SGDET3D++: GEOMETRY-GROUNDED SEMANTICS FOR 4D RADAR–CAMERA 3D OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

4D radar complements dense image semantics with long-range geometry and radial motion, but existing radar–camera detectors largely solve *where* to align the modalities while leaving *whether* a piece of evidence supports an evolving object hypothesis implicit. An image token may describe an occluder, a nearby radar return may belong to another object, and a pose-aligned memory slot may carry incompatible motion. We formulate *hypothesis-conditioned evidence grounding*, which separates candidate access from evidence use: semantic, geometric, or temporal evidence is filtered or conditioned by the evolving 3D state before updating the corresponding query. SGDet3D++ instantiates this principle through Anchor-Grounded Semantic Retrieval (AGR), which conditions deformable image retrieval on pooled anchor-consistent radar support; Geometry-Consistent Anchor Refinement (GCR), which attentively aggregates individual associated returns; and Doppler-Verified Correspondence (DVC), which replaces history only when current radial motion contradicts it. SGDet3D++ improves the strongest compared method by 3.82 mAP and 6.82 ODS on OmniHD-Scenes and by 6.82 mAP and 9.22 NDS on ManTruckScenes, while also leading the listed methods in the TJ4DRadSet test comparison. Mechanism-targeted evaluations show that AGR improves strict AP in every projected-occlusion bin, the yaw-aligned box gate raises target-return purity from 29.95% to 58.87%, and DVC preserves 96.11% of motion-consistent history while retaining 75.90% contradiction recall. Code will be released.

1 INTRODUCTION

Reliable 3D object detection must localize dynamic objects under changing visibility, range, and traffic density. Surround-view cameras provide dense appearance and category cues, but remain ambiguous along the viewing ray. 4D imaging radar provides metric range, elevation, return strength, and relative radial motion, and remains effective in rain, fog, and night scenes. Recent radar–camera detectors exploit camera–radar complementarity by aligning image features with radar pillars, BEV grids, or object queries (Zheng et al., 2023; Liu et al., 2025; Lin et al., 2024; Yang et al., 2025; Gu et al., 2025; Chu et al., 2025). Although the aligned features make complementary cues available, scene-level alignment determines *where* evidence is placed in a shared representation but not *whether* it is admissible for the same evolving 3D hypothesis. We argue that this second decision—evidence–hypothesis compatibility—should be explicit at every object-level read.

Current fusion designs leave evidence–hypothesis consistency implicit. Dense BEV fusion projects both modalities into a shared ground-plane representation (Lin et al., 2024; Yang et al., 2025), while query- and proposal-level methods aggregate instance features after cross-modal association (Kim et al., 2023a; Chu et al., 2025; Zhong et al., 2025). Neither design verifies evidence against the evolving query's 3D radar geometry. A nearby image token may describe an occluder or background, while a radar return may arise from multipath or a reflector outside the object. Projection and proximity alone therefore cannot determine which evidence supports the query.

Unlike LiDAR, 4D radar provides sparse and noisy returns rather than dense object surfaces. Treating radar returns as a dense semantic map can introduce false support, whereas ignoring their object structure discards useful geometry and motion. We instead treat each return as potential support for a specific anchor rather than as unrestricted scene context. This view separates candidate access from

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.

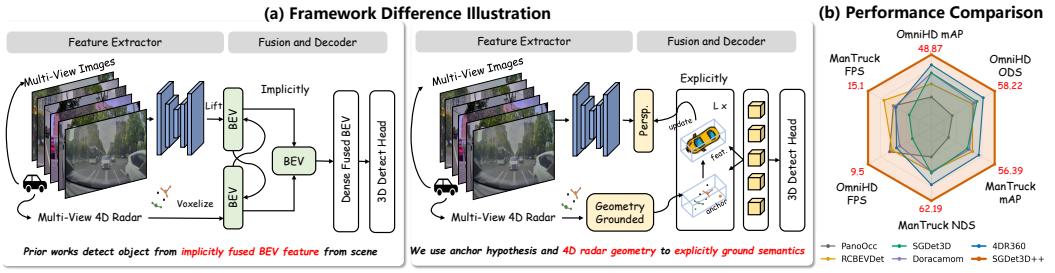


Figure 1: Conceptual and empirical comparison of radar–camera fusion strategies. *(a)* Comparison between implicit BEV fusion and geometry-grounded semantics via hypothesis-conditioned evidence grounding in SGDet3D++. *(b)* The radar chart compares representative methods on OmniHD-Scenes and ManTruckScenes and places SGDet3D++ in their accuracy–efficiency landscape.

evidence admission: proximity proposes evidence, while compatibility with the current hypothesis determines how it may affect a decision.

SGDet3D provides a useful starting point by combining radar-corrected depth, semantic radar pillars, and localization-aware object attention (Bai et al., 2025c). Yet its geometry and semantics are still coupled mainly through scene-level BEV features and projected supports. SGDet3D++ extends geometry-guided fusion through *hypothesis-conditioned evidence grounding*. Each sparse query uses its evolving 3D anchor to condition or verify radar evidence before semantic retrieval, geometric refinement, and temporal inheritance. The same principle thus governs three decoder decisions rather than introducing three unrelated feature blocks.

Figure 1(a) compares scene-level implicit BEV fusion with our hypothesis-conditioned design, in which each 3D anchor explicitly grounds the evidence retrieved by its query. Figure 1(b) compares the accuracy and efficiency of representative detectors on OmniHD-Scenes and ManTruckScenes. Figure 2 summarizes the SGDet3D++ architecture. AGR, GCR, and DVC ground semantic retrieval, box refinement, and temporal inheritance, respectively. At each decision, the current anchor determines which radar evidence may influence the query. Our contributions are summarized below.

- We formulate *hypothesis-conditioned evidence grounding* as a common object-decoder contract that separates candidate access from evidence admission: the evolving 3D state establishes semantic, geometric, or motion compatibility before evidence changes its query.
- We instantiate the principle at three decision interfaces. AGR uses pooled anchor-consistent radar statistics to condition image retrieval, GCR reads individual object-owned returns for box refinement, and DVC admits pose-aligned history with current radial motion.
- We validate accuracy and the proposed operating mechanism on three benchmarks. Controlled ablations isolate every module, while held-out decision-level diagnostics directly measure retrieval redistribution, return ownership, and motion-consistent memory inheritance.

2 Related Work

2.1 4D Radar-Based 3D Detection

Radar-only detectors encode sparse 4D-radar returns as points, voxels, or pillars. Point-feature aggregation and multi-view fusion improve local geometric descriptors (Xu et al., 2021; Yan & Wang, 2023; Bi et al., 2025; Wang et al., 2025), while pillar-oriented studies examine how sparse return support should be organized for detection (Palmer et al., 2023). Local–global interaction and stronger encoders further recover context from sparse measurements (Bai et al., 2025b; Xiao et al., 2026; Bai et al., 2025a). Radar nevertheless provides limited class semantics and incomplete object boundaries. We therefore use its range, reflectivity, and motion as object-level evidence rather than as a dense semantic substitute for images.

2.2 4D Radar and Camera Fusion 3D Detection

Radar–camera detectors commonly align the modalities in perspective space or a shared BEV representation. Projection and common BEV features complement image appearance with radar range (Long et al., 2023; Zheng et al., 2023; Kim et al., 2023b; Xiong et al., 2024; Kim et al., 2024; Baumann et al., 2024; Yan et al., 2026; Guan et al., 2025). Later methods strengthen BEV-level interaction with radar-aware fusion and multi-scale cross attention (Lin et al., 2024; Yang et al., 2025; Zhao et al., 2025; Wolters et al., 2025; Bai et al., 2026b). Other studies address radar sparsity through sampling, generation, and synchronization (Liu et al., 2025; Gu et al., 2025).

CRAFT associates radar returns with image proposals (Kim et al., 2023a), RaCFormer samples instance-relevant image and BEV features with object queries (Chu et al., 2025), and CVFusion aggregates point, image, and BEV features for proposal refinement (Zhong et al., 2025). In contrast, SGDet3D++ explicitly associates each query with anchor-consistent radar geometry before semantic retrieval, return-level refinement, and Doppler-verified temporal inheritance, extending SGDet3D (Bai et al., 2025c) from scene-level fusion to explicit object-level decisions.

3 Method

3.1 Overview

For each frame indexed by t, the detector receives synchronized multi-view images, 4D radar returns, camera calibration, and ego poses. The detector predicts 3D boxes with class labels and planar velocities, while depth serves only as an auxiliary geometric signal for feature construction and training. We write the input as

$$\mathcal{X}_t = \{I_t^k\}_{k=1}^{N_c} \cup \mathcal{R}_t \cup \mathcal{C}_t \cup \mathcal{T}_t, \qquad \mathcal{Y}_t = \{(b_n^{\mathrm{gt}}, c_n)\}_{n=1}^{N_t^{\mathrm{gt}}}, \tag{1}$$

where N_c is the number of cameras and I_t^k is the image from camera k. The sets $\mathcal{R}_t$, $\mathcal{C}_t$, and $\mathcal{T}_t$ contain the 4D radar returns, camera calibration parameters, and ego poses and timestamps for frame t, respectively. The target set contains N_t^{gt} objects. For object n, b_n^{gt} specifies the 3D location, extent, yaw, and planar velocity of object n, and c_n is its class label. SGDet3D++ realizes geometry-grounded semantics by associating radar support with each evolving hypothesis before semantic retrieval, geometric refinement, or temporal inheritance.

As shown in Figure 2, SGDet3D++ builds on the semantic–geometric principle of SGDet3D (Bai et al., 2025c), replacing dense image-to-BEV refinement with sparse object queries that sample image features at box-conditioned supports. Corrected-depth painting (Zhang et al., 2025) enriches occupied radar pillars, which the radar encoder scatters into the BEV feature map F_t^r. The image encoder produces the multi-camera, multi-level feature pyramid F_t^I. The front end couples image semantics with radar geometry, and the sparse object decoder grounds its evidence in the geometry of each object hypothesis. At layer ℓ, each anchor reads F_t^I and F_t^r through box-conditioned supports and updates its state according to

$$(f_i^{\ell+1}, b_i^{\ell+1}) = \mathcal{D}^{\ell}\big(f_i^{\ell}, b_i^{\ell}, F_t^I, F_t^r, \mathcal{R}_t, \mathcal{M}_{t-1}\big), \tag{2}$$

where f_i^{ℓ} and b_i^{ℓ} are the feature and 3D box state of query i before layer ℓ, and $\mathcal{M}_{t-1}$ is the object memory carried from the preceding frame. The operator $\mathcal{D}^{\ell}$ denotes the complete update performed by decoder layer ℓ from the query state, image and radar features, raw radar returns, and temporal memory shown in Eq. 2. AGR, GCR, and DVC perform semantic, geometric, and temporal grounding within the decoder update. For every operation, the current anchor defines radar support before the selected evidence modifies the object state. The decoder therefore selects evidence for each object instead of introducing another fused scene representation.

