modal sensor integration for all autonomous driving designs.

4.3.4. Long Trajectory Prediction.

In Tab. 4, we increase the prediction horizon to 6s and evaluate the performance of SpaRC-Drive in long-term trajectory

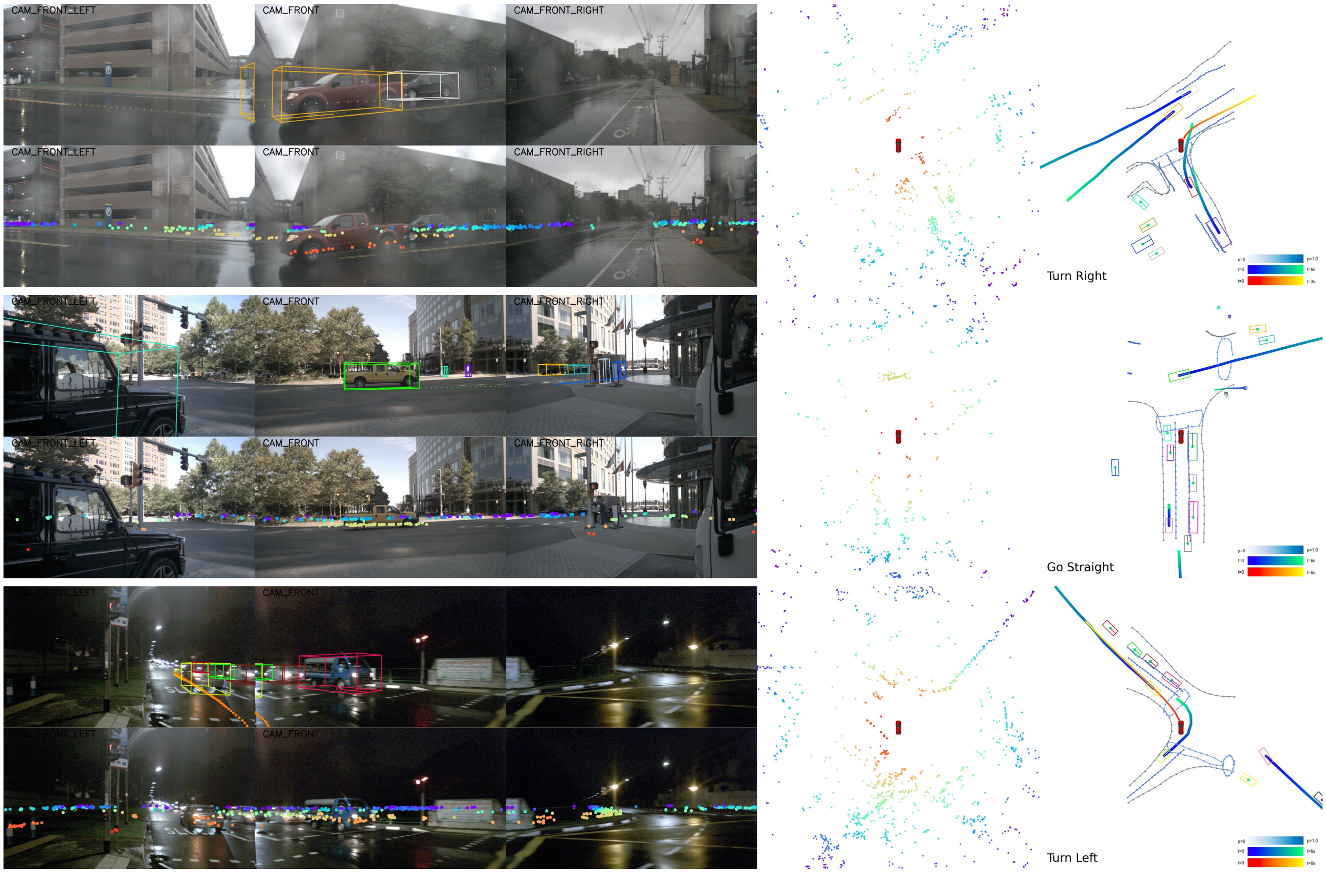

Figure 2. **Qualitative Examples** of produced trajectories and visualized radar points of a challenging turning scenario with a long horizon of six seconds (top: rain, middle: partially hidden objects, bottom: night). On the left, we show the front-facing camera views, predicted bounding boxes, and projected radar points. In the middle, we visualize the perceived radar points in a top-down Bird's-Eye-View at 50m range, color-coded by the distance to the ego vehicle. On the right, the corresponding predicted map elements, bounding boxes with motion forecasts and planned trajectories.

