arXiv:2607.09629v1 [cs.CV] 10 Jul 2026

4DR360°: State Reasoning for Joint 3D Detection and Occupancy Prediction in 4D Radar-Camera Full-Scene Perception

Xiaokai Bai[1], Lianqing Zheng[2], Runwei Guan[3], Songkai Wang[1], Siyuan Cao[1], Hui-liang Shen[1]

[1]College of Information Science and Electronic Engineering, Zhejiang University.
[2]School of Automotive Studies, Tongji University.
[3]Thrust of Artificial Intelligence, Hong Kong University of Science and Technology.
shawnnnkb@gmail.com

Abstract

Reliable autonomous driving requires full-scene perception that couples foreground objects with dense semantic layout. Recently, 4D millimeter-wave radar has emerged as a robust and affordable sensor, yet its sparse returns make radar-camera fusion necessary for comprehensive scene understanding. Existing radar-camera methods mainly optimize detection, while dual-task systems usually decode boxes and occupancy with limited interaction. To address this gap and advance radar-based multi-task learning, we propose 4DR360°, a 4D radar-camera framework for 360° full-scene perception, which models semantic occupancy as a persistent scene state rather than a terminal output. 4DR360° follows a cross-modal state reasoning paradigm, where the occupancy state is modeled and propagated through stages for coarse-to-fine feature aggregation. Specifically, State-guided BEV Enhancement (SBE) strengthens intra-frame BEV representation, while Doppler-guided Temporal Fusion (DTF) preserves state evidence over longer temporal horizons. Beyond the model, we further extend ManTruckScenes with satellite-map-based generated occupancy labels and pair it with OmniHD-Scenes in a unified cross-dataset detection-and-occupancy protocol. The resulting experiments cover accuracy, robustness, ablation, and efficiency under one radar-camera multi-task evaluation framework. Code and labels will be released upon acceptance.

Introduction

Autonomous driving requires 360° full-scene perception that jointly models foreground agents and surrounding scene layout. In this setting, 3D object detection estimates compact boxes for traffic participants (Liu et al. 2023; Bai et al. 2024, 2026c; Xia et al. 2026), whereas semantic occupancy prediction recovers dense free, occupied, and semantic voxel states around the ego vehicle (Cao and De Charette 2022; Li et al. 2023; Wei et al. 2023; Tian et al. 2023; Zheng et al. 2024). These two tasks describe the same scene at different granularities. Detection focuses on instance-level localization, while occupancy supplies the spatial support and layout context in which those instances exist. Therefore, reliable full-scene perception requires a unified treatment of object reasoning and scene layout.

Driven by recent datasets and radar-camera perception methods, 4D millimeter-wave radar is becoming an important sensing modality for autonomous driving (Palffy et al. 2022; Zheng et al. 2022; Paek, Kong, and Wijaya 2022; Zheng et al. 2023; Lin et al. 2024; Bai et al. 2024, 2026c; Xia et al. 2026). It provides stable geometric and motion cues under challenging conditions. However, its sparse returns still require camera semantics for complete scene understanding. Recent radar-camera methods therefore improve bird's-eye-view fusion, semantic-geometric interaction, sparse object representation, and temporal modeling (Zheng et al. 2023; Lin et al. 2024; Bai et al. 2024, 2026c; Xia et al. 2026). Nevertheless, most of these advances are either validated on front-view detection benchmarks such as View-of-Delft and TJ4DRadSet (Palffy et al. 2022; Zheng et al. 2022), or remain primarily optimized for box-level perception in surround-view settings. This field gap limits how 4D radar evidence is used for full-scene perception, where object reasoning and scene layout should be evaluated under a unified radar-camera protocol.

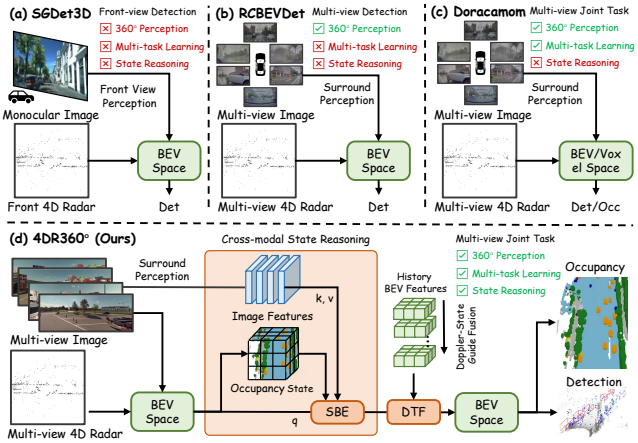

Figure 1. Comparison of representative 4D radar-camera paradigms. (a) SGDet3D is front-view and detection-only, (b) RCBEVDet extends detection to 360° perception, (c) Doracamom supports joint detection and occupancy prediction without explicit state reasoning, and (d) 4DR360° organizes 360° radar-camera multi-task perception through occupancy-state reasoning with SBE and DTF.

Figure 1 further shows how representative 4D radar-camera systems have progressively expanded this scope,

Copyright © 2026. All rights reserved.

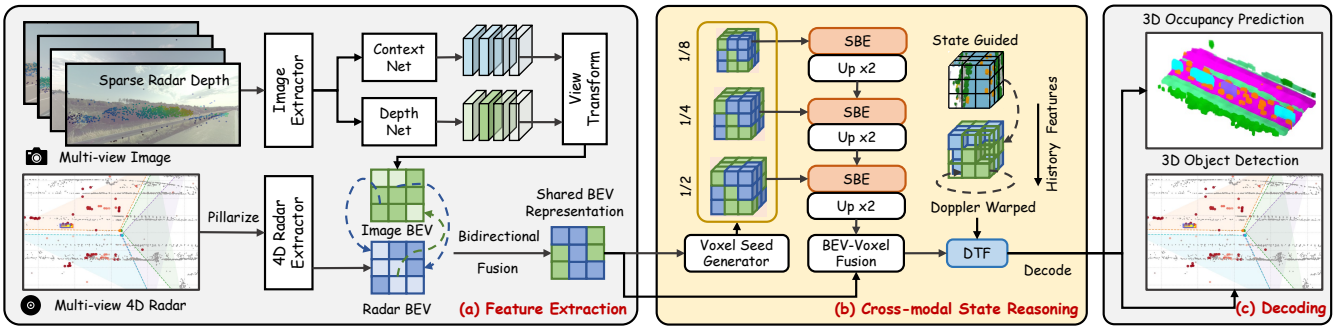

Figure 2. Overall framework of 4DR360°. The pipeline first extracts multi-view image BEV features and 4D radar BEV features, fuses them into a shared BEV representation, reasons over coarse-to-fine occupancy states with SBE and DTF, and decodes joint 3D detection and semantic occupancy outputs.

while a key limitation remains. SGDet3D (Bai et al. 2024) focuses on front-view detection. RCBEVDet (Lin et al. 2024) extends radar-camera detection to 360° perception. Doracamom (Zheng et al. 2026) further moves to joint detection and occupancy prediction. However, the representation form remains close to a shared BEV feature followed by a detection head and an occupancy branch.