3.2 Hypothesis-Conditioned Evidence Grounding

Let $h_i^{\ell} = (f_i^{\ell}, b_i^{\ell})$ denote the feature and 3D state of hypothesis i. We define evidence grounding as an operation-specific map

$$e_i^{m,\ell} = \Gamma_m\big(E_t^m; h_i^{\ell}\big), \qquad m \in \{\mathrm{sem}, \mathrm{geo}, \mathrm{temp}\}, \tag{3}$$

where E_t^m is the candidate evidence available to a semantic, geometric, or temporal decoder operation. Unlike unrestricted feature fusion, Γ_m must expose a physical compatibility relation between

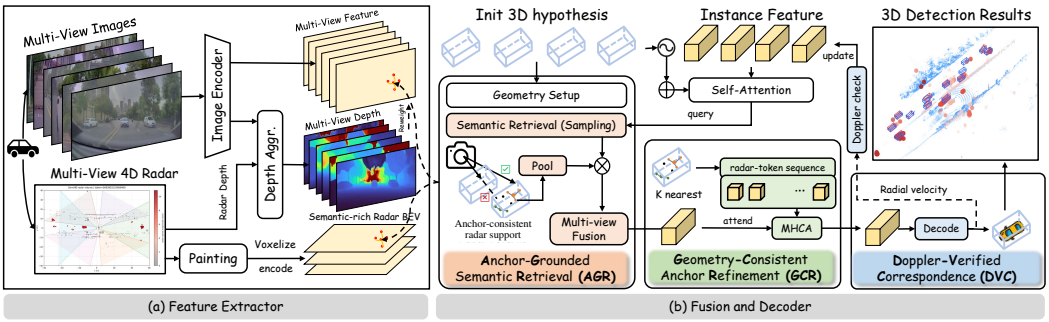


Figure 2: SGDet3D++ grounds object-query decoding in sparse 4D-radar evidence. AGR uses anchor-consistent radar support to reweight image tokens at box-conditioned supports, GCR supplies query-weighted local return evidence for box refinement, and DVC uses current radial Doppler to verify whether a historical object slot should be retained or replaced.

the candidate and the consuming hypothesis. AGR realizes Γ_{sem} by conditioning image-retrieval weights on anchor-consistent return statistics; GCR realizes Γ_{geo} by retaining box-owned returns as individual tokens; and DVC realizes Γ_{temp} by testing a history slot against current radial motion. This shared decision interface supports evidence-specific operators under a common evidence–hypothesis compatibility contract.

The decoder maintains N_q object hypotheses. At layer ℓ, hypothesis i has feature f_i^ℓ and box state $b_i^\ell = (x_i^\ell, y_i^\ell, z_i^\ell, w_i^\ell, l_i^\ell, h_i^\ell, \theta_i^\ell, v_{x,i}^\ell, v_{y,i}^\ell)$. The first three components define the box centre $o_i^\ell = (x_i^\ell, y_i^\ell, z_i^\ell)$. The next three give its width, length, and height, while the remaining components specify yaw and planar velocity. The current box generates 13 supports, including its centre, six fixed surface supports, and six learned object-relative supports. Let $s_j \in \mathbb{R}^3$ be the normalized object-relative offset of support j, let $R(\theta_i^\ell)$ be the rotation about the vertical axis defined by the current yaw, and let $\odot$ denote element-wise multiplication. Each support location is

$$p_{ij}^\ell = o_i^\ell + R(\theta_i^\ell)\big(s_j \odot (w_i^\ell, l_i^\ell, h_i^\ell)\big), \qquad j = 1, \ldots, 13. \tag{4}$$

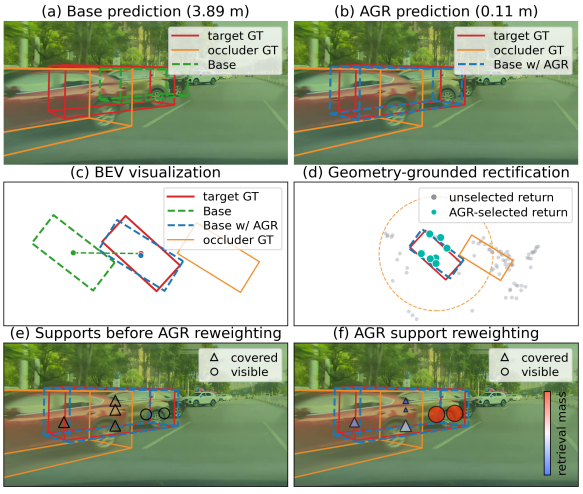


Figure 3: Qualitative AGR evidence in an occluded scene. (a–c) AGR reduces the target center error relative to Base. (d) The yaw-aligned box and radius gates restrict returns by anchor extent and association range; their intersection defines the anchor-consistent radar support. (e–f) Gated support shifts retrieval mass from occluder-covered supports (triangles) to visible supports (circles). Color indicates signed change from decrease (blue) to increase (red), and marker area its magnitude.

Each location p_{ij}^ℓ samples multi-view FPN features and the radar BEV feature map through deformable feature fusion. The subsequent self-attention, feed-forward, and refinement operations update both f_i^ℓ and b_i^ℓ. Thus, the location and extent of an anchor determine the accessible evidence rather than retrieving the same undifferentiated scene representation for every query.

The decoder's self-attention blocks use decoupled multi-head attention. Query and key projections include anchor embeddings, whereas value projections carry object features. Although decoupled attention allows a current query to interact with other current queries and pose-aligned historical queries, the attention mechanism does not establish whether a particular radar return supports that query. AGR and GCR therefore apply the association rule in Eq. 5 before image-feature retrieval and box refinement.

Anchor-Grounded Semantic Retrieval (AGR). AGR uses anchor-consistent

radar support to guide image retrieval. Figure 3 illustrates the effect of its two association gates. The yaw-aligned box gate removes returns outside the current anchor, and the radius gate prevents distant returns from entering the pooled radar descriptor (Figure 3(d)). Using the returns that pass both gates shifts retrieval mass from occluder-covered to visible supports in this case (Figure 3(e–f)) and yields the better-localized box in Figure 3(a–c). The aggregate evaluation quantifies retrieval redistribution and occlusion-stratified detection across the validation set. For an anchor b_i^ℓ, we write r for the 3D position of a radar return and transform the return into the yaw-aligned local coordinate $\widetilde{r}_{ir} = R(\theta_i^\ell)^\top (r - o_i^\ell)$. Let $u_i^\ell = (w_i^\ell, l_i^\ell, h_i^\ell)$ be the anchor extent. With anchor centre o_i^ℓ, margin δ, and radius ρ, AGR assigns return r to anchor i only if

$$\mathcal{R}_i^\ell = \left\{ r \,\middle|\, \left| R(\theta_i^\ell)^\top (r - o_i^\ell) \right| \preceq \tfrac{1}{2} u_i^\ell + \delta, \quad \|r - o_i^\ell\|_2 \le \rho \right\}. \tag{5}$$

The absolute value and inequality $\preceq$ are applied component-wise, so the first condition is a two-sided membership test in the yaw-aligned 3D box. The extent gate excludes returns outside the current anchor, while the radius gate prevents large or uncertain anchors from collecting distant returns. We call the selected set *anchor-consistent radar support*.

AGR pools the associated returns into an anchor-level evidence vector. For $n_i = |\mathcal{R}_i^\ell| > 0$, let P be the input return budget and use an overbar for the mean over $\mathcal{R}_i^\ell$. Let $S(r)$, $\mathrm{qual}(r)$, and $\mathrm{age}(r)$ denote the return strength, quality score, and sweep age of return r, respectively. The seven-dimensional descriptor, its D-dimensional embedding, and the resulting query update are

$$\begin{aligned} d_i^\ell &= \left[\bar{\widetilde{r}}_{i,x}, \bar{\widetilde{r}}_{i,y}, \bar{\widetilde{r}}_{i,z}, \frac{\log n_i}{\log P}, \overline{\frac{S(r)}{\max(\|r\|_2, 1\,\mathrm{m})^2}}, \overline{\mathrm{qual}(r)}, \overline{\mathrm{age}(r)} \right] \in \mathbb{R}^7, \\ e_i^\ell &= \Phi_{\mathrm{AGR}}(d_i^\ell) \in \mathbb{R}^D, \qquad \widehat{f}_i^\ell = f_i^\ell + \Phi_q([f_i^\ell \| e_i^\ell]). \end{aligned} \tag{6}$$

Here $P = 768$ is the maximum number of input returns and $D = 256$ is the decoder feature dimension in all reported experiments. The symbol $\|$ denotes feature concatenation, while Φ_{AGR} and Φ_q are learned MLPs that encode the pooled radar descriptor and compute its residual query update, respectively. The three local-coordinate means remain in metres. The count is logarithmically normalized to $[0, 1]$, and return strength $S(r)$—radar power on OmniHD-Scenes and RCS on ManTruckScenes—is normalized by the squared sensor range, clipped below at 1 m. The quality and sweep-age channels retain the shared loader's native units and are zero-filled when unavailable. The stated normalization keeps each physical cue on a documented scale before embedding.

The updated query $\widehat{f}_i^\ell$ reweights the projected image tokens without changing the 13 support locations p_{ij}^ℓ from Eq. 4.

$$\omega_{icsjg} = \mathrm{softmax}_{(c,s,j)} \left(\left[\Phi_\omega \left(\widehat{f}_i^\ell + \alpha_i^\ell + \gamma_c \right) \right]_{csjg} \right). \tag{7}$$

Here c, s, j, and g index the camera, FPN level, anchor support, and feature group, respectively. The vectors α_i^ℓ and γ_c are the learned embeddings of anchor i and camera c, and Φ_ω maps the combined query to one sampling logit for each (c, s, j, g) tuple. For every feature group g, the softmax normalizes the logits over all camera–level–support triples (c, s, j). AGR then aggregates the corresponding image features as

$$f_i^{I,\ell} = \sum_g \sum_{c,s,j} \omega_{icsjg} W_g F_{s,c}^I\big(\pi_c(p_{ij}^\ell)\big). \tag{8}$$

Here $F_{s,c}^I$ is the image feature map from camera c at FPN level s, π_c is the calibrated projection from a 3D point to camera c, and W_g projects the sampled feature into group g. AGR changes the contribution of each image token after anchor-consistent association while leaving the support locations fixed and introducing no dense radar branch or direct box correction. GCR then uses the same associated return set for box refinement while preserving selected returns as individual evidence.

Geometry-Consistent Anchor Refinement (GCR). After semantic retrieval, box refinement still faces sparse and noisy radar evidence. A valid object may produce only a few returns from different surfaces, with varying return strength and Doppler reliability. Even after the association gate in Eq. 5, not every retained return is equally useful for correcting the box. GCR therefore keeps several associated returns and lets the current object feature attentively select the radar evidence used for anchor refinement.

Table 1: OmniHD-Scenes 3D detection results from our adapted benchmark suite.

Method	Mod.	mAP	ODS	mATE	mASE	mAOE	mAVE	Car	Ped.	Rider	LVeh.	FPS
PointPillars'19	R	23.82	37.21	0.6752	0.2447	0.3776	0.6789	52.74	0.69	28.57	13.29	62.2
RadarPillarNet'23	R	24.88	37.81	0.6597	0.2389	0.3736	0.6982	52.99	2.06	29.45	15.02	60.3
BEVFormer'22	C	29.17	30.54	1.1046	0.2346	0.4889	1.0797	53.64	14.48	33.55	15.01	11.4
PanoOcc'24	C	29.17	28.55	1.1500	0.2446	0.6378	1.6066	51.58	15.82	35.02	14.26	5.5
BEVFusion'22	R+C	33.95	42.62	0.5730	0.2465	0.3814	0.7474	56.25	11.66	50.90	16.99	3.6
RCFusion'23	R+C	34.88	40.65	0.5676	0.2535	0.4011	0.9208	57.17	12.87	51.35	18.11	3.6
RCBEVDet'24	R+C	35.53	45.04	0.5138	0.2305	0.3914	0.6825	62.35	10.11	54.60	15.06	5.2
SGDet3D'25	R+C	41.73	47.70	0.5430	0.2369	0.3885	0.6849	63.30	19.00	58.30	26.30	3.4
RaGS'26	R+C	35.88	43.45	–	–	–	–	–	–	–	–	–
Doracamom-S'26	R+C	37.60	41.31	0.6724	0.2329	0.4359	0.8579	58.94	17.84	52.72	20.89	4.8
Doracamom'26	R+C	39.12	46.22	0.6646	0.2331	0.3545	0.6151	61.12	19.83	53.35	22.18	4.2
4DR360°'26	R+C	45.05	51.40	0.4705	0.2423	0.3762	0.6013	65.47	21.73	60.41	32.58	4.5
SGDet3D++ **(ours)**	R+C	**48.87**	**58.22**	**0.4298**	**0.1968**	**0.3355**	**0.3350**	**72.02**	**24.50**	**62.17**	**36.78**	**9.5**

Table 2: ManTruckScenes 3D detection results.

