CONSOLIDATE THEN REFINE: 3D OBJECT DETECTION WITH LIDAR AND 4D RADAR IN ADVERSE WEATHER

Anonymous authors
Paper under double-blind review

ABSTRACT

Adverse weather degrades LiDAR and camera observations, whereas 4D radar remains relatively stable but suffers from limited spatial resolution. Existing LiDAR–radar detectors mainly fuse the two modalities at the scene level, which couples their representations and prevents modality-specific evidence from being fully exploited. We develop a consolidate-then-refine framework that preserves modality-specific representations and associates their evidence at the object level. This framework consolidates independently generated LiDAR and radar candidates into unified object hypotheses and progressively refines their representations and geometry. On the official K-Radar v2.1 benchmark, our detector achieves 66.0% 3D mAP@0.5, surpassing the state of the art by 9.5 points.

1 INTRODUCTION

3D object detection is an important task in autonomous-driving perception. Fully autonomous driving requires vehicles to operate under diverse weather conditions. LiDAR provides accurate 3D geometric measurements that are essential for object localization and boundary estimation. Rain, fog, and snow, however, can degrade LiDAR returns and introduce additional noise, making it difficult to capture complete object geometry (Heinzler et al., 2019; Hahner et al., 2021; 2022). Cameras also undergo substantial degradation under these conditions (Nayar & Narasimhan, 1999; Sakaridis et al., 2021) and may not reliably compensate for object information missing from LiDAR observations. Consequently, 3D detection in adverse weather calls for a sensing strategy that provides more stable observations and complements LiDAR.

Radar provides a valuable source of observations for this purpose. It is less affected by adverse weather (Kutila et al., 2018; Waldschmidt et al., 2021); 4D radar further provides richer spatial observations, enabling its use in 3D object detection (Paek et al., 2022). However, the spatial resolution of automotive radar, particularly its angular resolution, is generally lower than that of LiDAR, making it difficult for radar alone to fully characterize an object's 3D geometry (Patole et al., 2017; Giuffrida et al., 2023). Combining LiDAR's precise geometry with radar's stable observations has thus become an important approach to object detection in adverse weather. However, LiDAR and radar differ substantially in their observation forms and information distributions, and fully exploiting useful information from both modalities remains challenging.

Existing LiDAR–radar fusion methods commonly establish cross-modal connections at an early stage based on local spatial relationships and use radar information to enhance LiDAR representations (Wang et al., 2023; Chae et al., 2024; Huang et al., 2025). Even for the same object, however, useful LiDAR and radar responses need not occur at the same position, and the forms of their responses also differ substantially. Cross-modal association based solely on spatial proximity may therefore introduce incorrect correspondences or omit spatially offset complementary evidence. Adverse weather further changes the response characteristics of the two modalities, exacerbating information mismatches and omissions during early fusion. Moreover, tightly coupling the two branches at this stage can relegate one modality to an auxiliary role and limit the ability of each branch to learn a distinct modality-specific representation.

For higher-level fusion, common methods aggregate features from different modalities into a unified BEV representation and use it as the basis for object prediction (Liu et al., 2023; Paek & Kong, 2025). Such scene-level fusion couples information from different modalities before object prediction and may dilute modality-specific information (Yang et al., 2022). Under adverse weather,

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.

imbalanced degradation across modalities can further reduce the performance of tightly coupled systems (Gao et al., 2023). Query-based fusion can instead retain independent modality representations, from which queries separately retrieve information relevant to potential objects (Chen et al., 2023; Yan et al., 2023). Generating candidates from the current frame's feature responses and using them to initialize object queries provides an object-location prior and reduces the localization burden on subsequent decoding (Bai et al., 2022). However, selecting candidates after multimodal features have been aggregated at dense BEV locations still couples the modalities before query initialization (van Geerenstein et al., 2024). Independently generating candidates from each modality and concatenating them into a query set, on the other hand, can introduce multiple queries for the same object and thereby increase query competition (Xie et al., 2023; Gao et al., 2024).