Yet introducing occupancy only after the shared BEV representation is formed makes the bottleneck more than task coverage. What remains missing is a stateful representation that lets occupancy shape object reasoning before final decoding. Multi-task perception offers a relevant perspective: SOGDet uses occupancy context for detection, and UniVision studies a unified formulation for joint 3D perception (Zhou et al. 2024; Hong et al. 2024). However, neither method is designed for 4D radar-camera multimodality. Existing joint radar-camera systems usually attach a detection head and an occupancy branch to a shared BEV feature (Liu et al. 2023; Zheng et al. 2026), leaving occupancy as a mostly terminal branch with limited influence on temporal representation and object reasoning. Evaluation is also concentrated on OmniHD-Scenes, so cross-dataset generality remains insufficiently examined. These limitations motivate a representation-level design that introduces occupancy as an intermediate scene state for 4D radar-camera multi-task perception. This state links radar-supported foreground geometry and motion cues with continuous scene layout, allowing dense occupancy evidence and temporal cues to enter the shared representation before final prediction.

To realize this idea, we propose 4DR360°, a 4D radar-camera framework for 360° full-scene perception. 4DR360° organizes the shared BEV representation through cross-modal state reasoning. First, it converts camera appearance and radar geometry into occupancy-aware state features before final decoding. It then applies State-guided BEV Enhancement (SBE) to refine the current BEV representation through state-guided deformable cross attention. Finally, Doppler-guided Temporal Fusion (DTF) uses ego alignment and motion-sensitive radar evidence to preserve reliable state history across frames. Through this design, occupancy state strengthens object-level features without replacing the detector representation. Beyond the model, we also extend the benchmark setting beyond OmniHD-Scenes, the main multiview 4D radar-camera dataset with joint detection and occupancy labels. Specifically, we construct occupancy annotations for ManTruckScenes from satellite-map priors and build a unified evaluation protocol across the two datasets. Together, these model and benchmark components lead to four contributions.

- We identify semantic occupancy as a persistent scene state for 4D radar-camera perception that unifies layout, temporal evidence, and object-level reasoning.
- We propose 4DR360°, a state-centered 4D radar-camera framework that performs cross-modal occupancy-state reasoning through SBE and DTF for feature refinement and temporal state modeling.
- We extend ManTruckScenes with satellite-map-based occupancy labels and establish a unified multi-task protocol for advanced 4D radar-camera perception.
- We define comprehensive experiments on accuracy, robustness, and efficiency, establishing a unified protocol and framework for dual-task 4D radar-camera perception and future radar-aware end-to-end or VLA research.

Related Work

4D Radar-Camera 3D Object Detection

4D radar-camera detection uses radar geometry and motion cues to strengthen camera perception under sparse or ambiguous observations. Benchmarks such as View-of-Delft, TJ4DRadSet, and K-Radar support this line of work (Palffy et al. 2022; Zheng et al. 2022; Paek, Kong, and Wijaya 2022). Early methods, including CenterFusion, CRAFT, and RADIANT, attach radar evidence to image proposals or object centers (Nabati and Qi 2021; Kim et al. 2023; Long et al. 2023). Later BEV methods, including RCFusion, RCBEVDet, LXL, SGDet3D, and HGSFusion, lift both modalities into shared BEV features for scene-level fusion (Zheng et al. 2023; Lin et al. 2024; Xiong et al. 2024; Bai et al. 2024; Gu et al. 2025). RaGS and R4Det further strengthen sparse representation, depth reasoning, and temporal modeling (Bai et al. 2026c; Xia et al.

2026). Recent extensions also explore local-global radar detection, sparse-to-dense radar learning, raw radar-tensor fusion, radar-camera depth estimation, cross-view instance awareness, and collaborative perception (Bai et al. 2025b,a; Guan et al. 2025; Zhang et al. 2025; Bai et al. 2026a,b).

These methods establish strong radar-camera detection baselines and mark the progression from front-view detection to 360° perception. However, their objective remains box-centric: radar evidence is fused for detection, whereas dense scene structure is rarely maintained as a reusable state for joint object and occupancy reasoning.

Multi-Task Learning for Perception

Multi-task perception seeks shared representations that support multiple driving outputs, with semantic occupancy providing a dense description of scene layout. Camera BEV and occupancy baselines such as MonoScene, VoxFormer, SurroundOcc, PanoOcc, and M-CONet from OpenOccupancy recover voxelized layout from multi-view imagery (Cao and De Charette 2022; Li et al. 2023; Wei et al. 2023; Wang et al. 2024, 2023; Huang et al. 2024). SOGDet and UniVision further show that occupancy context can support object reasoning (Zhou et al. 2024; Hong et al. 2024). These methods motivate joint perception, but they do not target 4D radar-camera perception, where sparse, motion-aware radar returns are naturally sensitive to foreground objects.

Benchmark support for 4D radar-camera multi-task perception is much narrower. OmniHD-Scenes is currently the only public 4D radar-camera benchmark with both 3D detection and semantic occupancy annotations (Zheng et al. 2024), and Doracamom provides a joint reference on this setting (Zheng et al. 2026). This leaves cross-dataset generality underexplored and ties future radar-aware end-to-end studies to one benchmark contract. We therefore extend ManTruckScenes with auditable occupancy labels and evaluate both datasets under one normalized radar-camera detection-and-occupancy protocol. Methodologically, we further differ from ordinary dual-head systems by propagating occupancy as an explicit scene state, together with Doppler motion cues, inside the shared representation rather than decoding it only at the output side.

Method

Overview

4DR360° formulates 4D radar-camera perception as cross-modal state reasoning, as summarized in Fig. 2. Given synchronized multi-view images and 4D radar sweeps at time t, Fig. 2 (a) constructs an image BEV through view transformation and fuses it with radar BEV using bidirectional deformable attention. The fused radar-camera BEV is then lifted into a coarse-to-fine voxel hierarchy, which carries the scene states refined in Fig. 2 (b). SBE injects current-frame occupancy structure, while DTF propagates temporally aligned state evidence with radar-motion cues. The detection and occupancy heads then read out from the shared state-refined representation, as shown in Fig. 2 (c). Implementation details are provided in the supplementary materials.

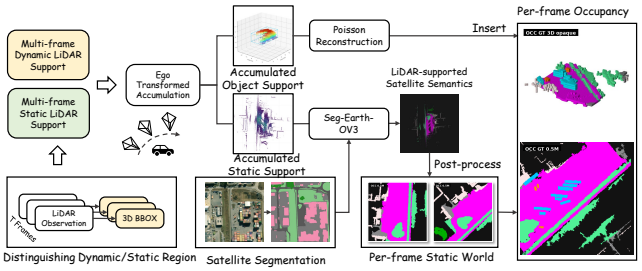

Figure 3. ManTruckScenes occupancy construction. LiDAR anchors static geometry, satellite priors provide conservative static semantics, and object-local replay inserts dynamic surfaces, which together enable auditable cross-dataset occupancy evaluation.