prediction. In both settings, full-set and T-nuScenes, we are able to significantly improve the trajectory consistencies in L2 and TPC, with strongly reduced collision rates. We can show that doubling the prediction horizon and overcoming partial occlusions in highly dynamic scenes shows a major potential for trajectory consistency and collision reduction. In a six second prediction horizon, we can see that the radar-based approach is able to predict more stable trajectories. SpaRC-Drive can capitalize on longer perception ranges, detecting partially occluded objects and better motion modeling.

4.3.5. Closed-Loop Planning Results

In Tab. 5, we generalize the findings of SpaRC-Drive to the closed-loop planning setting of Bench2Drive. Evaluating in open-loop, we again outperform the baseline SparseDrive by a trajectorydisplacement of 0.82 vs 0.87m. Moreover, in interactive scenarios like cut-ins, overtaking maneuvers or emergency brakings, SpaRC-Drive achieves a 20% higher success rate compared to SparseDrive. We will extend the evaluation to the full set of 220 routes of Bench2Drive in the camera-ready version.

4.4. Qualitative Analysis

Furhermore, we visualize the perception and planning performance of our model in challenging scenarios. Fig. 2 shows the perception and planning performance of our model visually in a challenging turning scenario with a long horizon of six seconds. We project the radar points onto the Bird's-Eye-View and front-facign camera views and visualize the predicted map elements, bounding boxes with motion forecasts and planned trajectories.

In Fig. 3, we compare our fusion design with the baseline SparseDrive and indicate, the synergies radar-fusion provides. The qualitative analysis validates that our radar fusion strategy addresses fundamental limitations of vision-centric approaches, particularly in scenarios where precise motion understanding and long-horizon prediction are essential for collision avoidance and safe autonomous driving operation.

Method	**L2 (m) ↓**				**Col. Rate (%) ↓**				**TPC (m) ↓**			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.	1s	2s	3s	Avg.
SparseDrive [34]	0.35	0.77	1.46	0.86	0.04	0.17	0.98	0.40	0.34	0.70	1.33	0.79
SpaRC-Drive (Ours)	**0.26**	**0.54**	**0.93**	**0.58**	**0.00**	**0.04**	**0.23**	**0.09**	**0.35**	**0.63**	**0.95**	**0.64**

Table 3. Planning results on the **Turning-nuScenes** validation dataset. We follow the VAD [14] evaluation metric.

Split	**Method**	**L2 (m) ↓**			**Col. Rate (%) ↓**			**TPC (m) ↓**		
		4s	5s	6s	4s	5s	6s	4s	5s	6s
nuScenes	SparseDrive [34]	1.75	2.32	2.95	0.87	1.54	2.33	1.33	1.66	1.99
	SpaRC-Drive (Ours)	**1.14**	**1.61**	**2.16**	**0.61**	**1.08**	**1.61**	**1.04**	**1.33**	**1.65**
T-nuScenes	SparseDrive [34]	2.07	2.71	3.36	0.91	1.71	2.57	1.54	2.31	2.90
	SpaRC-Drive (Ours)	**1.38**	**1.97**	**2.66**	**0.47**	**0.99**	**1.66**	**1.42**	**1.86**	**2.33**

Table 4. **Long trajectory planning** results on the **nuScenes** and **Turning-nuScenes** validation sets. We train models for 10 epochs for 6s-horizon prediction. We follow the VAD [14] evaluation metric.