Method	Mod.	mAP	NDS	mATE	mASE	mAOE	mAVE	Car	LVeh.	Trailer	Obs.	FPS
RadarPillarNet'23	R	26.61	37.73	0.5243	0.2413	0.1689	2.1915	44.43	30.66	28.65	2.68	64.2
LGDD'25	R	29.83	40.49	0.4922	0.2328	0.1229	1.5775	48.90	36.20	32.40	1.80	20.1
BEVFormer'22	C	28.30	35.47	0.7662	0.2583	0.1982	1.2747	39.56	28.10	19.66	25.87	10.7
PanoOcc'24	C	28.82	35.72	0.7686	0.2653	0.1920	1.2623	40.12	28.22	21.73	25.20	6.3
BEVFusion'22	R+C	30.97	40.65	0.5168	0.2486	0.1249	2.1377	47.24	36.09	30.15	10.40	4.6
RCFusion'23	R+C	31.61	40.57	0.5652	0.2483	0.1155	2.0670	48.51	36.84	28.58	12.49	5.4
LXL'24	R+C	34.70	43.04	0.4972	0.2470	0.1174	1.9877	51.95	41.14	31.41	14.30	5.2
RCBEVDet'24	R+C	41.72	46.85	0.4500	0.2658	0.1543	1.0407	61.93	38.00	34.38	32.58	6.5
HGSFusion'25	R+C	43.90	48.23	0.4607	0.2272	0.1668	1.9479	58.76	43.64	35.43	37.77	5.1
SGDet3D'25†	R+C	41.37	47.50	0.4252	0.2159	0.1523	1.6112	60.82	47.38	26.94	30.29	3.0
Doracamom'26	R+C	41.98	46.36	0.4874	0.2411	0.1982	1.5566	61.62	42.71	34.45	29.15	5.2
4DR360°'26	R+C	49.57	52.97	**0.3613**	0.2485	0.1571	0.9442	59.41	44.71	41.20	52.97	5.5
SGDet3D++ **(ours)**	R+C	**56.39**	**62.19**	0.3637	**0.2025**	**0.0814**	**0.5749**	**70.85**	**48.32**	**43.18**	**63.24**	**15.1**

Starting from $\mathcal{R}_i^\ell$, GCR keeps up to K_r nearest associated returns as $\mathcal{K}_i^\ell$. Each return is encoded with the cues needed for refinement.

$$h_{ir} = \Phi_{\text{ret}}\Big([\underbrace{\widetilde{r}_{ir}/\rho \| \|\widetilde{r}_{ir}\|_2/\rho}_{\text{anchor-local geometry}} \| \underbrace{S(r)/d(r)^2 \| \operatorname{qual}(r) \| \operatorname{age}(r) \| \nu(r)}_{\text{signal and radial motion}}] \Big), \quad r \in \mathcal{K}_i^\ell. \tag{9}$$

Here Φ_{ret} is the learned return encoder, ρ is the support radius, $d(r) = \|r\|_2$ is the sensor-to-return range, and $\nu(r)$ is the radial Doppler measurement. The quality and age functions have the same definitions as in Eq. 6. The anchor-local geometry indicates where the return lies with respect to the current box, while return strength, quality, sweep age, and Doppler describe the strength, reliability, recency, and radial motion of the measurement. Keeping the cues at the return level lets the model distinguish a strong surface return from a weak or motion-inconsistent return that passes the coarse association gate without agreeing with the current object state.

The encoded returns form a radar-token sequence, and the anchor feature provides the query.

$$\mathbf{T}_i^\ell = [h_{ir_1}^\top, \ldots, h_{ir_{K_r}}^\top]^\top, \qquad q_i^\ell = W_q f_i^\ell, \quad \mathbf{K}_i^\ell = W_k \mathbf{T}_i^\ell, \quad \mathbf{V}_i^\ell = W_v \mathbf{T}_i^\ell. \tag{10}$$

Here $\mathbf{T}_i^\ell$ is the return-token matrix, and W_q, W_k, and W_v are learned query, key, and value projections. The current object feature then aggregates the sparse radar-token sequence by attention.

$$\beta_i = \operatorname{softmax}\left(\frac{q_i^\ell (\mathbf{K}_i^\ell)^\top}{\sqrt{D}} + m_i \right), \quad g_i^\ell \quad = \beta_i \mathbf{V}_i^\ell. \tag{11}$$

Here D is the feature dimension defined above. Element m_{ir} of mask m_i is 0 for a valid selected return and a large negative value for a padded slot. The attention weights β_i let each anchor emphasize returns whose local geometry, signal strength, quality, and radial motion agree with its current object state, while suppressing less compatible returns. The resulting radar context g_i^ℓ is concatenated with f_i^ℓ and injected before the ordinary box refiner, so radar influences centre, size, orientation, and velocity through the refinement representation. Return strength and Doppler are therefore used as learnable evidence fields, not as hard thresholds or direct box residuals. If no anchor-consistent radar support is available, the anchor follows the original refinement path. AGR uses pooled radar evidence for semantic retrieval, while GCR attends to individual returns for box refinement.

Doppler-Verified Correspondence (DVC). DVC verifies temporal memory with current Doppler evidence. The decoder stores high-confidence object features and anchors, then transforms historical

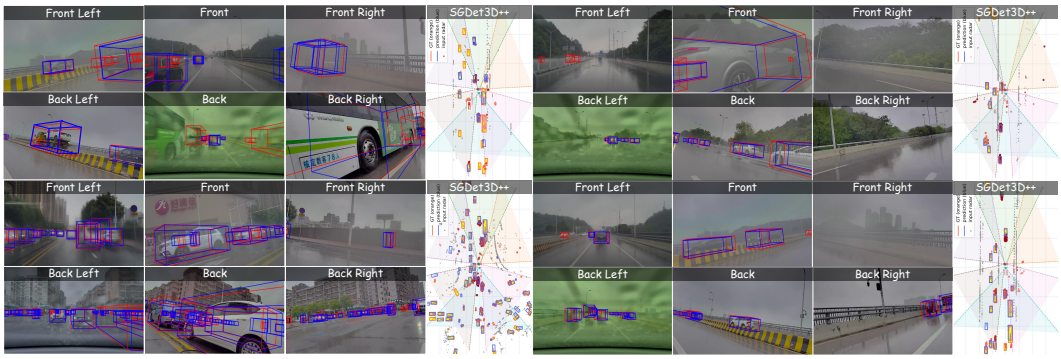


Figure 4: Qualitative visualization of SGDet3D++ on OmniHD-Scenes. Each example pairs multi-view image detections with BEV radar evidence colored by radial-speed magnitude. Red boxes denote ground truth and blue boxes denote predictions.

anchors to the current frame using ego pose and timestamp. Although ego-pose transformation removes displacement caused by ego motion, a stale or incorrect memory slot may remain spatially close to a current query. DVC therefore checks whether nearby radar returns have *relative radial* motion consistent with the motion predicted by the aligned historical anchor.

All quantities below are expressed in the current ego frame. For a return at 3D position r, define its unit line-of-sight direction as $\widehat{d}_r = r/\|r\|_2$. For a pose-aligned historical anchor a with centre c_a and ego-motion-compensated object velocity $v_a = (v_{x,a}, v_{y,a})$ resolved in the current ego axes, DVC first gathers current returns within a local support radius.

$$\mathcal{Q}_a = \{r \in \mathcal{R}_t \,|\, \|r - c_a\|_2 \le \rho_{\mathrm{DVC}}\}\,. \tag{12}$$

When $\mathcal{Q}_a$ is non-empty, DVC selects the spatially closest return $r_a^* = \arg\min_{r\in\mathcal{Q}_a} \|r - c_a\|_2$. DVC projects the anchor velocity onto the return direction as

$$\widehat{\nu}_a = \left\langle [v_{x,a}, v_{y,a}, 0]^\top, \widehat{d}_{r_a^*} \right\rangle. \tag{13}$$

The inner product $\langle\cdot,\cdot\rangle$ in Eq. 13 projects the anchor's planar velocity onto the selected return's line of sight, yielding the predicted radial velocity $\widehat{\nu}_a$. The value $\nu(r_a^*)$ is the loader-provided, ego-motion-compensated radial velocity of the selected return. DVC therefore compares two radial quantities rather than interpreting Doppler as an unconstrained BEV velocity. DVC marks a historical anchor as contradicted only when a nearby return exists and the radial-velocity mismatch exceeds threshold τ_v. The resulting test is

$$z_a = \mathbb{1}[\mathcal{Q}_a \neq \varnothing \;\wedge\; |\widehat{\nu}_a - \nu(r_a^*)| > \tau_v]\,. \tag{14}$$

where $\mathbb{1}[\cdot]$ is the indicator function and $z_a = 1$ denotes a contradiction. The memory contains M historical slots, which share bank indices with the first M current-query slots. If history slot k is contradicted, DVC substitutes the k-th freshly initialized current feature and anchor. Otherwise, DVC retains the history pair. The deterministic mapping uses neither nearest-object nor Hungarian matching. Missing radar evidence preserves history, and the update adds no learnable parameters while editing the existing temporal sequence without a second cache.

3.3 Detection Objective

The detection head supervises classification, box regression, centre quality, yaw quality, and planar velocity at decoder refinement layers. With $\mathcal{L}_{\mathrm{det}}$ denoting the weighted detection objective and $\mathcal{L}_{\mathrm{depth}}$ the multi-scale corrected-depth objective, the training loss is

$$\mathcal{L} = \mathcal{L}_{\mathrm{det}} + \lambda_{\mathrm{depth}}\mathcal{L}_{\mathrm{depth}}. \tag{15}$$

4 Experiments

Datasets and metrics. We evaluate 3D detection on OmniHD-Scenes (Zheng et al., 2026b) and ManTruckScenes (Fent et al., 2024). OmniHD-Scenes pairs six cameras with 4D radar and reports mAP, object detection score (ODS), four error metrics, and class AP for car, pedestrian, rider, and

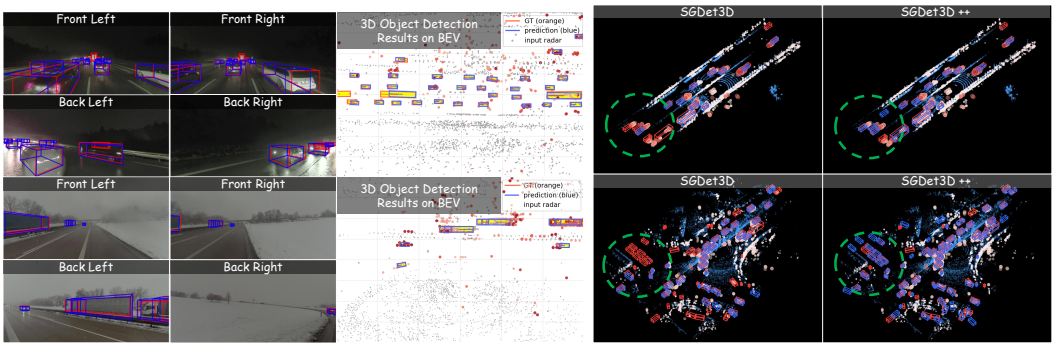


Figure 5: Qualitative evidence on ManTruckScenes. *Left:* SGDet3D++ aligns multi-view detections with radar-supported BEV predictions. *Right:* paired SGDet3D and SGDet3D++ outputs; green circles mark recovered detections and consolidated object boxes. Red boxes denote ground truth and blue boxes denote predictions.

large vehicle. ManTruckScenes complements it with commercial-vehicle scenes and reports mAP, NDS, the same errors, and AP for car, large vehicle, trailer, and obstacle.