ManTruckScenes Occupancy Construction

A single dataset provides limited evidence for full-scene radar-camera evaluation. OmniHD-Scenes provides native multiview 4D radar-camera detection and occupancy labels, but it is currently the only benchmark of this type. ManTruckScenes offers a complementary commercial-vehicle platform, yet it contains detection annotations only. Evaluating it only as a detector benchmark would leave the occupancy part of radar-camera full-scene perception untested across datasets. We therefore extend ManTruckScenes with an auditable occupancy export.

As illustrated in Fig. 3, we first aggregate multi-frame LiDAR observations into a static support world, so occupied geometry is anchored by measured surfaces. Satellite segmentation is then used only for conservative static semantics around supported regions, while dynamic actors are inserted by object-local surface replay instead of dense box filling. The resulting labels are used only as training and evaluation targets, with fixed label mapping, voxel range, masks, temporal metadata, and metrics within each dataset. The supplementary material provides the full generation and quality-control details.

Feature Extractor

The feature extractor in Fig. 2 (a) builds the radar-camera BEV evidence used to initialize the state. We omit the batch dimension. Given synchronized images from N cameras, a ResNet-FPN encoder extracts $\mathbf{F}_t^{2D} \in \mathbb{R}^{N\times C_i\times H_i\times W_i}$. A depth-aware view transformer produces the camera BEV feature, depth distribution, and image-view context:

$$(\mathbf{B}_t^c, \mathbf{D}_t, \mathbf{F}_t^{ctx}) = \mathcal{P}_{view}(\mathbf{F}_t^{2D}, \Pi_t), \tag{1}$$

where Π_t collects camera calibration and image augmentations. The outputs are image BEV $\mathbf{B}_t^c \in \mathbb{R}^{C\times H\times W}$, depth probabilities $\mathbf{D}_t \in \mathbb{R}^{N\times D\times H_i\times W_i}$, and multi-view context $\mathbf{F}_t^{ctx} \in \mathbb{R}^{N\times C\times H_i\times W_i}$. Following (Bai et al. 2024), we also use a geometric projection of $\mathbf{F}_t^{ctx}$ to strengthen the camera BEV.

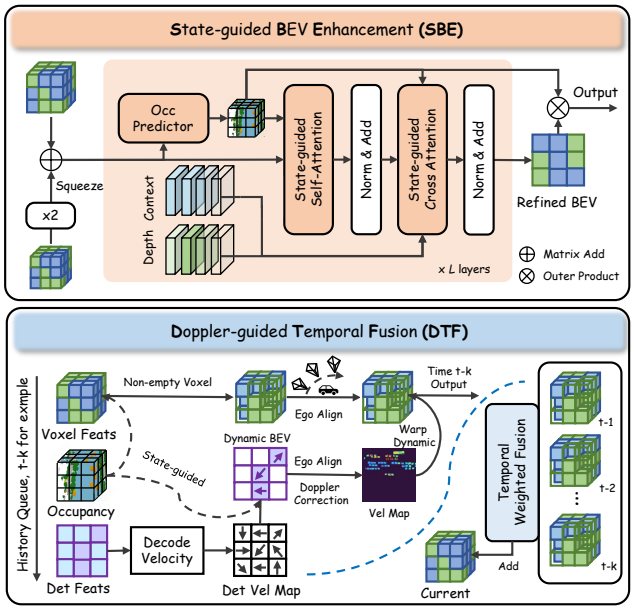

Figure 4. State reasoning modules. SBE predicts occupancy state before attention and uses non-empty confidence with depth-aware image context to refine features. DTF decodes velocity from detection features, uses occupancy confidence as non-empty support, applies Doppler correction, and performs dynamic warping and temporal weighted fusion.

The radar branch maps the synchronized 4D radar point set $\mathcal{R}_t = \{(x, y, z, \mathbf{a})\}$ to $\mathbf{B}_t^r \in \mathbb{R}^{C_r \times H \times W}$ with a pillarization encoder, where $\mathbf{a}$ contains Doppler and other radar attributes. The pillarization details are provided in the supplementary materials. Image and radar BEV features are then fused by bidirectional deformable BEV attention following (Lin et al. 2024), producing $\mathbf{B}_t^{rc}$. This shared BEV representation aligns camera layout cues with radar foreground evidence.

Finally, each pyramid level of $\mathbf{B}_t^{rc}$ is channel-adapted and lifted along height by a learned height MLP:

$$\mathbf{V}_t^{rc,s} = \mathrm{Reshape}\left(\mathcal{H}_s(\mathrm{Down}_s(\mathbf{B}_t^{rc}))\right), \tag{2}$$

where $\mathbf{V}_t^{rc,s} \in \mathbb{R}^{C_s \times Z_s \times H_s \times W_s}$. This produces multi-scale fused voxel features $\{\mathbf{V}_t^{rc,s}\}_{s=0}^{S}$, which provide the substrate for the following state reasoning modules.

Cross-modal State Representation

At stage s, 4DR360° represents the scene with a voxel feature $\mathbf{V}_t^{rc,s}$ and an occupancy-logit state $\mathbf{O}_t^s \in \mathbb{R}^{H_s \times W_s \times Z_s \times C_{occ}}$. The empty channel defines a non-empty confidence map $\mathcal{M}(\mathbf{O}) = 1 - \mathrm{softmax}(\mathbf{O})_{[\ldots,e]}$, where e denotes the empty class and $\mathbf{O}$ always denotes logits. $\mathcal{M}(\mathbf{O})$ is broadcast over channels on voxel features and height-pooled when applied to BEV features. This operator is the interface between occupancy prediction and feature reasoning, so the state is not a terminal side output. It is predicted before attention, resized across stages as a semantic prior, and converted back into voxel features for both final heads. We summarize the stage transition as

$$(\mathbf{O}_t^s, \widetilde{\mathbf{V}}_t^{rc,s}) = \Phi_{state}^s(\mathbf{V}_t^{rc,s}, \mathbf{F}_t^{ctx}, \mathbf{D}_t, \Pi_t, \mathbf{O}_t^{s-1}), \tag{3}$$

where Φ_{state}^s includes state prediction, SBE, and occupancy-aware voxel recovery. SBE internally produces the stage BEV feature $\mathbf{B}_t^{rc,s}$, which is lifted back into the recovered voxel state. The recovered $\widetilde{\mathbf{V}}_t^{rc,s}$ is upsampled as residual state evidence for the next finer scale.