Method	**Input**	**Open-loop**	**Closed-loop Metrics**	
		Avg. L2 ↓	Driving Score ↑	Success Rate (%) ↑
SparseDrive* [34]	C	0.87	39.9	10.0
SpaRC-Drive (Ours)	C + R	**0.82**	**55.6**	**30.0**

Table 5. **Open-loop** and **closed-loop** evaluation results on **Bench2Drive** (V0.0.3) using the base training set. We report the closed-loop simulation in the dev10 protocol. * indicates re-implementation and provided model checkpoint of [33].

4.5. Limitations

While our experiments demonstrate the benefits of radar fusion for end-to-end autonomous driving, several limitations remain. First, the radar data in both nuScenes and Bench2Drive provides only sparse point cloud representations, limiting the potential density of radar-based features. The sensing range is also restricted to 50m, which does not fully leverage radar's capability for long-range detection beyond 150m. Additionally, the nuScenes radar setup lacks height information, preventing full 4D radar perception. In the simulation environment of Bench2Drive, the radar sensor placement and extrinsic calibration are suboptimal compared to real-world setups. The simplified radar sensing principles in the CARLA simulator also do not fully capture the complex radar phenomenology of real sensors. To validate the full potential of radar-based perception for autonomous driving, extensive closed-loop testing with real-world radar-camera systems will be required.

4.6. Future Work

As next steps, we will explore more fusion mechanisms and extend the analysis also to dense-BEV based methods. While our current approach operates on pre-processed radar point clouds, future research directions include exploring raw radar tensor representations and investigating larger perception ranges, potentially up to 150m [6]. Additionally, the domain gap between simulated and real-world camera-radar data necessitates dedicated multi-modal planning-oriented datasets. We envision extending this work also to cooperative perception scenarios [43] on radar-camera-based V2X settings, further enhancing the robustness and safety of end-to-end autonomous driving systems [32, 39, 42, 51].

5. Conclusion

Multi-modal fusion, especially radar-based fusion, represents an overlooked yet promising research direction for end-to-end autonomous driving. Radar's unique characteristics—weather immunity, direct velocity measurement through Doppler effects, and long-range detection capabilities beyond 150m—enable significant improvements in scene understanding that are unavailable to vision-only approaches. These capabilities are highly synergistic with the overall planning requirements for safe autonomous driving.

In this paper, we introduce SpaRC-Drive, a novel query-based end-to-end camera-radar fusion framework that extends the sparse representation paradigm to planning-oriented autonomous driving. By integrating adaptive radar fusion strategies into a unified optimization pipeline, our approach addresses fundamental limitations of vision-centric methods, particularly in safety-critical scenarios where accurate motion understanding and long-horizon trajectory prediction are essential for collision avoidance.

References

[1] 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 *CVPR*, 2020. 2, 4

[2] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 2

[3] Shaoyu Chen, Bo Jiang, Hao Gao, Bencheng Liao, Qing Xu, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Vadv2: End-to-end vectorized autonomous driving via probabilistic planning. *arXiv preprint arXiv:2402.13243*, 2024. 2

[4] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE transactions on pattern analysis and machine intelligence*, 45(11):12878–12895, 2022. 1, 2

[5] Xiaomeng Chu, Jiajun Deng, Guoliang You, Yifan Duan, Houqiang Li, and Yanyong Zhang. Racformer: Towards