Implementation details. We train in MMDetection3D with AdamW and mixed precision on four RTX PRO 6000 GPUs. Within each dataset, all adapted methods share the validation split, synchronized inputs, spatial range, and official evaluator; full protocols are in the supplement. The ManTruckScenes SGDet3D result marked † is from 4DR360° (Bai et al., 2026c); the remaining SGDet3D results use our adapted control. "Base" disables AGR, GCR, and DVC. The SGDet3D benchmark row reports one designated run; the controlled Base result in Table 5 is instead the mean of three independent runs.

4.1 Main Results

Tables 1 and 2 compare 3D detection on both benchmarks, where SGDet3D++ performs best among the listed methods. On OmniHD-Scenes, it raises the strongest R+C baseline (Bai et al., 2026c) from 45.05 to 48.87 mAP and from 51.40 to 58.22 ODS, improves all four classes, and reduces mATE, mAOE, and mAVE. On ManTruckScenes, it raises the strongest listed baseline from 49.57 to 56.39 mAP and from 52.97 to 62.19 NDS, with the largest class gain on obstacles. These semantic, localization, and motion gains match the complementary roles of AGR, GCR, and DVC. Figures 4 and 5 align predictions with BEV radar evidence; the latter also shows fewer missed or fragmented detections than SGDet3D.

Table 3: Adverse-condition on OmniHD-Scenes.

Method	Mod.	mAP	ODS
PointPillars (Lang et al., 2019)	R	28.63	40.70
RadarPillarNet (Zheng et al., 2023)	R	29.78	41.00
LSS (Philion & Fidler, 2020)	C	21.21	25.62
BEVFormer (Li et al., 2022)	C	30.39	31.61
PanoOcc (Wang et al., 2024)	C	26.09	27.16
BEVFusion (Liu et al., 2023)	R+C	35.83	44.95
RCBEVDet (Lin et al., 2024)	R+C	37.49	47.32
Doracamom (Zheng et al., 2026a)	R+C	41.86	48.74
4DR360° (Bai et al., 2026c)	R+C	44.18	51.57
SGDet3D++	R+C	**49.95**	**58.86**

4.2 Robustness and Decision-Level Evidence

Adverse-condition detection. Table 3 evaluates the adverse OmniHD-Scenes subset. Radar-only methods remain competitive, radar–camera methods generally exceed camera-only detectors, and SGDet3D++ obtains the best mAP and ODS. AGR grounds image retrieval in radar support, while GCR and DVC retain anchor-local geometry and Doppler-validated memory. Under camera-extrinsic perturbations, SGDet3D++ also outperforms an independently trained SGDet3D control and degrades more slowly (Supplementary Table A11).

Mechanism-level evidence. Table 4 reports targeted tests: AGR uses the complete Base/Base+AGR heavy-occlusion comparison, while the gate and DVC use the frozen 43-scene split. Relative to radius-only selection, the yaw-aligned box gate raises

Table 4: Mechanism-targeted evidence for the three grounding decisions. Values are percentages.

Module	Diagnostic	Base	Ground	Δ
AGR	Heavy-occ. AP@1m ↑	10.33	11.89	+1.56
Gate	Target-return purity ↑	29.95	58.87	+28.92
DVC	Motion-conflict F1 ↑	0.62	10.61	+9.99

target-return purity from 29.95% to 58.87% and lowers foreign returns from 7.15% to 1.16%. With conflicts comprising only 0.31% of eligible histories, DVC reaches 10.61% F1 while preserving 96.11% of consistent histories and retaining 75.90% conflict recall. AGR shifts 8.68% of retrieval probability and changes the maximum-weight camera–level–support route for 15.77% of feature groups. Supplementary Section G.4 provides full diagnostics.

4.3 Ablation Study

Component contribution. Table 5 reports the overall ablation of SGDet3D++. Every component improves mAP and ODS over Base, with GCR giving the largest isolated gain. AGR and GCR together add 5.25 mAP, and DVC adds another 3.19 mAP. Supplementary Table A8 analyzes the cues within GCR.

To isolate the box gate in Equation 5, we remove only the yaw-aligned 3D box-membership test while keeping the radius threshold and all other settings fixed. The radius-only variant loses 1.89 mAP and 2.08 strict AP@1m, showing that box membership provides information beyond proximity. Supplementary Table A9 reports the full metrics.

Detection under projected occlusion. Table 6 tests AGR under projected image occlusion. We stratify visible GT objects by the fraction of their multi-view projection occluded by closer objects. With GCR and DVC disabled, AGR improves AP@1m in every stratum, by 4.82, 3.29, and 1.56 points in the light, moderate, and heavy bins, respectively. Supplementary Table A7 separately controls for memory capacity. At the same 600-slot budget, DVC improves mAP from 46.12 to 48.87 and reduces mAVE from 0.3820 to 0.3350, directly attributing the gain to Doppler verification.

Table 5: Overall ablations on OmniHD-Scenes.

Base	Overall Ablation			mAP↑	ODS↑
	AGR	GCR	DVC		
✓				40.43 ± 0.28	47.82 ± 0.54
✓	✓			43.94 ± 0.31	51.24 ± 0.27
✓		✓		44.56 ± 0.16	52.07 ± 0.22
✓			✓	42.29 ± 0.35	49.74 ± 0.39
✓	✓	✓		45.68 ± 0.23	54.59 ± 0.20
✓	✓	✓	✓	48.87 ± 0.45	58.22 ± 0.19

Table 6: Class-averaged AP@1m by projected occlusion on OmniHD-Scenes. N is the GT count.

Method	Light $< 20\%$	Moderate 20–50%	Heavy $\geq 50\%$
GT objects (N)	43,246	10,926	40,172
Base	41.15	15.67	10.33
Base w/ AGR	45.97	18.96	11.89
Full	**51.24**	**23.15**	**16.74**

Table 7: Accuracy and resource comparison.

Method	Resource				Metrics	
	Mem	Lat.	Param	GFLOPs	mAP↑	NDS↑
BEVFusion	3.7	217.4	41.5	782.2	30.97	40.65
RCFusion	3.8	185.2	36.3	684.8	31.61	40.57
LXL	4.6	192.3	41.5	782.2	34.70	43.04
SGDet3D†	9.9	333.3	65.9	1187.1	41.37	47.50
RCBEVDet	5.1	153.8	86.7	1650.9	41.72	46.85
Doracamom	9.6	192.3	55.8	724.2	41.98	46.36
HGSFusion	8.2	196.1	44.7	976.7	43.90	48.23
4DR360°	15.4	181.8	140.1	1936.8	49.57	52.97
SGDet3D++	4.2	66.2	60.7	306.3	56.39	62.19

Efficiency analysis. On one RTX A6000, Table 7 reports memory in GB, latency in ms, and parameters in millions. SGDet3D++ attains the best mAP and NDS with 66.2 ms latency, 4.2 GB memory, 60.7 M parameters, and 306.3 GFLOPs; it also has the lowest latency and GFLOPs among the compared methods.

5 Conclusion

In this work, we present SGDet3D++, an object-level framework that grounds each query's semantic retrieval, geometric refinement, and temporal inheritance in object-conditioned 4D radar evidence. At every stage, the current anchor selects radar evidence for the same object hypothesis, turning radar from a generic proximity cue into object-specific geometric and motion evidence. AGR retrieves radar-supported image features, GCR refines query geometry with anchor-local returns, and DVC retains only historical hypotheses consistent with current Doppler. Anchor-conditioned evidence use therefore links semantic discrimination, spatial correction, and temporal continuity within one decoding path. Across two benchmarks, SGDet3D++ improves detection, localization, and motion over the strongest baselines. Robustness and resource analyses further show reliable and efficient detection. Future work will extend object-level grounding to vision–language–action models.

AI USE STATEMENT

Generative AI tools were used for literature organization, reviewer-style feedback, and language editing. The authors verified all citations, technical descriptions, analyses, and revisions, and take responsibility for the final content of this work.

REFERENCES

Xiaokai Bai, Jiahao Cheng, Songkai Wang, Yixuan Luo, Lianqing Zheng, Xiaohan Zhang, Si-Yuan Cao, and Hui-Liang Shen. SD4R: Sparse-to-Dense Learning for 3D Object Detection with 4D Radar. In *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, pp. 4362–4368, 2025a. doi: 10.1109/ITSC60802.2025.11423715.

Xiaokai Bai, Qin Yang, Zili Zhou, Fuyi Zhang, Zhe Wu, Si-Yuan Cao, Lianqing Zheng, Beinan Yu, Fang Wang, Jie Bai, and Hui-Liang Shen. LGDD: Local-Global Synergistic Dual-Branch 3D Object Detection Using 4D Radar. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 13318–13325. IEEE, 2025b. doi: 10.1109/IROS60139.2025.11246504.

Xiaokai Bai, Zhu Yu, Lianqing Zheng, Xiaohan Zhang, Zili Zhou, Xue Zhang, Fang Wang, Jie Bai, and Hui-Liang Shen. SGDet3D: Semantics and Geometry Fusion for 3D Object Detection Using 4D Radar and Camera. *IEEE Robotics and Automation Letters (RA-L)*, 10(1):828–835, 2025c. doi: 10.1109/LRA.2024.3513041.

Xiaokai Bai, Lianqing Zheng, Si-Yuan Cao, Xiaohan Zhang, Zhe Wu, Beinan Yu, Fang Wang, Jie Bai, and Hui-Liang Shen. Boosting Instance Awareness via Cross-View Correlation with 4D Radar and Camera for 3D Object Detection. *IEEE Transactions on Multimedia (TMM)*, 2026a. doi: 10.1109/TMM.2026.3684935.

Xiaokai Bai, Lianqing Zheng, Runwei Guan, Si-Yuan Cao, Songkai Wang, and Hui-Liang Shen. RC-GeoCP: Geometric Consensus for Radar-Camera Collaborative Perception. *arXiv:2603.00654*, 2026b.

Xiaokai Bai, Lianqing Zheng, Runwei Guan, Songkai Wang, Siyuan Cao, and Hui-Liang Shen. 4DR360: State Reasoning for Joint 3D Detection and Occupancy Prediction in 4D Radar-Camera Full-Scene Perception. *arXiv:2607.09629*, 2026c.

Xiaokai Bai, Chenxu Zhou, Lianqing Zheng, Jianan Liu, Si-Yuan Cao, Xiaohan Zhang, Yiming Li, Zhengzhuang Zhang, and Hui-Liang Shen. 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 (CVPR)*, pp. 4983–4992. IEEE, 2026d.

Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection with Transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1090–1099, 2022.

Nicolas Baumann, Michael Baumgartner, Edoardo Ghignone, Jonas Kühne, Tobias Fischer, Yung-Hsu Yang, Marc Pollefeys, and Michele Magno. CR3DT: Camera-RADAR Fusion for 3D Detection and Tracking. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4926–4933. IEEE, 2024. doi: 10.1109/IROS58592.2024.10801848.

Xin Bi, Caien Weng, Panpan Tong, Baojie Fan, and Arno Eichberge. MAFF-Net: Enhancing 3D Object Detection With 4D Radar via Multi-Assist Feature Fusion. *IEEE Robotics and Automation Letters (RA-L)*, 10(5):4284–4291, 2025. doi: 10.1109/LRA.2025.3550707.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17081–17091, 2025.

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. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:62062–62082, 2024. doi: 10.52202/079017-1982.