State-guided BEV Enhancement (SBE)

Given the state interface above, SBE converts the predicted occupancy state into current-frame spatial guidance for radar-camera scene understanding. It is the stage-wise spatial decoder in the upper panel of Fig. 4. At stage s, it takes the voxel seed $\mathbf{V}_t^{rc,s}$, image-view context $\mathbf{F}_t^{ctx}$, depth distribution $\mathbf{D}_t$, camera geometry Π_t, and the previous coarser state $\mathbf{O}_t^{s-1}$ when available. An occupancy predictor first estimates $\mathbf{O}_t^s$, from which $\mathcal{M}(\mathbf{O}_t^s)$ is obtained.

The predicted state guides two attention operations. State-guided self-attention uses non-empty confidence to bias BEV query offsets and weights, so foreground and uncertain occupied regions receive more modeling capacity. State-guided cross-attention then projects height anchors to camera views and uses depth consistency and non-empty confidence to weight the sampled context. For a BEV location (x, y), let $\mathcal{X}_{x,y}$ be its height anchors. The image value sampled at anchor $\mathbf{x}$ and view j is

$$\mathbf{g}_{j\mathbf{x}} = \mathbf{W}_v \mathtt{Bilinear}\big(\mathbf{F}_{t,j}^{ctx}, \mathcal{P}(\mathbf{x}, \Pi_{t,j}) + \Delta\mathbf{u}_{j\mathbf{x}}\big), \tag{4}$$

and the state-guided aggregation is

$$\bar{\mathbf{b}}_t^{rc,s}(x, y) = \sum_{j=1}^{N} \sum_{\mathbf{x} \in \mathcal{X}_{x,y}} \alpha_{j\mathbf{x}} d_{j\mathbf{x}} \mathcal{M}(\mathbf{O}_t^s)(\mathbf{x})\, \mathbf{g}_{j\mathbf{x}}, \tag{5}$$

where $\alpha_{j\mathbf{x}}$ is the deformable attention weight, $d_{j\mathbf{x}}$ is obtained by interpolating $\mathbf{D}_{t,j}$ at the projected pixel and anchor-depth bin, $\Delta\mathbf{u}_{j\mathbf{x}}$ is the learned image-plane offset, and $\mathcal{M}(\mathbf{O}_t^s)(\mathbf{x})$ provides a continuous non-empty confidence rather than a hard occupied-anchor mask.

The aggregated response is added to the stage BEV query and refined by a lightweight convolutional block, yielding $\mathbf{B}_t^{rc,s}$. The BEV feature is then lifted back to the voxel lattice. Let $\mathcal{L}_s(\mathbf{B}_t^{rc,s})$ denote the height-lifted stage BEV feature. The recovered voxel state is

$$\widetilde{\mathbf{V}}_t^{rc,s} = \mathbf{V}_t^{rc,s} + \mathcal{L}_s(\mathbf{B}_t^{rc,s}) \odot \mathcal{M}(\mathbf{O}_t^s), \tag{6}$$

where the second term gives explicit non-empty weighting. The recovered voxel feature is upsampled and passed to the next finer stage. Multi-scale occupancy supervision is applied to these stage states, while detailed SBE variants are reported in the supplementary materials.

Doppler-guided Temporal Fusion (DTF)

DTF is the temporal interface of the final voxel state, as shown in the lower panel of Fig. 4. The lower panel implements a compact motion-conditioning path: detection features decode a BEV velocity map, the occupancy state indexes non-empty dynamic support, and radar Doppler

Methods	Mod.	mAP↑	ODS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	Car↑	Ped.↑	Rider↑	LVeh.↑	FPS↑
PointPillars (CVPR 2019)	R	23.82	37.21	0.6752	0.2447	0.3776	0.6789	52.74	0.69	28.57	13.29	**62.2**
RadarPillarNet (IEEE TIM 2023)	R	24.88	37.81	0.6597	0.2389	0.3736	0.6982	52.99	2.06	29.45	15.02	60.3
BEVFormer (ECCV 2022)	C	29.17	30.54	1.1046	0.2346	0.4889	1.0797	53.64	14.48	33.55	15.01	11.4
PanoOcc (CVPR 2024)	C	29.17	28.55	1.1500	0.2446	0.6378	1.6066	51.58	15.82	35.02	14.26	5.5
BEVFusion (NeurIPS 2022)	C&R	33.95	42.62	0.5730	0.2465	0.3814	0.7474	56.25	11.66	50.90	16.99	3.6
RCFusion (IEEE TIM 2023)	C&R	34.88	40.65	0.5676	0.2535	0.4011	0.9208	57.17	12.87	51.35	18.11	3.6
RCBEVDet (CVPR 2024)	C&R	35.53	45.04	0.5138	**0.2305**	0.3914	0.6825	62.35	10.11	54.60	15.06	5.2
SGDet3D (RAL 2025)	C&R	41.73	47.70	0.5430	0.2369	0.3885	0.6849	63.30	19.00	58.30	26.30	3.4
RaGS (CVPR 2026)	C&R	35.88	43.45	–	–	–	–	–	–	–	–	–
Doracamom-S (TCSVT 2026)	C&R	37.60	41.31	0.6724	0.2329	0.4359	0.8579	58.94	17.84	52.72	20.89	4.8
Doracamom (TCSVT 2026)	C&R	39.12	46.22	0.6646	0.2331	**0.3545**	0.6151	61.12	19.83	53.35	22.18	4.2
4DR360° **(ours)**	C&R	**45.05**	**51.40**	**0.4705**	0.2423	0.3762	**0.6013**	**65.47**	**21.73**	**60.41**	**32.58**	4.5

Table 1. OmniHD-Scenes 3D detection results.

Methods	Mod.	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	Car↑	LVeh.↑	Trailer↑	Obs.↑	FPS↑
RadarPillarNet (IEEE TIM 2023)	R	26.61	37.73	0.5243	0.2413	0.1689	2.1915	44.43	30.66	28.65	2.68	**64.2**
LGDD (IROS 2025)	R	29.83	40.49	0.4922	0.2328	0.1229	1.5775	48.90	36.20	32.40	1.80	20.1
BEVFormer (ECCV 2022)	C	28.30	35.47	0.7662	0.2583	0.1982	1.2747	39.56	28.10	19.66	25.87	10.7
PanoOcc (CVPR 2024)	C	28.82	35.72	0.7686	0.2653	0.1920	1.2623	40.12	28.22	21.73	25.20	6.3
BEVFusion (NeurIPS 2022)	C&R	30.97	40.65	0.5168	0.2486	0.1249	2.1377	47.24	36.09	30.15	10.40	4.6
RCFusion (IEEE TIM 2023)	C&R	31.61	40.57	0.5652	0.2483	**0.1155**	2.0670	48.51	36.84	28.58	12.49	5.4
LXL (IEEE TIV 2024)	C&R	34.70	43.04	0.4972	0.2470	0.1174	1.9877	51.95	41.14	31.41	14.30	5.2
RCBEVDet (CVPR 2024)	C&R	41.72	46.85	0.4500	0.2658	0.1543	1.0407	**61.93**	38.00	34.38	32.58	6.5
HGSFusion (AAAI 2025)	C&R	43.90	48.23	0.4607	0.2272	0.1668	1.9479	58.76	43.64	35.43	37.77	5.1
SGDet3D (RAL 2025)	C&R	41.37	47.50	0.4252	**0.2159**	0.1523	1.6112	60.82	**47.38**	26.94	30.29	3.0
Doracamom (TCSVT 2026)	C&R	41.98	46.36	0.4874	0.2411	0.1982	1.5566	61.62	42.71	34.45	29.15	5.2
4DR360° **(ours)**	C&R	**49.57**	**52.97**	**0.3613**	0.2485	0.1571	**0.9442**	59.41	44.71	**41.20**	**52.97**	5.5