Based on these observations, we develop a consolidate-then-refine framework for object-level LiDAR–radar detection in adverse weather. Separate LiDAR and radar branches first encode their observations into modality-specific multiscale features. We tailor the radar encoder for sparse observations on a regular lattice to capture their height-wise structure and relative power distribution. Each modality then produces object-center candidates. Center voting consolidates these candidates in a common physical space into unified object hypotheses, which may be supported by one or both modalities. Using each hypothesis center as a reference, the first decoder separately retrieves object-relevant information from the two feature branches and predicts a preliminary 3D box. Conditioned on this estimate, the second decoder retrieves further object evidence from both modalities and refines the prediction. By retaining modality-specific representations and aggregating available observations around progressively refined object geometry, the framework improves the stability of 3D object detection in adverse weather.

Our main contributions are:

- We consolidate independently generated LiDAR and radar center candidates into object hypotheses through cross-modal center voting. This object-level association preserves modality-specific feature streams while providing unified hypotheses for decoding.
- We refine each consolidated object hypothesis with a two-stage geometry-conditioned decoder. The first stage retrieves modality-specific features around the hypothesis center and estimates a preliminary 3D box; the second uses the resulting box to gather additional evidence and refine the prediction.
- We evaluate our method on the official K-Radar v2.1 benchmark, where it achieves 66.0% 3D mAP@0.5 and surpasses the previous best by 9.5 points. Ablation studies quantify the contributions of candidate consolidation and progressive refinement, while evaluation across weather conditions characterizes the method's performance under adverse weather.

2 Related Work

2.1 Multimodal Fusion for 3D Object Detection

Cross-modal interaction in multimodal 3D detection can operate on either scene representations or object representations. When constructing scene representations, information from one modality can augment the detection input of another (Vora et al., 2020; Wang et al., 2021), and features from different modalities can interact during encoding (Li et al., 2022b; Wang et al., 2023). After encoding, modality-specific BEV features can be fused into a unified BEV feature map (Liang et al., 2022), from which a detection head can directly produce predictions. At the object-representation level, query-based detectors use object queries to represent potential objects and aggregate object-relevant evidence from modality features (Bai et al., 2022; Yang et al., 2022; Chen et al., 2023; Yan et al., 2023; Xie et al., 2023). The updated queries provide instance-specific representations for classification and box regression. Predicted centers or box proposals can also provide geometric references for subsequent feature extraction and guide object-related feature sampling (Zhou et al., 2022; Nguyen et al., 2022). Existing query-based multimodal detectors have mainly focused on LiDAR–camera fusion, while query-based LiDAR–radar detection under adverse weather remains less explored.

2.2 LiDAR–Radar Detection under Adverse Weather

LiDAR–radar detection under adverse weather commonly uses stable radar observations to complement degraded LiDAR information. At the sparse-voxel level, 3D-LRF aggregates neighboring radar features and gates their injection into LiDAR features according to weather conditions (Chae et al., 2024). REL uses radar to generate virtual LiDAR points that supplement degraded geometry (Tong et al., 2026). DLR-Fusion models radar power and Doppler separately and updates the features of both modalities through iterative interaction (Chae et al., 2025). L4DR combines radar foreground selection with bidirectional interaction in shared pillars and uses gating to fuse multi-scale BEV features (Huang et al., 2025). ASF applies cross-sensor attention across BEV feature maps from camera, LiDAR, and radar (Paek & Kong, 2025). These methods primarily rely on spatial association within voxel or pillar neighborhoods and direct fusion on BEV feature maps. When LiDAR and radar responses to the same object are spatially offset, local association can miss useful information; direct fusion on BEV feature maps also makes it difficult to explicitly aggregate object evidence distributed across different locations.