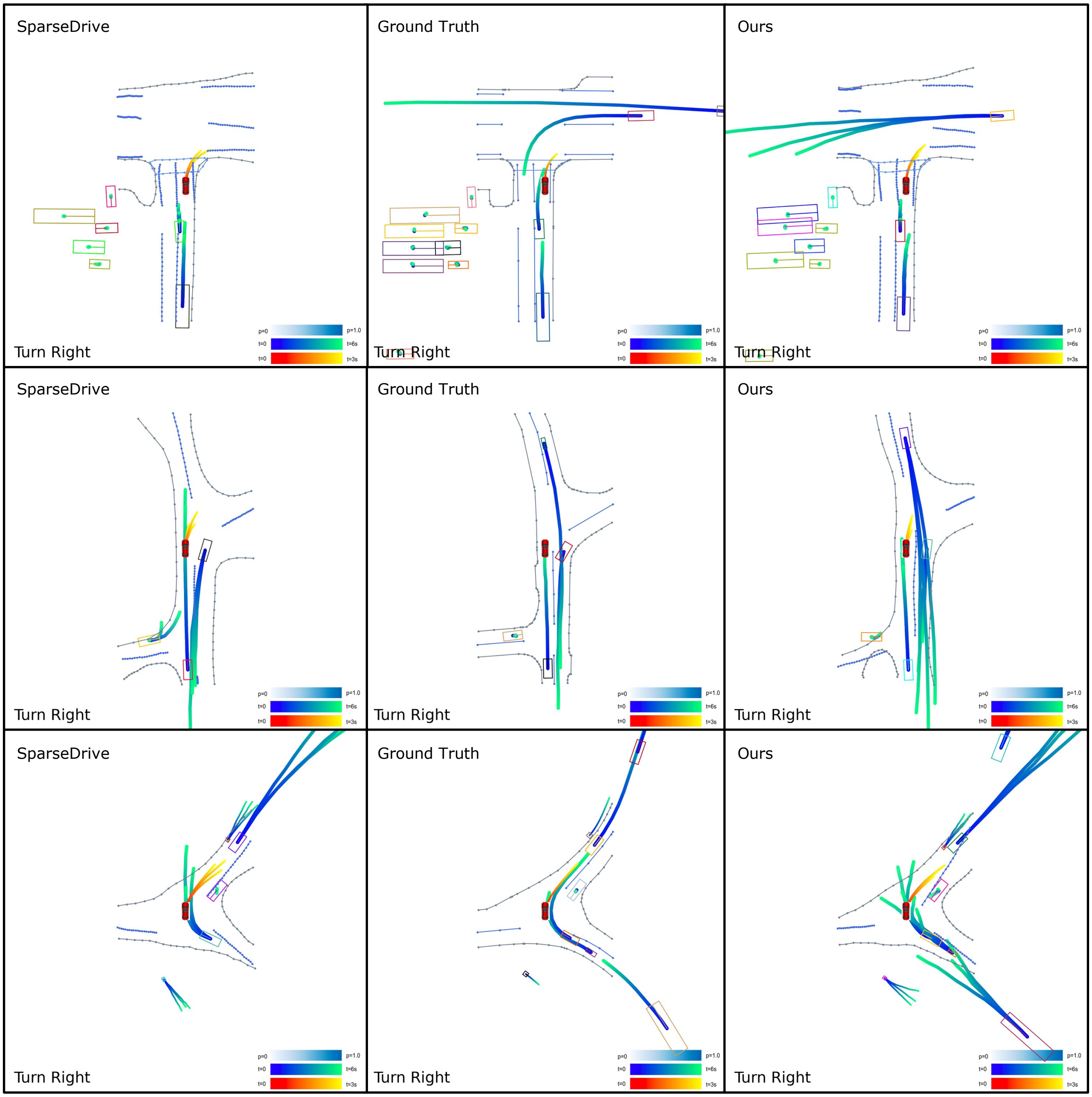

Figure 3. **Qualitative Comparison** of SpaRC-Drive with SparseDrive on challenging turning scenarios in a crowded city environment (top-3 multi-mode trajectories). The first row shows a T-crossing scenario, where SpaRC-Drive successfully detects a oncoming vehicle at high speed. The second row shows a night scenario, where the vison baseline does not detect the oncoming scooter, whereas our approach correctly forecasts the trajectory of the camouflaged vehicle. and the Last scene emphasizes a dynamic turning scenario, where the radar-based approach is able to detect partially occluded vehicles at long range.

high-quality 3d object detection via query-based radar-camera fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17081–17091, 2025. 2, 3