Table 2. ManTruckScenes 3D detection results.

corrects the motion cue before dynamic warping. Let $\mathbf{V}_t$ be the final occupancy-state feature and $\mathbf{O}_t$ be its occupancy logits. We use $\mathcal{I}(\mathbf{O}, \cdot)$ to denote state-indexed feature selection, which suppresses empty or irrelevant support before temporal retrieval. For a history horizon T_h, DTF caches $\mathcal{H}_t = \{(\mathbf{V}_{t-k}, \mathbf{O}_{t-k}, \mathbf{T}_{t-k}, \mathbf{U}_{t-k})\}_{k=1}^{T_h}$, where $\mathbf{T}_{t-k}$ is the ego pose and $\mathbf{U}_{t-k}$ is the BEV motion field derived from box velocity and radar Doppler cues. Its input-output contract is

$$\widetilde{\mathbf{V}}_t = \mathcal{F}_{DTF}\big(\mathbf{V}_t, \mathcal{H}_t\big) \tag{7}$$

where DTF aligns historical state features and motion fields to the current ego frame.

DTF performs state-aware history retrieval in three steps. First, the state index $\mathcal{I}(\mathbf{O}_{t-k}, \cdot)$ provides the non-empty voxel path in Fig. 4, and the selected historical features are ego-aligned:

$$\widehat{\mathbf{V}}_{t,k} = \mathcal{T}_{t-k\to t}(\mathcal{I}(\mathbf{O}_{t-k}, \mathbf{V}_{t-k})). \tag{8}$$

Second, the Doppler-corrected velocity map drives dynamic warping after ego alignment. The motion field $\mathbf{U}_{t-k}$ follows the lower path in Fig. 4: detection features decode velocity, occupancy confidence selects dynamic foreground support, and radar Doppler corrects the motion cue. With interval Δt_k, the displacement field is

$$\Delta\mathbf{p}_{t,k} = \mathrm{clip}\left(\Delta t_k \mathcal{T}_{t-k\to t}(\mathcal{I}(\mathbf{O}_{t-k}, \mathbf{U}_{t-k})), \tau\right), \tag{9}$$

where τ limits unrealistically large shifts. The ego-aligned state is sampled from the source location that moves into each current BEV cell:

$$\bar{\mathbf{V}}_{t,k} = \mathcal{S}(\widehat{\mathbf{V}}_{t,k}, -\Delta\mathbf{p}_{t,k}), \tag{10}$$

where $\mathcal{S}$ denotes BEV-plane bilinear sampling shared by all height bins. Indexing removes unsupported regions, but the reliability of the remaining memory still varies across voxels. We therefore derive $\bar{\mathbf{C}}_{t,k}$ by applying the same ego alignment and dynamic warp to $\mathcal{M}(\mathbf{O}_{t-k})$. Third, temporal fusion averages the warped memories with occupancy confidence and temporal decay:

$$\mathbf{M}_t = \frac{\sum_{k=1}^{T_h} \lambda^k \bar{\mathbf{C}}_{t,k} \odot \bar{\mathbf{V}}_{t,k}}{\sum_{k=1}^{T_h} \lambda^k \bar{\mathbf{C}}_{t,k} + \epsilon}. \tag{11}$$

The current state is then refined by a lightweight 3D convolutional update, $\widetilde{\mathbf{V}}_t = \mathbf{V}_t + \mathtt{Conv}_{3D}\big(\mathbf{M}_t\big)$. The same Doppler-aligned update is height-pooled and adapted to the detection BEV, so detection and occupancy receive a shared radar-guided temporal state. Compared with generic BEV temporal aggregation, DTF separates three measurable factors: temporal horizon, state-aware memory retrieval, and radar-conditioned motion compensation.

Head and Training Objective

The final heads read out the same refined state with task-specific adapters, corresponding to Fig. 2(c). Detection projects the voxel state to full-resolution BEV and adds it to the adapted shared radar-camera BEV before CenterHead decoding. Occupancy decodes the voxel state to the full

Methods	Mod.	SC IoU	mIoU	Car	Ped.	LVeh.	Cyc.	Obs.	Fence	Drive	Veg.	Man.	Rider	Side.
BEVFormer (ECCV 2022)	C	28.42	16.23	22.73	5.45	18.21	3.09	3.87	21.54	48.15	17.77	5.48	14.70	17.58
PanoOcc (CVPR 2024)	C	26.36	15.20	22.42	5.91	17.98	3.11	3.36	21.46	50.47	11.20	1.80	13.58	15.90
BEVFusion (NeurIPS 2022)	C&R	27.02	16.24	27.02	4.78	21.59	1.55	2.78	25.21	44.35	13.06	4.25	21.71	12.32
M-CONet (ICCV 2023)	C&R	27.74	16.08	25.21	3.42	21.46	0.88	0.58	29.88	34.48	19.57	8.98	17.53	14.89
SGDet3D (RAL 2025)	C&R	31.15	21.19	28.74	10.35	21.77	5.92	9.02	36.13	47.90	20.72	12.12	24.45	15.96
Doracamom-S (TCSVT 2026)	C&R	31.46	19.49	30.10	6.71	24.31	2.85	6.55	25.77	49.72	19.57	8.72	23.60	16.53
Doracamom (TCSVT 2026)	C&R	33.96	21.81	30.81	7.22	24.70	4.49	7.84	34.49	52.00	21.68	11.49	24.33	20.86
4DR360° **(ours)**	C&R	**35.05**	**24.71**	**33.71**	**11.29**	**27.41**	**6.50**	**10.19**	**41.98**	**52.52**	**22.38**	**14.52**	**29.75**	**21.54**

Table 3. OmniHD-Scenes semantic occupancy results.