3 Method

3.1 Framework Overview

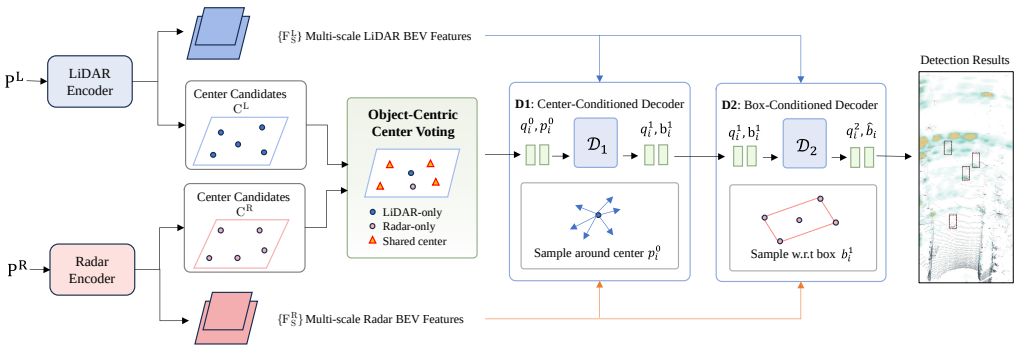

Figure 1: Overview of the proposed consolidate-then-refine framework for LiDAR–radar 3D object detection in adverse weather. Center voting consolidates LiDAR and radar candidates into unified object hypotheses, which are progressively refined through center- and box-conditioned decoding.

The overall framework comprises modality-specific encoders, unified object hypothesis construction, and a progressive geometry-conditioned decoder. The LiDAR and radar encoders separately produce multiscale BEV features and object center candidates. Center voting aggregates candidates that point to the same object in a unified physical coordinate system, forming unified object hypotheses. The support set of each hypothesis can contain candidates from one or both modalities. The supporting candidates of each hypothesis are pooled to initialize an object query, while the two multiscale feature streams retain their branch identities and serve as modality-specific feature memories for subsequent decoding.

The query representations need to aggregate object evidence from the LiDAR and radar features. Global cross-attention requires each query to search the entire scene for object evidence; numerous unrelated locations make the relevant features harder to isolate. The progressively refined object estimate, however, already provides spatial priors that can constrain retrieval. The object hypothesis center supplies an initial position, and the preliminary 3D box further characterizes the object extent and orientation. We therefore adopt multiscale deformable attention (Zhu et al., 2021), using the object center and preliminary box in sequence to guide local sampling and query updates in the two decoder layers.

Figure 1 illustrates the complete data flow. The inputs are a LiDAR point cloud P^L and sparse radar power measurements P^R on a regular 3D lattice, synchronized and calibrated to a unified coordinate system. The two encoders construct multiscale BEV feature sets F_S^L and F_S^R and predict the corresponding object center responses. The center-voting module aggregates the two sets of

center candidates into unified object hypotheses represented by centers p_i^0 and query features q_i^0. The center-conditioned decoder uses p_i^0 as its reference to update the query and predict a preliminary 3D box b_i^1. The box-conditioned decoder then uses b_i^1 to organize a second round of feature sampling, further updates the query, and outputs the final 3D box $\hat{b}_i$. The preliminary box is parameterized as $b_i^1 = (x_i^1, y_i^1, z_i^1, l_i^1, w_i^1, h_i^1, \theta_i^1)$, and the final box uses the same parameterization.

3.2 Multimodal Feature Encoding

The LiDAR point cloud is first partitioned into regular voxels, and the point attributes within each voxel are mean-pooled to form sparse 3D features. Sparse 3D convolutions aggregate local geometry through progressive downsampling, and height compression projects the 3D features onto the BEV plane. A 2D backbone then expands the receptive field in the BEV plane and produces the multiscale feature set F_S^L (Yan et al., 2018). The high-resolution features preserve fine spatial structure, while the low-resolution features encode broader scene context.