Zijian Gu, Jianwei Ma, Yan Huang, Honghao Wei, Zhanye Chen, Hui Zhang, and Wei Hong. HGS-Fusion: Radar-Camera Fusion with Hybrid Generation and Synchronization for 3D Object Detection. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pp. 3185–3193. AAAI, 2025. doi: 10.1609/aaai.v39i3.32328.

Runwei Guan, Jianan Liu, Shaofeng Liang, Fangqiang Ding, Shanliang Yao, Xiaokai Bai, Daizong Liu, Tao Huang, Guoqiang Mao, and Hui Xiong. Wavelet-based multi-view fusion of 4d radar tensor and camera for robust 3d object detection. *arXiv preprint arXiv:2512.22972*, 2025.

Jisong Kim, Minjae Seong, Geonho Bang, Dongsuk Kum, and Jun Won Choi. RCM-Fusion: Radar-Camera Multi-Level Fusion for 3D Object Detection. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 18236–18242. IEEE, 2024.

Youngseok Kim, Sanmin Kim, Jun Won Choi, and Dongsuk Kum. CRAFT: Camera-Radar 3D Object Detection with Spatio-Contextual Fusion Transformer. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 37, pp. 1160–1168, 2023a.

Youngseok Kim, Juyeb Shin, Sanmin Kim, In-Jae Lee, Jun Won Choi, and Dongsuk Kum. CRN: Camera Radar Net for Accurate, Robust, Efficient 3D Perception. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17615–17626, 2023b. doi: 10.1109/ICCV51070.2023.01615.

Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. PointPillars: Fast Encoders for Object Detection from Point Clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12697–12705, 2019.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BEVFormer: Learning Bird's-Eye-View Representation from Multi-camera Images via Spatiotemporal Transformers. In *European Conference on Computer Vision (ECCV)*, pp. 1–18, 2022.

Xuewu Lin, Zixiang Pei, Tianwei Lin, Lichao Huang, and Zhizhong Su. Sparse4D v3: Advancing End-to-End 3D Detection and Tracking. *arXiv:2311.11722*, 2023.

Zhiwei Lin, Zhe Liu, Zhongyu Xia, Xinhao Wang, Yongtao Wang, Shengxiang Qi, Yang Dong, Nan Dong, Le Zhang, and Ce Zhu. RCBEVDet: Radar-Camera Fusion in Bird's Eye View for 3D Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14928–14937. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01414.

Hongsi Liu, Jun Liu, Guangfeng Jiang, and Xin Jin. MSSF: A 4D Radar and Camera Fusion Framework With Multi-Stage Sampling for 3D Object Detection in Autonomous Driving. *IEEE Transactions on Intelligent Transportation Systems (T-ITS)*, 26(6):8641–8656, 2025. doi: 10.1109/TITS.2025.3554313.

Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L 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)*, pp. 2774–2781, 2023.

Yunfei Long, Abhinav Kumar, Daniel Morris, Xiaoming Liu, Marcos Castro, and Punarjay Chakravarty. RADIANT: Radar-Image Association Network for 3D Object Detection. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 37, pp. 1808–1816, 2023.

Ramin Nabati and Hairong Qi. CenterFusion: Center-based Radar and Camera Fusion for 3D Object Detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1527–1536, 2021.

Patrick Palmer, Martin Krueger, Richard Altendorfer, Ganesh Adam, and Torsten Bertram. Reviewing 3D Object Detectors in the Context of High-Resolution 3+1D Radar. *arXiv:2308.05478*, 2023.

Jonah Philion and Sanja Fidler. Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D. In *European Conference on Computer Vision (ECCV)*, pp. 194–210, 2020.

Nicolas Scheiner, Florian Kraus, Nils Appenrodt, Jürgen Dickmann, and Bernhard Sick. Object Detection for Automotive Radar Point Clouds – A Comparison. *AI Perspectives (AI Perspect.)*, 3 (1):6, 2021. doi: 10.1186/s42467-021-00012-z.

Xingzheng Wang, Jiahui Li, Jianbin Wu, Shaoyong Wu, and Lihua Li. DADAN: Dynamic-Augmented and Density-Aware Network for Accurate 3D Object Detection With 4D Radar. *IEEE Sensors Journal (IEEE Sens. J.)*, 25(6):10272–10286, 2025. doi: 10.1109/JSEN.2025.3528986.

Yuqi Wang, Yuntao Chen, Xingyu Liao, Lue Fan, and Zhaoxiang Zhang. PanoOcc: Unified Occupancy Representation for Camera-Based 3D Panoptic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17158–17168, 2024.

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 *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7467–7474. IEEE, 2025. doi: 10.1109/ICRA55743.2025.11127412.

Yichun Xiao, Runwei Guan, Jin Jin, and Fangqiang Ding. HyperDet: 3D Object Detection with Hyper 4D Radar Point Clouds. *arXiv:2602.11554*, 2026.

Weiyi Xiong, Jianan Liu, Tao Huang, Qing-Long Han, Yuxuan Xia, and Bing Zhu. LXL: LiDAR Excluded Lean 3D Object Detection with 4D Imaging Radar and Camera Fusion. *IEEE Transactions on Intelligent Vehicles (T-IV)*, 9(1):79–92, 2024. doi: 10.1109/TIV.2023.3321240.

Baowei Xu, Xinyu Zhang, Li Wang, Xiaomei Hu, Zhiwei Li, Shuyue Pan, Jun Li, and Yongqiang Deng. RPFA-Net: a 4D RaDAR Pillar Feature Attention Network for 3D Object Detection. In *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, pp. 3061–3066, 2021.

Hang Yan, Peifeng Cheng, Xiaokai Bai, Luping Wang, and Wuming Wu. Disadet: Depth interaction and semantic alignment for 4d radar-camera 3d object detection. *Displays*, pp. 103666, 2026.

Qiao Yan and Yihan Wang. MVFAN: Multi-View Feature Assisted Network for 4D Radar Object Detection. In *International Conference on Neural Information Processing (ICONIP)*, pp. 493–511. Springer, 2023.

Sheng Yang, Tong Zhan, Shichen Qiao, Jicheng Gong, Qing Yang, Jian Wang, and Yanfeng Lu. ZFusion: An Effective Fuser of Camera and 4D Radar for 3D Object Perception in Autonomous Driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3807–3816, 2025. doi: 10.1109/CVPRW67362.2025.00362.

Fuyi Zhang, Zhu Yu, Chunhao Li, Runmin Zhang, Xiaokai Bai, Zili Zhou, Si-Yuan Cao, Fang Wang, and Hui-Liang Shen. Structure-aware radar-camera depth estimation. In *2025 IEEE International Conference on Robotics and Automation*, pp. 13028–13035. IEEE, 2025.

Haocheng Zhao, Runwei Guan, Taoyu Wu, Ka Lok Man, Limin Yu, and Yutao Yue. UniBEVFusion: Unified Radar-Vision BEVFusion for 3D Object Detection. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6321–6327. IEEE, 2025. doi: 10.1109/ICRA55743.2025.11128067.

Lianqing Zheng, Zhixiong Ma, Xichan Zhu, Bin Tan, Sen Li, Kai Long, Weiqi Sun, Sihan Chen, Lu Zhang, Mengyue Wan, Libo Huang, and Jie Bai. TJ4DRadSet: A 4D Radar Dataset for Autonomous Driving. In *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, pp. 493–498, 2022. doi: 10.1109/ITSC55140.2022.9922539.

Lianqing Zheng, Sen Li, Bin Tan, Long Yang, Sihan Chen, Libo Huang, Jie Bai, Xichan Zhu, and Zhixiong Ma. RCFusion: Fusing 4-D Radar and Camera with Bird's-Eye View Features for 3-D Object Detection. *IEEE Transactions on Instrumentation and Measurement (TIM)*, 72:1–14, 2023. doi: 10.1109/TIM.2023.3280525.

Lianqing Zheng, Jianan Liu, Runwei Guan, Long Yang, Shouyi Lu, Yuanzhe Li, Xiaokai Bai, Jie Bai, Zhixiong Ma, Hui-Liang Shen, and Xichan Zhu. 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 (TCSVT)*, 36(6):7630–7645, 2026a. doi: 10.1109/TCSVT.2026.3657111.

Lianqing Zheng, Long Yang, Qunshu Lin, Wenjin Ai, Minghao Liu, Shouyi Lu, Jianan Liu, Hongze Ren, Jingyue Mo, Xiaokai Bai, Jie Bai, Zhixiong Ma, and Xichan Zhu. OmniHD-Scenes: A Next-Generation Multimodal Dataset for Autonomous Driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 48(6):7032–7049, 2026b. doi: 10.1109/TPAMI.2026.3663672.

Hanzhi Zhong, Zhiyu Xiang, Ruoyu Xu, Jingyun Fu, Peng Xu, Shaohong Wang, Zhihao Yang, Tianyu Pu, and Eryun Liu. CVFusion: Cross-View Fusion of 4D Radar and Camera for 3D Object Detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28188–28197. IEEE, October 2025. doi: 10.1109/ICCV51701.2025.02617.

APPENDIX

A OVERVIEW

The appendix provides additional evidence on object-level, radar-grounded access to image semantics in BEV fusion (Liu et al., 2023; Lin et al., 2024) and query-based temporal detection (Lin et al., 2023; Bai et al., 2025c). Evaluator-level and qualitative results describe recognition, localization, and motion under the official protocols and under degraded visual conditions. Fine-grained analyses examine object scale, commercial-vehicle geometry, radar support, projected occlusion, components, and key sensitivities. The final sections document the evaluation protocol, supplementary formulations, and implementation details.

An additional TJ4DRadSet test-set comparison evaluates transfer from the two surround-view benchmarks to a front-view radar–camera setting.

All AP values are percentages and lower error values are better. The method uses radar–camera detection inputs and depth supervision, without occupancy labels, satellite priors, or scene-level occupancy supervision.

B EVALUATOR-LEVEL RESULTS

Table A1 reports results on OmniHD-Scenes (Zheng et al., 2026b) and ManTruckScenes (Fent et al., 2024) under their official validation protocols. OmniHD-Scenes emphasizes fine-grained urban objects and adverse visual conditions, whereas ManTruckScenes emphasizes commercial vehicles and elongated object extents. We therefore report recognition, localization, and motion terms rather than comparing mAP alone. For dataset D with classes $\mathcal{C}_D$ and matching thresholds $\mathcal{T}_D$, the reported class AP and mAP are

$$\mathrm{AP}_c^D = \frac{1}{|\mathcal{T}_D|} \sum_{\tau \in \mathcal{T}_D} \mathrm{AP}_c^D(\tau), \qquad \mathrm{mAP}^D = \frac{1}{|\mathcal{C}_D|} \sum_{c \in \mathcal{C}_D} \mathrm{AP}_c^D \,. \tag{16}$$

Here $\mathcal{T}_{\mathrm{Omni}} = \{1, 2, 3, 4\}$ m and $\mathcal{T}_{\mathrm{ManTruck}} = \{0.5, 1, 2, 4\}$ m. The different matching thresholds make the two summary rows unsuitable for direct comparison.

Table A1: Evaluation summary on the official validation sets.

Dataset	Samples	mAP	Score	mATE	mAVE
OmniHD	3600	48.87	58.22	0.4298	0.3350
ManTruck	2475	56.39	62.19	0.3637	0.5749

ManTruckScenes reports a higher mAP and composite score but a larger mAVE than OmniHD-Scenes. The benchmarks also differ in class composition because ManTruckScenes contains more elongated commercial vehicles, whereas OmniHD-Scenes contains more fine-grained urban instances. Since lateral and extent errors can coexist with successful category matching, mATE and mAVE are reported alongside recognition to expose localization and motion behavior.