Methods	Mod.	SC IoU	mIoU	Car	LVeh.	Trailer	EgoTr.	Obs.	Drive	O.Flat	Terr.	Veg.	Man.
BEVFormer (ECCV 2022)	C	21.04	23.35	27.74	29.06	36.22	76.22	10.76	22.41	10.65	12.37	5.13	2.94
PanoOcc (CVPR 2024)	C	22.75	24.43	28.68	29.51	37.31	76.21	11.16	25.26	12.32	13.31	6.21	4.33
BEVFusion (NeurIPS 2022)	C&R	24.74	24.69	31.68	31.68	38.89	75.37	5.84	24.30	12.95	15.14	9.59	1.46
RCFusion (IEEE TIM 2023)	C&R	21.91	26.02	35.95	35.11	47.18	75.47	6.26	19.24	11.99	14.97	9.54	4.44
LXL (IEEE TIV 2024)	C&R	22.73	27.92	37.18	38.14	47.64	76.39	9.87	25.78	12.61	18.22	10.16	3.21
RCBEVDet (CVPR 2024)	C&R	24.02	28.69	40.39	37.75	48.94	76.62	18.99	22.61	13.08	14.74	9.57	4.21
HGSFusion (AAAI 2025)	C&R	23.87	29.36	42.10	40.86	51.62	76.69	17.30	21.32	12.09	15.81	10.49	5.30
SGDet3D (RAL 2025)	C&R	24.48	28.74	36.63	42.22	50.26	66.00	15.87	27.34	13.14	18.42	11.36	**6.13**
Doracamom (TCSVT 2026)	C&R	24.93	28.65	42.59	37.58	50.45	76.56	3.28	**28.09**	13.50	18.05	11.55	4.85
4DR360° **(ours)**	C&R	**25.07**	**31.65**	**43.61**	**43.56**	**53.90**	**77.65**	**25.58**	24.41	**13.85**	**18.75**	**11.86**	3.33

Table 4. ManTruckScenes semantic occupancy results.

Method	Mod.	mAP↑	ODS↑	mIoU↑
BEVFormer-S (ECCV 2022)	C	28.11	29.36	15.03
BEVFormer (ECCV 2022)	C	30.39	31.61	15.84
PanoOcc (CVPR 2024)	C	26.09	27.16	14.02
BEVFusion (NeurIPS 2022)	C&R	35.83	44.95	15.36
M-CONet (ICCV 2023)	C&R	–	–	15.30
RCBEVDet (CVPR 2024)	C&R	37.49	47.32	–
Doracamom-S (TCSVT 2026)	C&R	38.75	43.47	18.81
Doracamom (TCSVT 2026)	C&R	41.86	48.74	20.30
4DR360° **(ours)**	C&R	**44.18**	**51.57**	**23.17**

Table 5. OmniHD-Scenes adverse-condition robustness.

voxel grid, adds a height-lifted radar-camera BEV prior, and uses the final stage occupancy logits as a soft semantic prior before 3D classification. Further implementation details are provided in the supplementary materials.

Training supervises both terminal predictions and intermediate state. We therefore use a shared occupancy loss for the state branch and the final occupancy branch. At stage s, the occupancy loss is $\mathcal{L}^s_{occ} = \mathcal{L}^s_{focal} + \mathcal{L}^s_{geo} + \mathcal{L}^s_{sem}$, where the three terms denote weighted focal loss, geometric scaling loss, and semantic scaling loss. The overall objective is formulated as

$$\mathcal{L} = \mathcal{L}_{det} + \lambda_{depth}\mathcal{L}_{depth} + \lambda\mathcal{L}^S_{occ} + \sum_i w_i\mathcal{L}^i_{occ}. \quad (12)$$

$\mathcal{L}_{det}$ supervises the detection branch, $\mathcal{L}^S_{occ}$ supervises the final occupancy output, and the weighted intermediate losses supervise state predictions at each stage. $\mathcal{L}_{depth}$ is used when depth supervision is available.

Experiments

Experimental Setup

Datasets and Metrics. We evaluate 4DR360° on multi-view OmniHD-Scenes and ManTruckScenes. OmniHD-Scenes uses six cameras and six 4D radars and provides 4-class detection labels with 11-class semantic occupancy. ManTruckScenes provides a complementary commercial-vehicle setting with the generated 10-class occupied semantic export under four cameras setting. The masks, ignored regions, spatial range, and label mapping are fixed before training and shared by all methods. Detection is reported with mAP, dataset-specific overall score (ODS for OmniHD-Scenes and NDS for ManTruckScenes), mATE, mASE, mAOE, mAVE, per-class AP, and FPS when available. Occupancy is reported with SC IoU, mIoU, and per-class IoU. We also evaluate night and rainy OmniHD-Scenes subsets.

Implementation Details. All methods are adapted under one MMDetection3D data contract with synchronized radar-camera inputs, temporal metadata, detection boxes, occupancy labels, masks, and voxel ranges. OmniHD-Scenes uses a $160 \times 240 \times 16$ grid, while ManTruckScenes uses a $216 \times 216 \times 16$ grid. Training uses AdamW. All ablations are built on the adapted RCBEVDet baseline.

3D Object Detection Results

Tables 1 and 2 show that the detection gain is not limited to one dataset. On OmniHD-Scenes, 4DR360° improves over the strongest prior radar-camera row by 3.32 mAP and 3.70 ODS, with the largest class gain on large vehicles. On ManTruckScenes, the margin is larger: 4DR360° exceeds

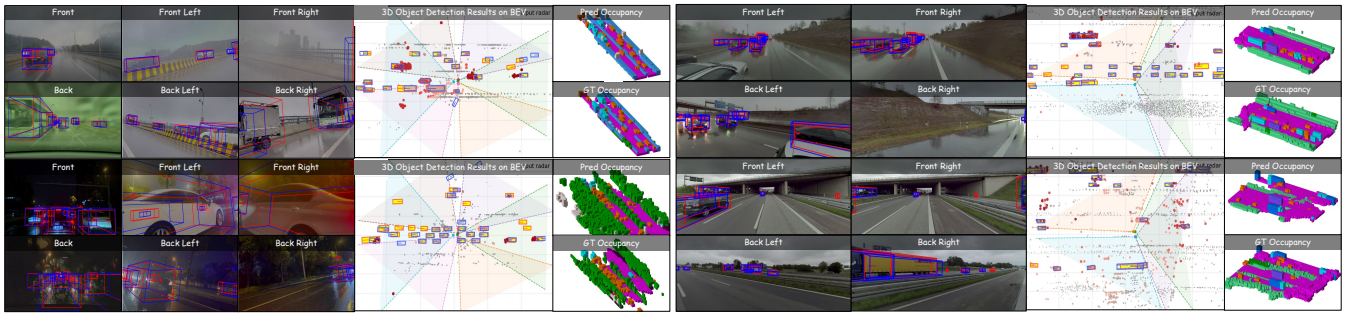

Figure 5. Qualitative results on OmniHD-Scenes (left) and ManTruckScenes (right).