The radar branch takes sparse power measurements on a regular 3D lattice as input (Paek et al., 2022). Existing LiDAR–radar methods commonly represent these lattice observations as 3D points augmented with power and then employ point-set, voxel, or pillar encoders (Chae et al., 2024; Huang et al., 2025). These encoders are designed for irregular geometric point sets and derive local representations from point distributions and set aggregation. Our radar encoder instead exploits this lattice organization by grouping observations at the same planar position (x, y) into an ordered sequence along height. It combines the resulting vertical representation with column-level geometric, occupancy, and power statistics to form a feature at each occupied BEV location. A 2D backbone then produces the multiscale feature set F_S^R.

After multiscale encoding, each modality predicts a class-specific center heatmap H^m and within-cell offset O^m on a high-resolution BEV grid, where $m \in \{L, R\}$ (Yin et al., 2021). Heatmap peaks, corrected by the offsets, form center candidates in continuous physical coordinates. Section 3.3 combines candidates from both modalities in a class-agnostic voting field to establish query locations, while the multiscale feature sets F_S^m remain modality-specific memories for query updates.

3.3 Object-Centric Center Voting

LiDAR and radar center candidates for the same object are generally close but do not coincide exactly. We associate them through a spatial voting field that accounts for local representations and spatial consistency. For each candidate j, we sample multiscale features from its source modality and local features from both modalities at its BEV location. A source-specific voting head first predicts a vote center and a 2D spatial scale from the candidate's native representation. A joint correction head then updates both quantities using the bimodal features and predicts an objectness weight. We denote the corrected center, scale, and objectness by μ_j, σ_j, and α_j, respectively. The corrected vote center represents the object center supported by the candidate and may differ from its original BEV location.

Each candidate then casts a Gaussian vote $G_j(u)$ over planar locations u, centered at μ_j and shaped by σ_j. Weighted by α_j, the votes from all valid LiDAR and radar candidates form a class-agnostic spatial voting field through the probabilistic union

$$V(u) = 1 - \prod_{j \in \mathcal{C}} \big(1 - \alpha_j G_j(u)\big), \tag{1}$$

where $\mathcal{C}$ denotes the set of valid candidates. Votes from spatially consistent candidates reinforce one another where they overlap, producing a shared peak even when their original candidate locations differ. Local maxima of the voting field define object modes $\hat{u}_i$. For each mode, r_{ij} measures the relative contribution of candidate j:

$$r_{ij} = \frac{\alpha_j G_j(\hat{u}_i)}{\sum_{\ell \in \mathcal{C}} \alpha_\ell G_\ell(\hat{u}_i)}. \tag{2}$$

Let f_j denote the joint bimodal feature at the candidate location. The vote centers and candidate features are aggregated using r_{ij} to obtain the initial center and query representation of mode i

$$p_i^0 = \sum_{j \in \mathcal{C}} r_{ij} \mu_j, \qquad q_i^0 = \sum_{j \in \mathcal{C}} r_{ij} f_j + e_i. \tag{3}$$

Here, p_i^0 is the aggregated object center, q_i^0 is the corresponding query representation, and e_i encodes the composition of contributing modalities, the effective number of contributing candidates, and the aggregate spatial scale. Each valid mode is then converted into a unified object hypothesis (p_i^0, q_i^0), which initializes one object query for the subsequent progressive decoder. Unused slots in the fixed query set remain masked during decoding.

3.4 Progressive Geometry-Conditioned Refinement

The unified object hypotheses (p_i^0, q_i^0) generated by center voting initialize the object queries, which are progressively updated by two geometry-conditioned decoder layers. In each layer, the current query first aggregates local context separately from the LiDAR and radar multiscale BEV features. The decoder then fuses both contexts with the query representation, applies inter-query self-attention, and passes the result through a feed-forward network. The first layer predicts a preliminary 3D box from the planar center estimate, and the second uses the geometry of this box to further refine the query and box parameters.