C TJ4DRADSET TEST-SET EVALUATION

To evaluate transfer to a front-view radar–camera benchmark, Table A2 reports AP40 on the TJ4DRadSet test set (Zheng et al., 2022). We use the moderate-difficulty values from the official evaluator, with IoU thresholds of 0.50 for car and truck and 0.25 for pedestrian and cyclist. Published comparison values are consolidated from the MSSF, SIFormer, CVFusion, and RaGS benchmark tables (Liu et al., 2025; Bai et al., 2026a; Zhong et al., 2025; Bai et al., 2026d); dashes denote class-wise results that are not reported by the cited sources.

SGDet3D++ reaches 46.44 $\mathrm{AP}_{\mathrm{3D}}$ and 53.72 $\mathrm{AP}_{\mathrm{BEV}}$. It improves over RaGS by 4.49 and 2.68 points, respectively. Among the listed methods, SGDet3D++ obtains the best aggregate AP in both views, the best car and truck AP in both views, and the best pedestrian AP in 3D. MSSF-PP retains the best cyclist AP in both views and the best pedestrian AP in BEV. This complementary front-view evaluation shows that object-level grounding remains competitive beyond the surround-view rigs used by the two main benchmarks.

Table A2: Comparison with state-of-the-art approaches on the TJ4DRadSet test set. AP40 at moderate difficulty is reported in percent. R and C denote 4D radar and camera, respectively. Best values among the listed methods are bold.

Method	Mod.	AP_{3D} (%)					AP_{BEV} (%)				
		Car	Ped.	Cyc.	Truck	mAP	Car	Ped.	Cyc.	Truck	mAP
ImVoxelNet'22	C	22.55	13.73	9.67	13.87	14.96	26.10	14.21	10.99	17.18	17.12
PointPillars'19	R	21.26	28.33	52.47	11.18	28.31	38.34	32.26	56.11	18.19	36.23
CenterPoint'21	R	22.03	25.02	53.32	15.92	29.07	33.03	27.87	58.74	25.09	36.18
RadarPillarNet'23	R	28.45	26.24	51.57	15.20	30.37	45.72	29.19	56.89	25.17	39.24
SMURF'24	R	28.47	26.22	54.61	22.64	32.99	43.13	29.19	58.81	32.80	40.98
RCFusion'23	R+C	29.72	27.17	54.93	23.56	33.85	40.89	30.95	58.30	28.92	39.76
FUTR3D'23	R+C	–	–	–	–	32.42	–	–	–	–	37.51
BEVFusion'23	R+C	–	–	–	–	32.71	–	–	–	–	41.12
LXL'24	R+C	–	–	–	–	36.32	–	–	–	–	41.20
UniBEVFusion'25	R+C	44.26	27.92	51.11	27.75	37.76	50.43	29.57	56.48	35.22	42.92
MSSF-PP'25	R+C	46.78	31.11	**57.15**	21.24	39.07	64.13	**34.33**	**63.05**	30.64	48.04
HGSFusion'25	R+C	–	–	–	–	37.21	–	–	–	–	43.23
CVFusion'25	R+C	51.54	29.49	49.41	29.55	40.00	58.07	31.65	51.29	35.29	44.07
SGDet3D'25	R+C	59.43	26.57	51.30	30.00	41.82	66.38	29.18	53.72	39.36	47.16
RaGS'26	R+C	–	–	–	–	41.95	–	–	–	–	51.04
SGDet3D++ **(ours)**	R+C	**60.11**	**31.56**	53.90	**40.19**	**46.44**	**74.63**	34.27	56.97	**49.00**	**53.72**

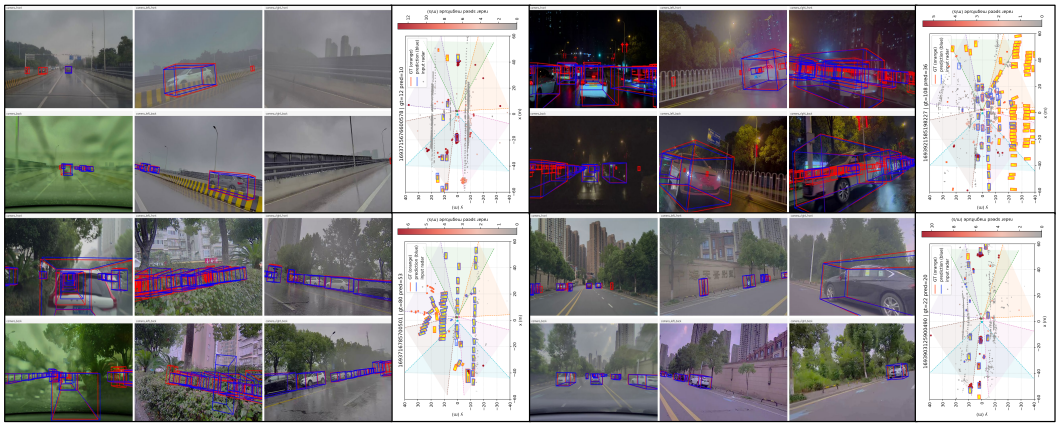


Figure A1: OmniHD-Scenes qualitative comparisons in four visual conditions. Each cell contains the BEV evidence and the synchronized camera views for one sample. Within each BEV/camera block, the left panel is the adapted SGDet3D baseline and the right panel is the full SGDet3D++ model. Orange boxes denote ground truth and blue boxes denote predictions.

D Additional Qualitative Evidence

Figures A1 and A2 provide a qualitative audit on synchronized camera and BEV views. OmniHD-Scenes covers rainy daytime, black night, dense cloudy daytime, and normal daytime scenes. The paired baseline/full panels show that, when contrast is weak or nearby objects overlap in projection, the full model more often preserves hypotheses that remain plausible in both the camera image and local radar-supported BEV region. The dense-cloudy case further shows improved separation of neighbouring object extents, while the normal-day case serves as a control. ManTruckScenes complements this comparison with night, rain or low-contrast, snow or haze, and clear-day traffic. Across these examples, the full model retains selective long-range vehicle hypotheses without filling the road corridor with unsupported boxes. The figures provide qualitative evidence of object-level semantic access and geometry checking. The studies below quantify the aggregate effect.

E FINE-GRAINED OMNIHD-SCENES EVALUATION

The OmniHD-Scenes evaluator averages class AP across four center-distance thresholds. Table A3 keeps these thresholds separate. Car and rider AP remain high under strict localization, while pedestrian and large-vehicle AP increase more visibly as the matching radius relaxes. The gap reflects two sparse-radar regimes because small objects are often weakly observed, whereas large vehicles may create separated reflections around the same object extent. We summarize localization sensitivity by the AP spread

$$\Delta_c^D = \mathrm{AP}_c^D(\tau_{\max}) - \mathrm{AP}_c^D(\tau_{\min}), \tag{17}$$

where a larger Δ_c^D indicates that more detections are recovered only after relaxing the center-distance criterion.

The strict 1m criterion distinguishes semantic recognition from precise geometric support. Car and rider retain 65.65% and 58.69% AP, respectively, while pedestrian AP rises from 18.64% at 1m to 29.33% at 4m and large-vehicle AP rises from 26.13% to 42.23%. The two gaps reflect different radar regimes. Pedestrians yield few weak returns, whereas a large vehicle can yield returns from separated surfaces, and the class-wise ATE values follow the same ordering. To accommodate both regimes, GCR aggregates multiple anchor-consistent returns instead of relying on a single geometric correction, following the broader principle that radar should resolve uncertain visual hypotheses (Nabati & Qi, 2021; Bai et al., 2022), while using the richer 4D evidence available to SGDet3D++.

Table A3: OmniHD-Scenes class AP under official center-distance thresholds. AP columns are percentages. ATE and AVE are class-wise errors.

Class	1m	2m	3m	4m	AP	ATE	AVE
Car	65.65	71.81	74.49	76.13	72.02	0.286	0.242
Ped.	18.64	23.26	26.77	29.33	24.50	0.595	0.390
Rider	58.69	61.92	63.47	64.59	62.17	0.261	0.349
L.Veh.	26.13	37.70	41.07	42.23	36.78	0.577	0.360

F FINE-GRAINED MANTRUCKSCENES EVALUATION

ManTruckScenes uses a different threshold set, including a stricter 0.5m matching radius. Table A4 shows that the final model keeps strong obstacle AP even at 0.5m and obtains the best averaged AP on every class in the main table. The large-vehicle and trailer rows expose a typical truck-scene difficulty. Recognition improves strongly under looser matching, but exact centre localization remains harder for elongated vehicles.

Table A4: ManTruckScenes class AP under official center-distance thresholds. AP columns are percentages. ATE and AVE are class-wise errors.

Class	0.5m	1m	2m	4m	AP	ATE	AVE
Car	49.93	70.20	79.25	84.01	70.85	0.300	0.715
L.Veh.	21.05	42.24	59.88	70.10	48.32	0.477	0.686
Trailer	22.83	40.84	49.74	59.30	43.18	0.418	0.746
Obs.	54.60	63.07	66.94	68.33	63.24	0.259	0.153

The table exposes two complementary behaviours. Road obstacles retain 54.60% AP at 0.5m and 63.24% on average, even under the strictest matching criterion. In contrast, large vehicles and trailers gain 49.05 and 36.47 points, respectively, between the strictest and loosest thresholds. Their ATE values of 0.477 and 0.418 show that centre localization, rather than category recognition, remains the dominant challenge. Their higher AVE further exposes motion error. DVC directly targets this temporal dimension by using radial Doppler to validate correspondence.

G ADDITIONAL ABLATION STUDIES

Beyond the overall component study in the main paper, the supplementary ablations examine visual-condition robustness, strict localization, radar-return budget, Doppler tolerance, and depth supervision. To retain a direct comparison, every row changes only the named factor under matched optimization, data, and evaluation protocols.

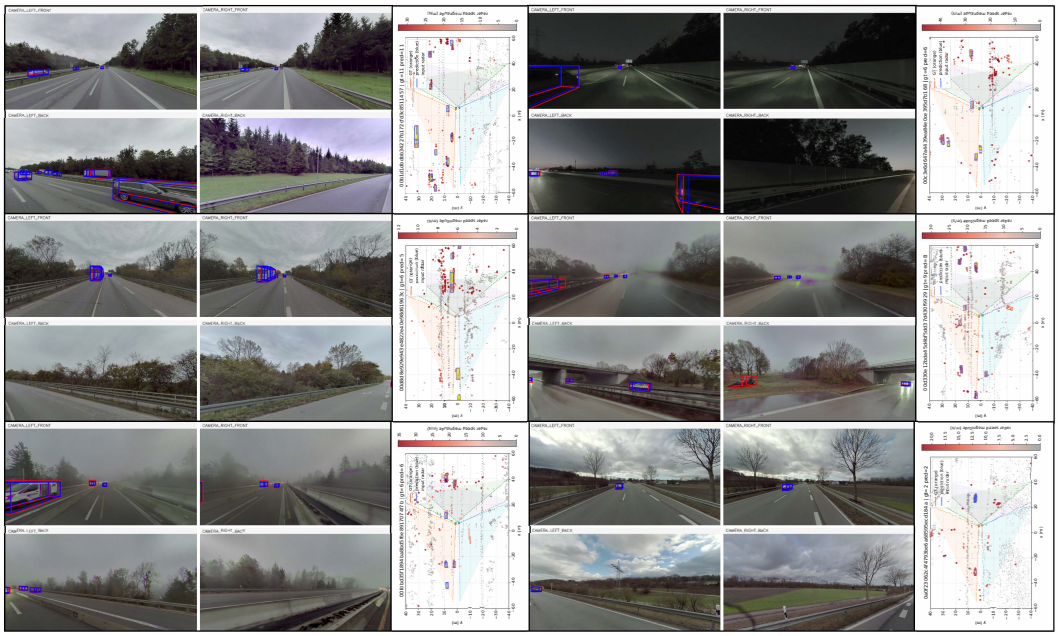


Figure A2: ManTruckScenes qualitative examples from SGDet3D++. Each cell aligns the BEV prediction with the camera views for the same sample. Orange boxes denote ground truth and blue boxes denote model predictions.