Base	State	SBE	DTF	mAP↑	NDS↑	mIoU↑
✓	–	–	–	41.72	46.85	28.69
✓	✓	–	–	45.12	49.24	29.62
✓	✓	✓	–	47.45	50.89	30.22
✓	✓	✓	✓	49.57	52.97	31.65

Table 6. Component ablation.

Variant	Fr.	State	Dop.	mAP↑	NDS↑	mIoU↑
Single	1	–	–	47.45	50.89	30.22
BEV-TF	2	–	–	46.98	50.26	29.91
DTF	2	✓	✓	48.62	52.05	30.74
DTF	4	✓	✓	49.57	52.97	31.65

Table 7. Temporal horizon ablation.

HGSFusion by 5.67 mAP and 4.74 NDS, while improving the obstacle AP by 15.20 over the best baseline. The error metrics further localize the source of the gain. mATE and mAVE drop to 0.3613 and 0.9442 on ManTruckScenes, indicating better localization and motion estimation.

Semantic Occupancy Prediction Results

Tables 3 and 4 show complementary occupancy behavior. On OmniHD-Scenes, 4DR360° reaches best results, improving over Doracamom by 1.09 SC IoU and 2.90 mIoU. The gains are strongest on thin or foreground classes, while driveable surface and sidewalk also improves. On ManTruckScenes, 4DR360° gives the best SC IoU and improves mIoU by +2.29 over HGSFusion.

Adverse Conditions

Table 5 evaluates night and rainy OmniHD-Scenes subsets. 4DR360° reaches best performance, improving over Doracamom by 2.32, 2.83, and 2.87 points. These adverse scenes emphasize the regime where 4D radar is complementary to cameras, since radar returns remain sparse but stable when image appearance is degraded by illumination or weather.

Ablation Studies

Ablations are conducted on ManTruckScenes. Further details are reported in the supplementary materials. Detailed resource and efficiency comparisons are provided in the supplementary materials, showing that the accuracy gains do not come from a resource increase.

Component contribution. Table 6 shows that the first improvement appears as soon as occupancy is organized as a coarse-to-fine state path. This row adds intermediate state prediction, stage propagation, and occupancy-aware voxel recovery, bringing 3.40 mAP, 2.39 NDS, and 0.93 mIoU over the baseline. Then, SBE adds another 2.33 mAP, 1.65 NDS, and 0.60 mIoU, confirming that feeding the explicit state back to BEV matters for both boxes and voxels. DTF further adds 2.12 mAP, 2.08 NDS, and 1.43 mIoU, giving a total improvement of 7.85/6.12/2.96 over the baseline.

Variant	Fr.	Ego	State	Dop.	mAP↑	NDS↑	mIoU↑	mAVE↓
No Temp.	1	–	–	–	47.45	50.89	30.22	1.0213
Ego Align	4	✓	–	–	47.12	50.45	29.96	0.9986
State Mem.	4	✓	✓	–	49.06	52.28	30.92	0.9824
DTF	4	✓	✓	✓	49.57	52.97	31.65	0.9442

Table 8. Doppler ablation.

Temporal horizon. Table 7 shows that simply adding history is harmful: two-frame BEV-TF drops below the single-frame model by 0.47 mAP and 0.31 mIoU. DTF reverses this trend because it selects and aligns history through the learned state. Two-frame DTF improves over single-frame by 1.17 mAP and 0.52 mIoU, while four-frame DTF gives the best result, showing that temporal context is useful only when history is retrieved with state and motion guidance.

Doppler effectiveness. Table 8 makes the motion effect explicit. Ego alignment alone reduces mAVE from 1.0213 to 0.9986 but hurts mAP, NDS, and mIoU, showing that ego compensation cannot handle moving objects by itself. Explicit state memory recovers the metrics, and then Doppler gives the decisive motion correction: mAP, NDS, and mIoU further increase by 0.51, 0.69, and 0.73, while mAVE drops from 0.9824 to 0.9442.

Conclusion

In this work, we presented 4DR360°, a systematic 4D radar-camera framework for 360° multi-view multi-task perception. Its main contribution is occupancy-state reasoning, which models semantic occupancy as a persistent scene state rather than a terminal output for radar-camera representation learning. SBE and DTF make this state useful within and across frames, allowing foreground evidence, dense layout, and radar motion to reinforce a common full-scene representation. This changes occupancy from an auxiliary prediction into the interface that couples object detection with comprehensive scene understanding. Experiments on OmniHD-Scenes and ManTruckScenes show consistent

gains in detection and occupancy prediction, while the latter occupancy export provides a second benchmark contract for future 4D radar-camera multi-task studies. Future work will focus on radar-aware end-to-end or VLA research.

References

Bai, X.; Cheng, J.; Wang, S.; Luo, Y.; Zheng, L.; Zhang, X.; Cao, S.-Y.; and Shen, H.-L. 2025a. SD4R: Sparse-to-Dense Learning for 3D Object Detection with 4D Radar. In *IEEE International Conference on Intelligent Transportation Systems*, 4362–4368.

Bai, X.; Yang, Q.; Zhou, Z.; Zhang, F.; Wu, Z.; Cao, S.-Y.; Zheng, L.; Yu, B.; Wang, F.; Bai, J.; and Shen, H.-L. 2025b. LGDD: Local-Global Synergistic Dual-Branch 3D Object Detection Using 4D Radar. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 13318–13325.

Bai, X.; Yu, Z.; Zheng, L.; Zhang, X.; Zhou, Z.; Zhang, X.; Wang, F.; Bai, J.; and Shen, H.-L. 2024. SGDet3D: Semantics and Geometry Fusion for 3D Object Detection Using 4D Radar and Camera. *IEEE Robotics and Automation Letters*, 1–8.

Bai, X.; Zheng, L.; Cao, S.-Y.; Zhang, X.; Wu, Z.; Yu, B.; Wang, F.; Bai, J.; and Shen, H.-L. 2026a. Boosting Instance Awareness via Cross-View Correlation with 4D Radar and Camera for 3D Object Detection. *arXiv preprint arXiv:2602.20632*.

Bai, X.; Zheng, L.; Guan, R.; Cao, S.-Y.; and Shen, H.-L. 2026b. RC-GeoCP: Geometric Consensus for Radar-Camera Collaborative Perception. *arXiv preprint arXiv:2603.00654*.

Bai, X.; Zhou, C.; Zheng, L.; Cao, S.-Y.; Liu, J.; Zhang, X.; Li, Y.; Zhang, Z.; and Shen, H.-L. 2026c. RaGS: Unleashing 3D Gaussian Splatting from 4D Radar and Monocular Cue for 3D Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Cao, A.-Q.; and De Charette, R. 2022. MonoScene: Monocular 3D Semantic Scene Completion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3991–4001.