Both decoder layers use the same form of multiscale feature aggregation. For decoder layer $t \in \{1, 2\}$ and modality $m \in \{L, R\}$, let $F_s^m(x)$ denote the interpolated feature from BEV level s at physical location x. The context from modality m is

$$z_i^{m,t} = \sum_{h=1}^{H} W_h^{m,t} \left[\sum_{s=1}^{S} \sum_{k=1}^{K} \omega_{i,h,s,k}^{m,t} F_s^m \left(r_{i,k}^t + \Delta p_{i,h,s,k}^{m,t} \right) \right]. \tag{4}$$

Here, h, s, and k index the attention head, feature level, and sampling point, respectively. The query predicts attention weights $\omega_{i,h,s,k}^{m,t}$ and sampling offsets $\Delta p_{i,h,s,k}^{m,t}$, while $W_h^{m,t}$ is the output projection of the h-th attention head. The reference locations of the two layers are

$$r_{i,k}^1 = p_i^0, \qquad r_{i,k}^2 = \mathcal{G}_k\left(b_i^1\right), \tag{5}$$

where $\mathcal{G}_k(b_i^1)$ denotes the k-th sampling anchor derived from the preliminary box. The first layer aggregates multiscale LiDAR and radar features separately around p_i^0 (Zhou et al., 2022). After the LiDAR and radar contexts are fused with q_i^0, inter-query self-attention and a feed-forward network produce q_i^1, which predicts the class, quality, and preliminary 3D box b_i^1. The second layer incorporates the 3D geometry of b_i^1 into the query positional encoding and updates the query using bimodal context sampled around the geometric anchors defined by the box. Prediction heads applied to the fused query q_i^2 produce updated class and quality predictions together with residual corrections to the center, dimensions, and orientation of b_i^1, yielding the final 3D box $\hat{b}_i$.

3.5 Training Objectives

The training objective supervises the candidate representations and object queries at two levels. The total loss is

$$\mathcal{L} = \sum_{m \in \{L,R\}} \mathcal{L}_{\text{center}}^m + \mathcal{L}_{\text{vote}} + \sum_{t=1}^{2} \lambda_t \mathcal{L}_{\text{det}}^{(t)} + \lambda_{\text{dn}} \mathcal{L}_{\text{dn}}. \tag{6}$$

The loss $\mathcal{L}_{\text{center}}^m$ supervises the center candidates of modality $m \in \{L, R\}$. For each annotated box whose BEV extent contains valid observations from modality m, we construct the corresponding center-heatmap and within-cell-offset targets (Yin et al., 2021) and apply Gaussian focal loss and Smooth L1 loss, respectively. For candidates within the BEV extent of an annotated box, the target center supervises both the source-specific vote and the jointly corrected vote center. Both terms use a Gaussian negative log-likelihood based on the predicted center and spatial scale. Candidate objectness is supervised with binary cross-entropy balanced between positive and negative samples, and the class-agnostic voting field uses Gaussian focal loss. Together, these terms constitute $\mathcal{L}_{\text{vote}}$.

At the query level, each decoder layer performs Hungarian matching using costs for class, box geometry, and rotated 3D IoU. The detection loss $\mathcal{L}_{\text{det}}^{(t)}$ comprises focal classification loss, an IoU-supervised quality loss, and box-regression losses over the center, dimensions, orientation, BEV corners, and vertical overlap. The first-layer detection loss provides intermediate supervision, while the second-layer loss supervises the final predictions. The second-layer loss additionally includes

a rotated 3D adaptation of DIoU (Zheng et al., 2020) to constrain box overlap and center distance. During training, we construct denoising queries from perturbed ground-truth boxes and supervise class, quality, and box-parameter recovery at both decoder layers (Li et al., 2022a). The loss $\mathcal{L}_{\mathrm{dn}}$ is used only during training.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