G.1 Condition-Split Robustness

The qualitative examples in Section D are paired with the official adverse validation subsets supplied with each benchmark. Table A5 reports overall performance together with a strict matching score, thereby exposing localization under tighter matching. Poor illumination and precipitation can weaken camera evidence, whereas radar continues to provide range, return-strength, and radial-motion cues (Scheiner et al., 2021). The full and adverse rows are evaluated with the same model and dataset-specific evaluator.

On OmniHD-Scenes, the adverse subset retains 49.95 mAP and 58.86 ODS, although weak radar support and degraded image evidence coincide with lower pedestrian AP than on the full set. Even with the lower pedestrian AP, the subset retains a 43.19 strict score at the 1m threshold. The supplied ManTruckScenes adverse subset differs from the full set in class, range, and scene composition, and SGDet3D++ reaches 71.92 mAP and 72.48 NDS within that setting.

Table A5: Official full and adverse-subset evaluation. Strict is class-averaged AP@1m for OmniHD-Scenes and AP@0.5m for ManTruckScenes.

Dataset / split	N	mAP	Score	Strict	Focus AP	Focus
OmniHD / full	3600	48.87	58.22	42.28	24.50	Ped.
OmniHD / adverse	1860	49.95	58.86	43.19	19.10	Ped.
ManTruck / full	2475	56.39	62.19	37.10	48.32	L.Veh.
ManTruck / adverse	360	71.92	72.48	52.91	46.72	L.Veh.

G.2 Validation Composition and Occlusion-Stratified Gains

Figure A3 jointly characterizes the 94,344 eligible OmniHD-Scenes objects and the model advantage across their projected- occlusion strata. For a ground-truth object i, let $\mathcal{R}_i^{\mathrm{gt}}$ denote returns selected by the same box-and-radius association rule as Eq. 5. The class-conditional radar-support rate is

$$s_c = \frac{\sum_i \mathbb{1}[y_i = c \ \wedge \ |\mathcal{R}_i^{\mathrm{gt}}| > 0]}{\sum_i \mathbb{1}[y_i = c]}. \tag{18}$$

Anchor-consistent radar support is present for 59.1% of eligible objects. The rate ranges from 35.4% for pedestrians to 73.6% for large vehicles, showing that radar evidence is informative but uneven across classes. The light, moderate, and heavy strata contain 43,246, 10,926, and 40,172 objects, respectively. Relative to Base, the Full model improves class-averaged AP@1m by 10.09, 7.48, and 6.41 points across these strata. The consistent advantage, including under heavy projected occlusion, spans the full severity range.

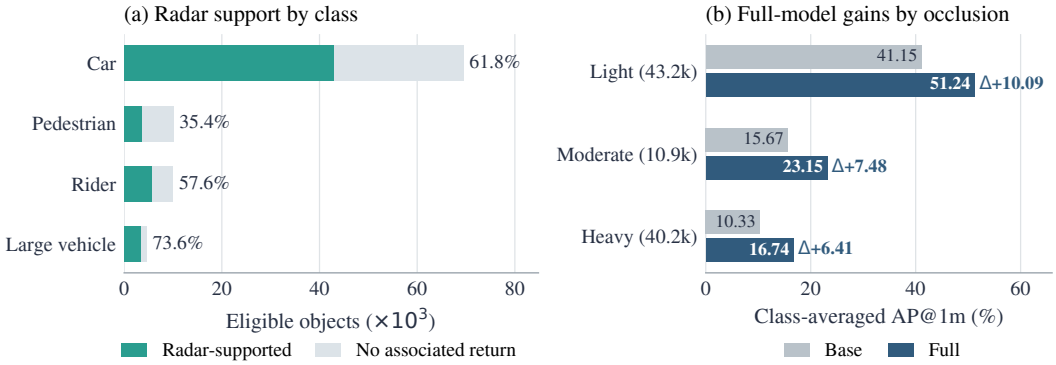


Figure A3: OmniHD-Scenes validation statistics. (a) Eligible ground-truth objects by class, partitioned by anchor-consistent radar support. Percentages denote s_c. (b) Class-averaged AP@1m for Base and Full in the projected-occlusion bins of Eq. 19. Parentheses report GT counts and Δ denotes Full minus Base. Full improves Base in every stratum.

G.3 Occlusion-Stratified Evaluation Protocol

The main-paper occlusion study derives severity from annotation geometry rather than the binary camera-visibility flag. For a GT object i visible in camera c, let P_i^c be the image-clipped convex hull of its projected 3D box and let z_i^c be its centre depth in that camera. We define the multi-view projected-occlusion ratio as

$$o_i = \frac{\sum_{c \in \mathcal{V}_i} \left| P_i^c \cap \left(\bigcup_{z_j^c < z_i^c}^{j} P_j^c \right) \right|}{\sum_{c \in \mathcal{V}_i} |P_i^c|}, \tag{19}$$

where $\mathcal{V}_i$ contains the cameras in which object i has a valid clipped projection. Only closer GT boxes contribute to the occluding union. We use three fixed bins. Light uses $o_i < 0.2$, moderate uses $0.2 \le o_i < 0.5$, and heavy uses $o_i \ge 0.5$. We report the number of GT objects in every bin. The ratio measures projected overlap, not latent physical visibility or pixel-level instance occlusion.

For each bin, we compute class-averaged AP at the strict 1m centre-distance threshold while keeping the model fixed across severity levels. GT objects outside the active bin are ignored rather than removed as background, and predictions matched to those ignored objects are also ignored. Ignoring both groups prevents detections of valid objects in another severity bin from being counted as false positives. The Base and Base w/ AGR variants isolate the effect of AGR, whereas the full model includes the complete evidence chain.

G.4 Held-Out Grounding Diagnostics

We evaluate the local prediction of each grounding operation independently of aggregate detection AP. A stable scene split assigns $\mathrm{SHA1}(\text{scene token}) \bmod 10 \in \{0, 1, 2\}$ to metric development and the other buckets to confirmation. The 17 exploration scenes are used once to freeze the definitions; Table A6 reports the disjoint 43-scene confirmation set. Confidence intervals use 2,000 scene-level bootstrap resamples with seed 2027.

For AGR, retrieval TV is the total-variation distance between the final decoder's image-retrieval distribution and a same-checkpoint counterfactual in which the AGR query update is removed.

Table A6: Decision-level grounding diagnostics on the frozen OmniHD-Scenes confirmation split. All values and confidence intervals are percentages. TV denotes total variation; Preserve is the rate of retaining GT motion-consistent history.

Module	Diagnostic	Control	Grounded	Δ	95% CI of Δ
AGR	Retrieval TV ↑	0.00	8.68	+8.68	[8.11, 9.25]
AGR	Maximum-route change ↑	0.00	15.77	+15.77	[14.47, 17.10]
Gate	Target-return purity ↑	29.95	58.87	+28.92	[27.67, 30.23]
Gate	Foreign-return rate ↓	7.15	1.16	−5.99	[−6.84, −5.04]
DVC	Motion-conflict precision ↑	0.31	5.70	+5.39	[3.68, 7.03]
DVC	Motion-conflict recall ↑	100.00	75.90	−24.10	–
DVC	Motion-conflict F1 ↑	0.62	10.61	+9.99	[6.97, 12.85]
DVC	Consistent-history preserve ↑	0.00	96.11	+96.11	[95.05, 96.97]

Route change is the fraction of feature groups whose maximum-weight camera–level–support tuple changes. To evaluate the shared return gate used by AGR and GCR, the control selects the nearest K returns within the same radius but omits the yaw-aligned box test. Purity is the fraction of selected returns owned by the matched GT target, and foreign rate is the fraction lying in another GT box. Both are micro-averaged over class-aware true positives at the strict 1m threshold.

For DVC, a historical anchor is first matched to a tracked previous-frame GT object. We retain cases with a current continuation, finite GT velocity, and a nearby return inside that continuation's oriented GT box. A GT motion conflict occurs when the historical anchor and current GT radial velocities differ by more than the configured 2.5 m/s tolerance along the selected return ray. The proximity control rejects every eligible history with a nearby return. This strict protocol yields 26,845 eligible histories, including 83 conflicts.

The confirmation results reproduce every frozen directional prediction. AGR redistributes a non-trivial amount of retrieval probability and changes the dominant route, while its controlled detector improves AP@1m in every projected-occlusion bin. The yaw-aligned box gate almost doubles target-return purity and removes most foreign-object returns before GCR attention. DVC replaces 1,105 of 26,845 eligible histories (4.12%), with TP/FP/FN/TN counts of 63/1,042/20/25,720. It preserves 96.11% of consistent histories, retains 75.90% conflict recall, and reaches a Doppler-residual AUROC of 0.913.

G.5 Focused Hyperparameter Ablations

Table A7 separates Doppler verification from raw memory capacity. Ordinary 600-slot memory adds only 0.44 mAP over no memory and leaves mAVE nearly unchanged. With the same capacity, DVC reaches 48.87 mAP and reduces mAVE from 0.3820 to 0.3350. Across verified memories, 600 slots gives the best mAP and mAVE, while 1200 slots gives the highest ODS.

Table A7: DVC memory ablation on OmniHD-Scenes.

Mem.	DVC	mAP	ODS	mAVE
0		45.68	54.59	0.3821
300	✓	47.73	57.18	0.3514
600		46.12	54.79	0.3820
600	✓	**48.87**	58.22	**0.3350**
1200	✓	48.41	**58.42**	0.3428

Table A8 isolates the evidence used by GCR before anchor refinement. Local XYZ first improves geometry, RCS adds a further gain when combined with XYZ, and current-frame Doppler provides the strongest configuration. The unchanged RCS-only row shows that signal strength is useful in conjunction with local geometry rather than as a standalone confidence gate. Unlike DVC, the Doppler term here operates within the current frame.

Table A8: Internal GCR ablation on OmniHD-Scenes. Dop. denotes relative radial velocity. AGR and DVC are fixed.

Ref.	XYZ	RCS	Dop.	mAP	ODS	mAOE	mATE
✓				46.72	55.46	0.3718	0.4635
✓		✓		46.72	55.46	0.3718	0.4635
✓	✓			47.34	56.02	0.3584	0.4517
✓	✓	✓		48.06	57.15	0.3427	0.4388
✓	✓	✓	✓	**48.87**	**58.22**	**0.3355**	**0.4298**

Table A9 examines sensitivity to four factors. Score denotes ODS and Strict denotes class-averaged AP@1m. AGR varies the association radius and extent gate, GCR varies the number of attended returns, DVC varies the radial-velocity tolerance, and the depth variants isolate the projection supervision supplied to the detector. Each comparison changes one factor while keeping the remaining architecture and optimization protocol fixed.

The matched controls reveal distinct operating ranges. AGR peaks at $\rho = 5$. A tighter radius can miss surface returns, whereas a looser one admits neighboring evidence, and removing the box gate causes the largest AGR drop. For GCR, performance rises from $K = 4$ to $K = 8$ but saturates at $K = 16$, where the larger return set can include more noise.

DVC peaks at the default tolerance. A strict threshold rejects history under Doppler noise, whereas a loose threshold can retain inconsistent hypotheses. Removing verification most degrades motion estimation. By contrast, the depth rows vary projection supervision rather than evidence verification. Removing all depth supervision is most harmful, whereas retaining LiDAR depth but dropping radar depth is less damaging.