[6] Felix Fent, Fabian Kuttenreich, Florian Ruch, Farija Rizwin, Stefan Juergens, Lorenz Lechermann, Christian Nissler, Andrea Perl, Ulrich Voll, Min Yan, et al. Man truckscenes: A multimodal dataset for autonomous trucking in diverse conditions. *arXiv preprint arXiv:2407.07462*, 2024. 7

[7] Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi

Yan, and Dacheng Tao. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *European Conference on Computer Vision*, pages 533–549. Springer, 2022. 2

[8] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *CVPR*, pages 17853–17862, 2023. 1, 2, 4, 5

[9] Junjie Huang, Guan Huang, Zheng Zhu, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. In *arXiv preprint arXiv:2112.11790*, 2021. 1

[10] Bernhard Jaeger, Kashyap Chitta, and Andreas Geiger. Hidden biases of end-to-end driving models. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2023. 4

[11] Xiaosong Jia, Li Chen, Penghao Wu, Jia Zeng, Junchi Yan, Hongyang Li, and Yu Qiao. Towards capturing the temporal dynamics for trajectory prediction: a coarse-to-fine approach. In *Conference on Robot Learning*, pages 910–920. PMLR, 2023. 2

[12] Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. *NeurIPS*, 2024. 2, 4

[13] Xiaosong Jia, Junqi You, Zhiyuan Zhang, and Junchi Yan. Drivetransformer: Unified transformer for scalable end-to-end autonomous driving. *ICLR*, 2025. 4

[14] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8350, 2023. 2, 4, 5, 7

[15] Jisong Kim, Minjae Seong, Geonho Bang, Dongsuk Kum, and Jun Won Choi. Rcm-fusion: Radar-camera multi-level fusion for 3d object detection. In *ICRA*, pages 18236–18242. IEEE, 2024. 2

[16] Jisong Kim, Minjae Seong, and Jun Won Choi. Crt-fusion: Camera, radar, temporal fusion using motion information for 3d object detection. *Advances in Neural Information Processing Systems*, 37:108625–108648, 2024. 2

[17] Youngseok Kim, Sanmin Kim, Jun Won Choi, and Dongsuk Kum. CRAFT: Camera-Radar 3D Object Detection with Spatio-Contextual Fusion Transformer. In *AAAI*, 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 *ICCV*, 2023. 2

[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 *CVPR*, pages 12697–12705, 2019. 2

[20] Hongyang Li, Chonghao Sima, Jifeng Dai, Wenhai Wang, Lewei Lu, Huijie Wang, Jia Zeng, Zhiqi Li, Jiazhi Yang, Hanming Deng, et al. Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe. *IEEE TPAMI*, 2023. 2

[21] Qifeng Li, Xiaosong Jia, Shaobo Wang, and Junchi Yan. Think2drive: Efficient reinforcement learning by thinking with latent world model for autonomous driving (in carla-v2). In *European Conference on Computer Vision*, pages 142–158. Springer, 2024. 2

[22] Zhenxin Li, Kailin Li, Shihao Wang, Shiyi Lan, Zhiding Yu, Yishen Ji, Zhiqi Li, Ziyue Zhu, Jan Kautz, Zuxuan Wu, et al. Hydra-mdp: End-to-end multimodal planning with multi-target hydra-distillation. *arXiv preprint arXiv:2406.06978*, 2024. 2

[23] Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahan Li, Jan Kautz, Tong Lu, and Jose M Alvarez. Is ego status all you need for open-loop end-to-end autonomous driving? In *CVPR*, pages 14864–14873, 2024. 1, 4, 5

[24] Bencheng Liao, Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Chang Huang. Maptr: Structured modeling and learning for online vectorized hd map construction. *arXiv preprint arXiv:2208.14437*, 2022. 1

[25] Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, et al. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12037–12047, 2025. 4, 5

[26] 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 *CVPR*, pages 14928–14937, 2024. 2

[27] Yicheng Liu, Tianyuan Yuan, Yue Wang, Yilun Wang, and Hang Zhao. Vectormapnet: End-to-end vectorized hd map learning. In *International Conference on Machine Learning*, pages 22352–22369. PMLR, 2023. 1

[28] 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 *ICRA*, 2023. 1, 2

[29] Yunfei Long, Abhinav Kumar, Daniel Morris, Xiaoming Liu, Marcos Castro, and Punarjay Chakravarty. Radiant: Radar-image association network for 3d object detection. In *AAAI*, 2023. 2

[30] Ramin Nabati and Hairong Qi. Centerfusion: Center-based radar and camera fusion for 3d object detection. In *WACV*, pages 1527–1536, 2021.