K-Radar is a large-scale multisensor autonomous-driving dataset that covers diverse road and weather conditions. It provides spatially and temporally calibrated 4D radar, LiDAR, camera, and RTK-GPS/IMU measurements (Paek et al., 2022). We use the v2.1 annotations and evaluate Sedan and the combined Bus/Truck class on the official test set. To examine different overlap requirements, we report BEV AP, 3D AP, and the corresponding two-class mAP at IoU 0.3 and 0.5. All methods are evaluated using the same 3D-box coordinate convention, a prediction confidence threshold of 0.3, and the revised 41-point AP protocol. We additionally correct the box-coordinate conversion in the evaluation pipeline, so the reported values may differ from those in prior work; Appendix A.3 provides details.

The comparisons include LiDAR-only PointPillars (Lang et al., 2019), a LiDAR–radar adaptation of BEVFusion (BEVFusion†) (Liang et al., 2022), L4DR (Huang et al., 2025), and ASF (Paek & Kong, 2025). BEVFusion† uses independent PointPillars branches for LiDAR and radar and feeds their fused BEV features to the detection head. For L4DR and ASF, we use the weights released by their authors; ASF additionally uses camera. All predictions are re-evaluated on the same test set. The full model and all ablation variants use the same LiDAR–radar input range and synchronized geometric augmentation. They are trained for 20 epochs using AdamW with an initial learning rate of 2×10^{-4}, a batch size of 4, and FP32 precision.

4.2 MAIN RESULTS

Table 1: Detection results on K-Radar v2.1 at IoU 0.3 (%). Best results are in bold and second-best results are underlined. L, R, and C denote LiDAR, 4D radar, and camera, respectively.

Method	Input	Sedan BEV AP	Sedan 3D AP	Bus/Truck BEV AP	Bus/Truck 3D AP	BEV mAP	3D mAP
PointPillars	L	78.28	<u>77.78</u>	73.90	66.84	76.09	72.31
BEVFusion†	L+R	78.38	77.68	<u>78.74</u>	<u>74.71</u>	<u>78.56</u>	<u>76.19</u>
L4DR [AAAI 2025]	L+R	<u>79.06</u>	76.42	72.54	67.60	75.80	72.01
ASF [NeurIPS 2025]	L+C+R	78.02	75.27	75.66	68.96	76.84	72.12
Ours	L+R	**87.28**	**84.74**	**82.53**	**77.35**	**84.91**	**81.05**

BEV mAP and 3D mAP are computed from unrounded per-class AP.

Table 2: Detection results on K-Radar v2.1 at IoU 0.5 (%). Other notation follows Table 1.

Method	Input	Sedan BEV AP	Sedan 3D AP	Bus/Truck BEV AP	Bus/Truck 3D AP	BEV mAP	3D mAP
PointPillars	L	77.64	56.53	67.22	49.46	72.43	53.00
BEVFusion†	L+R	<u>77.83</u>	57.05	<u>73.13</u>	51.35	<u>75.48</u>	54.20
L4DR [AAAI 2025]	L+R	76.45	<u>58.08</u>	69.27	<u>54.86</u>	72.86	<u>56.47</u>
ASF [NeurIPS 2025]	L+C+R	75.19	57.82	71.36	47.40	73.27	52.61
Ours	L+R	**84.40**	**68.50**	**79.49**	**63.43**	**81.95**	**65.96**

Tables 1 and 2 compare the detectors under two overlap criteria. Our method achieves the best result in every reported metric at both thresholds. At IoU 0.3, it reaches 84.91 BEV mAP and 81.05 3D mAP, improving over BEVFusion†, the best-performing baseline, by 6.35 and 4.86 points,

respectively. At IoU 0.5, it attains 81.95 BEV mAP and 65.96 3D mAP, exceeding BEVFusion† by 6.47 points in BEV and L4DR by 9.49 points in 3D.