Table A9: Hyperparameter ablations on OmniHD-Scenes.

Mod.	Variant	Value	mAP	Score	Strict	mAVE
AGR	tight radius	$\rho = 3$	47.91	57.41	41.45	0.3462
AGR	default	$\rho = 5$	48.87	58.22	42.28	0.3350
AGR	loose radius	$\rho = 7$	48.15	57.65	41.82	0.3421
AGR	no box gate	–	46.98	56.43	40.20	0.3635
GCR	small budget	$K = 4$	48.23	57.47	41.80	0.3440
GCR	default	$K = 8$	48.87	58.22	42.28	0.3350
GCR	large budget	$K = 16$	48.42	57.79	42.05	0.3392
DVC	strict Doppler	$\tau_v = 1.5$	48.36	57.86	42.01	0.3410
DVC	default	$\tau_v = 2.5$	48.87	58.22	42.28	0.3350
DVC	loose Doppler	$\tau_v = 3.5$	48.41	57.90	42.00	0.3431
DVC	no Doppler	∞	46.12	54.79	39.42	0.3820
Depth	no depth loss	–	45.90	54.70	38.91	0.3874
Depth	LiDAR only	–	47.85	57.06	41.32	0.3517
Depth	no radar depth	–	47.33	56.61	40.86	0.3598
Depth	default	–	48.87	58.22	42.28	0.3350

Anchor-consistent extent constraint versus proximity. The no-box-gate variant removes the extent test while retaining the radius, capacity, and training protocol. Under the matched setting, removing the extent test loses 1.89 mAP, 1.79 ODS, and 2.08 strict AP@1m while increasing mAVE from 0.3350 to 0.3635, which measures the contribution of object-shaped support beyond distance-only association.

G.6 Transferability of AGR and GCR to LXL

We further test whether object-level grounding transfers to a different radar–camera feature construction. LXL first forms a dense radar–camera BEV representation (Xiong et al., 2024). We adapt it in the same way as SGDet3D in Section 3.1. Specifically, dense image-to-BEV refinement is replaced by sparse object queries that sample multi-view image features at box-conditioned supports, while the LXL feature encoders are retained. The resulting single-pass model is denoted LXL-Adapted, with AGR and GCR placed at the same semantic retrieval and geometric refinement positions used in our main model. The two LXL-Adapted variants share the feature encoders, object-query decoder, inputs, and optimization protocol and differ only in whether AGR and GCR are enabled.

Because the two LXL-Adapted rows differ only in whether AGR and GCR are enabled, their performance gap measures the contribution of object-level grounding on the adapted architecture. Enabling both modules improves mAP and NDS by 3.24 and 3.09 points and reduces mATE and mAVE by 0.0439 and 0.6941, extending the recognition, localization, and motion gains to LXL feature construction.

Table A10: Transfer of AGR and GCR to LXL.

Method	AGR+GCR	mAP ↑	NDS ↑	mATE ↓	mAVE ↓
LXL	–	34.70	43.04	0.4972	1.9877
LXL-Adapted	–	35.12	44.09	0.4644	1.8241
LXL-Adapted	✓	**38.36**	**47.18**	**0.4205**	**1.1300**

H Reproducibility and Evaluation Protocol

All comparisons follow the official validation split and evaluator of each benchmark. OmniHD-Scenes averages class AP over 1, 2, 3, and 4 m center-distance thresholds, whereas ManTruckScenes

uses 0.5, 1, 2, and 4 m. Because the benchmarks also differ in class taxonomy, metric comparisons remain within each dataset.

Calibration robustness. Figure A4 visualizes the joint rotation–translation perturbations from Table A11. Both systems degrade as the disturbance grows, but SGDet3D++ remains above Control throughout. At $\pm 10^\circ$ and ± 1.0 m, SGDet3D++ retains 41.70 mAP and 52.76 ODS, exceeding Control by 10.67 and 12.56 points.

Adapted-baseline protocol. For every repository-trained baseline, we preserve the published backbone, fusion mechanism, decoder, pretrained initialization, data augmentation, optimizer, and learning-rate schedule. Adaptation is restricted to the benchmark-facing interface, which covers the official train/validation split, synchronized camera and radar inputs, three radar sweeps for radar-based methods, calibration and ego-pose metadata, the spatial range $[-60, -40, -3, 60, 40, 5]$ m, class mapping, and official evaluator. No component from SGDet3D++ is introduced into a competing method, so the comparison preserves method-specific capacity and optimization under a common data and evaluation contract. Table A10 separately evaluates transfer, with LXL following the SGDet3D adaptation described in Section 3.1. Its dense image-to-BEV refinement is replaced by sparse object-query decoding at box-conditioned supports. The two LXL-Adapted rows differ only in whether AGR and GCR are enabled. In the component studies, Base is the adapted SGDet3D control with AGR, GCR, and DVC disabled. Results taken from cited sources remain marked as reported values and are not mixed with repository-trained comparisons.

Table A11: Performance of SGDet3D (Control) and SGDet3D++ (Ours) under camera-extrinsic disturbances on OmniHD-Scenes.

Disturbance	mAP↑		ODS↑	
	Control	Ours	Control	Ours
None	40.41	48.87	47.81	58.22
$\pm 2^\circ$, ± 0.2 m	38.70	48.65	46.03	57.93
$\pm 5^\circ$, ± 0.5 m	36.17	46.46	43.36	56.14
$\pm 10^\circ$, ± 1.0 m	31.03	41.70	40.20	52.76

Training and model selection. All adapted methods are trained with their native optimization recipes under the same distributed training regime. Every repository-trained method uses the terminal checkpoint of its prescribed schedule, without ranking intermediate checkpoints by validation or test performance. The terminal checkpoint is evaluated unchanged on the full validation set, adverse-condition subsets, occlusion strata, calibration disturbances, and resource profiling. Controlled variants share the same preprocessing, radar sweep count, backbone, decoder depth, optimizer, schedule, and checkpoint rule. The main component study reports mean ± standard deviation over three independent runs, whereas focused sensitivity studies provide diagnostic comparisons.

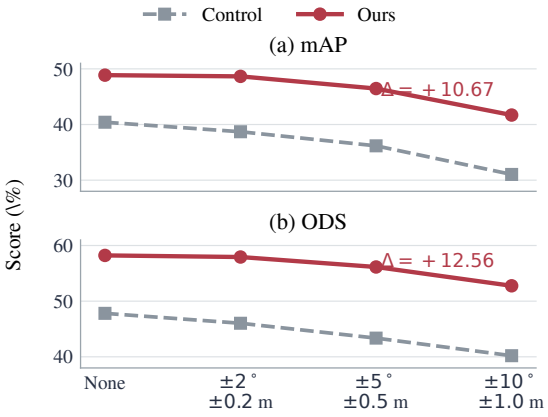


Figure A4: Performance under joint camera-extrinsic disturbances on OmniHD-Scenes. Values reproduce Table A11. SGDet3D++ exceeds Control in mAP and ODS at every tested level.

Resource-profiling protocol. All measurements use one NVIDIA RTX A6000 and batch size one. Latency runs each frozen checkpoint in evaluation mode without gradients on the first 1,296 ManTruckScenes validation samples. Five calls are discarded as warm-up, and the remaining 1,291 are accumulated. CUDA synchronization brackets the complete detector call, including scatter and post-processing executed inside it, while data-loader waiting and the CPU input pipeline are excluded. Latency is $1000/\mathrm{FPS}$, and FP16 follows each frozen configuration. Training memory is peak reserved memory during one forward–backward–update step after one warm-up step, cache clearing, and peak-statistic reset. Parameter counts include the full model. THOP measures GFLOPs on one evaluation forward.

I ADDITIONAL METHOD DETAILS

The reported models use 900 object queries and six decoder layers. The memory retains $M = 600$ hypotheses. AGR and GCR use association radius $\rho = 5\,\text{m}$ and box margin $\delta = 0.5\,\text{m}$, while GCR keeps at most $K_r = 8$ returns. DVC uses $\rho_{\text{DVC}} = 3\,\text{m}$ and radial-velocity tolerance $\tau_v = 2.5\,\text{m/s}$.

Architectural distinction from prior object-centric fusion. Object-centric hypotheses have different roles in the closest systems. CRAFT starts from camera proposals, performs polar-coordinate radar association, and uses cross-modal attention to predict a fusion score and proposal offsets (Kim et al., 2023a). RaCFormer's circular queries sample radar-enhanced perspective, camera BEV, and radar BEV features, with Doppler dynamics encoded implicitly in the radar BEV stream (Chu et al., 2025). CVFusion generates proposals from a fused radar–camera BEV map before second-stage aggregation of radar points, image views, and BEV grids (Zhong et al., 2025). In contrast, every SGDet3D++ decoder layer recomputes a yaw-aligned 3D return set from its evolving anchor *before* semantic retrieval. AGR pools this set, GCR preserves its individual returns, and DVC applies current Doppler as an explicit temporal accept–reject test. The prior hypotheses determine where to aggregate or refine features, whereas our model additionally verifies which local radar evidence may influence each of the three decisions. Published comparisons use the complete detectors because their proposal paths differ, while matched component attribution uses the common Base scaffold.

Identity-preserving evidence adaptation. AGR and GCR enter the object decoder through residual mappings. Writing e_i^ℓ for the pooled AGR evidence in Eq. 6, g_i^ℓ for the GCR context in Eq. 11, and n_i for the number of returns in the anchor-consistent radar support, the residual updates are

$$\widehat{f}_i^{\ell} = f_i^{\ell} + \Psi_{\text{AGR}}\big([f_i^{\ell} \| e_i^{\ell}]\big)\,, \qquad f_i^{\ell,+} = f_i^{\ell} + \mathbb{1}[n_i > 0]\Psi_{\text{GCR}}\big([f_i^{\ell} \| g_i^{\ell}]\big)\,. \tag{20}$$

Both residual mappings are initialized at zero. Hence $\widehat{f}_i^{\ell} = f_i^{\ell}$ and $f_i^{\ell,+} = f_i^{\ell}$ at the beginning of optimization, preserving the original object representation while the model learns how much radar evidence to incorporate. GCR also preserves the input feature for anchors with no anchor-consistent radar support.

Temporal confidence and replacement. Before DVC is applied, historical hypotheses are retained only when the timestamp gap satisfies $m_t = \mathbb{1}[|t_t - t_{t-1}| \le T_{\max}]$, with $T_{\max} = 2$ seconds. For the first M inherited hypotheses, temporal confidence combines the previous estimate $c_{t-1,k}$ and current classification logit $s_{t,k}$ as

$$\widetilde{c}_{t,k} = \max(\gamma c_{t-1,k}, \sigma(s_{t,k}))\,, \qquad \gamma = 0.6, \quad k \le M, \tag{21}$$

and retains the M feature–anchor pairs with the highest confidence. DVC then verifies each pose-aligned historical pair while preserving its correspondence order. Using the contradiction indicator z_k from Eq. 14, the pre-attention update is

$$(\bar{f}_{t,k}, \bar{b}_{t,k}) = (1 - z_k)(f_{t,k}^{\text{hist}}, b_{t,k}^{\text{hist}}) + z_k(f_{t,k}^{\text{cur}}, b_{t,k}^{\text{cur}}), \quad k = 1, \ldots, M. \tag{22}$$

Thus a contradiction selects the same-index current slot, whereas a missing return gives $z_k = 0$ and preserves the historical one.

Decision roles. AGR conditions projected image retrieval, GCR attends to individual supporting returns for box refinement, and DVC replaces pose-aligned history whose nearest current return contradicts its predicted radial motion. Because the M historical slots retain their order, a contradiction selects the same-index current slot before temporal attention without an extra matching objective. The separation between semantic retrieval, geometric refinement, and temporal verification keeps radar from becoming a generic proximity prior.