Gu, Z.; Ma, J.; Huang, Y.; Wei, H.; Chen, Z.; Zhang, H.; and Hong, W. 2025. HGSFusion: Radar-Camera Fusion with Hybrid Generation and Synchronization for 3D Object Detection. In *AAAI Conference on Artificial Intelligence*, volume 39, 3185–3193.

Guan, R.; Liu, J.; Liang, S.; Ding, F.; Yao, S.; Bai, X.; Liu, D.; Huang, T.; Mao, G.; and Xiong, H. 2025. Wavelet-based Multi-View Fusion of 4D Radar Tensor and Camera for Robust 3D Object Detection. *arXiv preprint arXiv:2512.22972*.

Hong, Y.; Liu, Q.; Cheng, H.; Ma, D.; Dai, H.; Wang, Y.; Cao, G.; and Ding, Y. 2024. UniVision: A Unified Framework for Vision-Centric 3D Perception. *arXiv preprint arXiv:2401.06994*.

Huang, Y.; Zheng, W.; Zhang, Y.; Zhou, J.; and Lu, J. 2024. GaussianFormer: Scene as Gaussians for Vision-Based 3D Semantic Occupancy Prediction. In *European Conference on Computer Vision*, 376–393. Springer.

Kim, Y.; Kim, S.; Choi, J. W.; and Kum, D. 2023. CRAFT: Camera-Radar 3D Object Detection with Spatio-Contextual Fusion Transformer. In *AAAI Conference on Artificial Intelligence*, volume 37, 1160–1168.

Li, Y.; Yu, Z.; Choy, C.; Xiao, C.; Alvarez, J. M.; Fidler, S.; Feng, C.; and Anandkumar, A. 2023. VoxFormer: Sparse Voxel Transformer for Camera-based 3D Semantic Scene Completion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9087–9098.

Lin, Z.; Liu, Z.; Xia, Z.; Wang, X.; Wang, Y.; Qi, S.; Dong, Y.; Dong, N.; Zhang, L.; and Zhu, C. 2024. RCBEVDet: Radar-camera Fusion in Bird's Eye View for 3D Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14928–14937.

Liu, Z.; Tang, H.; Amini, A.; Yang, X.; Mao, H.; Rus, D. L.; and Han, S. 2023. BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation. In *IEEE International Conference on Robotics and Automation*, 2774–2781.

Long, Y.; Kumar, A.; Morris, D.; Liu, X.; Castro, M.; and Chakravarty, P. 2023. RADIANT: Radar-Image Association Network for 3D Object Detection. In *AAAI Conference on Artificial Intelligence*, volume 37, 1808–1816.

Nabati, R.; and Qi, H. 2021. CenterFusion: Center-based Radar and Camera Fusion for 3D Object Detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 1527–1536.

Paek, D.-H.; Kong, S.-H.; and Wijaya, K. T. 2022. K-Radar: 4D Radar Object Detection for Autonomous Driving in Various Weather Conditions. *Advances in Neural Information Processing Systems*, 35: 3819–3829.

Palffy, A.; Pool, E.; Baratam, S.; Kooij, J. F.; and Gavrila, D. M. 2022. Multi-Class Road User Detection with 3+1D Radar in the View-of-Delft Dataset. *IEEE Robotics and Automation Letters*, 7(2): 4961–4968.

Tian, X.; Jiang, T.; Yun, L.; Mao, Y.; Yang, H.; Wang, Y.; Wang, Y.; and Zhao, H. 2023. Occ3D: A Large-Scale 3D Occupancy Prediction Benchmark for Autonomous Driving. *Advances in Neural Information Processing Systems*.

Wang, X.; Zhu, Z.; Xu, W.; Zhang, Y.; Wei, Y.; Chi, X.; Ye, Y.; Du, D.; Lu, J.; and Wang, X. 2023. OpenOccupancy: A Large Scale Benchmark for Surrounding Semantic Occupancy Perception. In *IEEE/CVF International Conference on Computer Vision*, 17850–17859.

Wang, Y.; Chen, Y.; Liao, X.; Fan, L.; and Zhang, Z. 2024. PanoOcc: Unified Occupancy Representation for Camera-Based 3D Panoptic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17158–17168.

Wei, Y.; Zhao, L.; Zheng, W.; Zhu, Z.; Zhou, J.; and Lu, J. 2023. SurroundOcc: Multi-Camera 3D Occupancy Prediction for Autonomous Driving. *arXiv preprint arXiv:2303.09551*.

Xia, Z.; Tang, Y.; Wang, Y.; Wang, Z.; and Qin, W. 2026. R4Det: 4D Radar-Camera Fusion for High-Performance 3D Object Detection. *arXiv preprint arXiv:2603.11566*.

Xiong, W.; Liu, J.; Huang, T.; Han, Q.-L.; Xia, Y.; and Zhu, B. 2024. LXL: LiDAR Excluded Lean 3D Object Detection with 4D Imaging Radar and Camera Fusion. *IEEE Transactions on Intelligent Vehicles*, 9(1): 79–92.

Zhang, F.; Yu, Z.; Li, C.; Zhang, R.; Bai, X.; Zhou, Z.; Cao,

S.-Y.; Wang, F.; and Shen, H.-L. 2025. Structure-Aware Radar-Camera Depth Estimation. In *IEEE International Conference on Robotics and Automation*, 13028–13035.

Zheng, L.; Li, S.; Tan, B.; Yang, L.; Chen, S.; Huang, L.; Bai, J.; Zhu, X.; and Ma, Z. 2023. RCFusion: Fusing 4-D Radar and Camera with Bird's-Eye View Features for 3-D Object Detection. *IEEE Transactions on Instrumentation and Measurement*, 72: 1–14.

Zheng, L.; Liu, J.; Guan, R.; Yang, L.; Lu, S.; Li, Y.; Bai, X.; Bai, J.; Ma, Z.; Shen, H.-L.; and Zhu, X. 2026. Doracamom: Joint 3D Detection and Occupancy Prediction with Multi-view 4D Radars and Cameras for Omnidirectional Perception. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zheng, L.; Ma, Z.; Zhu, X.; Tan, B.; Li, S.; Long, K.; Sun, W.; Chen, S.; Zhang, L.; Wan, M.; et al. 2022. TJ4DRadSet: A 4D Radar Dataset for Autonomous Driving. In *IEEE International Conference on Intelligent Transportation Systems*, 493–498.

Zheng, L.; Yang, L.; Lin, Q.; Ai, W.; Liu, M.; Lu, S.; Liu, J.; Ren, H.; Mo, J.; Bai, X.; et al. 2024. OmniHD-Scenes: A Next-Generation Multimodal Dataset for Autonomous Driving. *arXiv preprint arXiv:2412.10734*.

Zhou, Q.; Cao, J.; Leng, H.; Yin, Y.; Kun, Y.; and Zimmermann, R. 2024. SOGDet: Semantic-Occupancy Guided Multi-view 3D Object Detection. In *AAAI Conference on Artificial Intelligence*.