[31] Felix Nobis, Maximilian Geisslinger, Markus Weber, Johannes Betz, and Markus Lienkamp. A deep learning-based radar and camera sensor fusion architecture for object detection. In *2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*, 2019. 2

[32] Rui Song, Chenwei Liang, Hu Cao, Zhiran Yan, Walter Zimmer, Markus Gross, Andreas Festag, and Alois Knoll. Collaborative semantic occupancy prediction with hybrid feature fusion in connected automated vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17996–18006, 2024. 7

[33] Ziying Song, Caiyan Jia, Lin Liu, Hongyu Pan, Yongchang Zhang, Junming Wang, Xingyu Zhang, Shaoqing Xu, Lei

Yang, and Yadan Luo. Don't shake the wheel: Momentum-aware planning in end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22432–22441, 2025. 1, 4, 5, 7

[34] Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. Sparsedrive: End-to-end autonomous driving via sparse scene representation. *arXiv preprint arXiv:2405.19620*, 2024. 1, 2, 3, 4, 5, 7

[35] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *ICCV*, 2023. 1, 5

[36] Philipp Wolters, Johannes Gilg, Torben Teepe, Fabian Herzog, Felix Fent, and Gerhard Rigoll. Sparc: Sparse radar-camera fusion for 3d object detection. *arXiv preprint arXiv:2411.19860*, 2024. 1, 3

[37] 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. *arXiv preprint arXiv:2403.07746*, 2024. 2

[38] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *CVPR*, pages 4840–4851, 2024. 3

[39] Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, et al. V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13712–13722, 2023. 7

[40] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *CVPR*, 2023. 1

[41] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *CVPR*, pages 11784–11793, 2021. 1

[42] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, et al. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21361–21370, 2022. 7

[43] Haibao Yu, Wenxian Yang, Jiaru Zhong, Zhenwei Yang, Siqi Fan, Ping Luo, and Zaiqing Nie. End-to-end autonomous driving through v2x cooperation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9598–9606, 2025. 7

[44] Tianyuan Yuan, Yicheng Liu, Yue Wang, Yilun Wang, and Hang Zhao. Streammapnet: Streaming mapping network for vectorized online hd map construction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7356–7365, 2024. 1

[45] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In *European conference on computer vision*, pages 659–675. Springer, 2022. 1

[46] Bozhou Zhang, Nan Song, Xin Jin, and Li Zhang. Bridging past and future: End-to-end autonomous driving with historical prediction and planning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6854–6863, 2025. 4

[47] Diankun Zhang, Guoan Wang, Runwen Zhu, Jianbo Zhao, Xiwu Chen, Siyu Zhang, Jiahao Gong, Qibin Zhou, Wenyuan Zhang, Ningzi Wang, et al. Sparsead: Sparse query-centric paradigm for efficient end-to-end autonomous driving. *arXiv preprint arXiv:2404.06892*, 2024. 1

[48] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International journal of computer vision*, 129(11):3069–3087, 2021. 1

[49] Wenzhao Zheng, Ruiqi Song, Xianda Guo, Chenming Zhang, and Long Chen. Genad: Generative end-to-end autonomous driving. In *European Conference on Computer Vision*, pages 87–104. Springer, 2024. 4

[50] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018. 2

[51] Walter Zimmer, Gerhard Arya Wardana, Suren Sritharan, Xingcheng Zhou, Rui Song, and Alois C Knoll. Tumtraf v2x cooperative perception dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22668–22677, 2024. 7