Compared with BEV IoU, 3D IoU additionally constrains vertical center and height, making accurate 3D localization more sensitive to errors in the vertical dimension. Under adverse weather, degraded LiDAR geometry and noisy 4D radar elevation measurements make these vertical constraints harder to satisfy. Accordingly, 3D AP under the stricter IoU threshold provides a more sensitive test of progressive box refinement.

L4DR and ASF couple modality features within fixed spatial partitions and produce 3D boxes from the resulting fused representation in a single pass. L4DR exchanges LiDAR and radar features within shared pillars, then combines the raster-aligned BEV branches through multiscale gating (Huang et al., 2025). ASF applies cross-sensor attention within fixed local BEV patches and reconstructs a single fused map (Paek & Kong, 2025). In both cases, cross-modal responses at nearby but nonidentical locations must be reconciled implicitly during scene-level encoding, and the resulting box estimate receives no subsequent geometric refinement. ASF additionally uses camera and remains 13.35 points below our 3D mAP at IoU 0.5.

Our method addresses these constraints at the object level. Center voting consolidates nearby modality-specific candidates without requiring cell-wise coincidence. The center-conditioned decoder predicts the initial geometry, and the box-conditioned decoder then retrieves features relative to that estimate and corrects it. At IoU 0.5, the 3D AP gains over L4DR are 10.42 points for Sedan and 8.57 points for Bus/Truck, showing that the benefit extends across both categories.

4.3 Ablation Studies

Table 3: Component ablations on K-Radar v2.1 at IoU 0.5 (%). Each variant changes one design choice while retaining the remaining components; Δ mAP is measured relative to the full model and computed from unrounded values.

Ablation variant	Sedan 3D AP	Bus/Truck 3D AP	3D mAP	Δ mAP
Fused-heatmap initialization	64.96	58.20	61.58	-4.38
Full-resolution global attention	65.90	61.07	63.49	-2.48
Center-conditioned D2 sampling	67.91	62.86	65.38	-0.58
PointPillars-style radar encoder	65.88	60.15	63.01	-2.95
Full model	**68.50**	**63.43**	**65.96**	–

Table 3 evaluates four targeted variants, each modifying one design choice while retaining the remaining components of the full model. To examine object-hypothesis construction, the fused-heatmap variant aligns and concatenates the LiDAR and radar BEV features, predicts a shared center heatmap and subcell offsets, and initializes the 64 decoder queries from its strongest local maxima. This replaces the modality-specific center heads and center voting while keeping both decoder layers unchanged. The resulting 4.38-point 3D mAP decrease supports separate candidate generation followed by object-level consolidation. Appendix Table B1 further characterizes the localization changes introduced by the two voting stages.

Two variants evaluate the spatial priors used for decoder retrieval. The full-resolution global-attention variant replaces the modality-specific deformable readers in both decoder layers with cross-attention over every spatial cell at both scales of the LiDAR and radar BEV pyramids. Center voting, two-stage prediction, and the modality-specific memories remain unchanged. This variant reaches 63.49 3D mAP, a 2.48-point decrease, supporting the use of spatially constrained retrieval over unrestricted full-scene attention. The center-conditioned D2 variant retains the deformable readers but uses only the center updated by D1 to position the second-stage samples. Its 0.58-point decrease isolates the additional gain from using the preliminary dimensions and orientation in D2. The larger decrease from global attention supports restricting retrieval with the current object geometry, while the D2 comparison identifies the further contribution of box extent and orientation.

The final variant evaluates the radar representation. It replaces the proposed vertical-column encoder with a PointPillars-style PFN that max-pools raw (x, y, z, p) observations within vertical pillars, while retaining radar center prediction, center voting, and both decoder layers. The resulting 2.95-

point decrease in 3D mAP supports preserving the lattice-organized vertical power structure before BEV encoding.

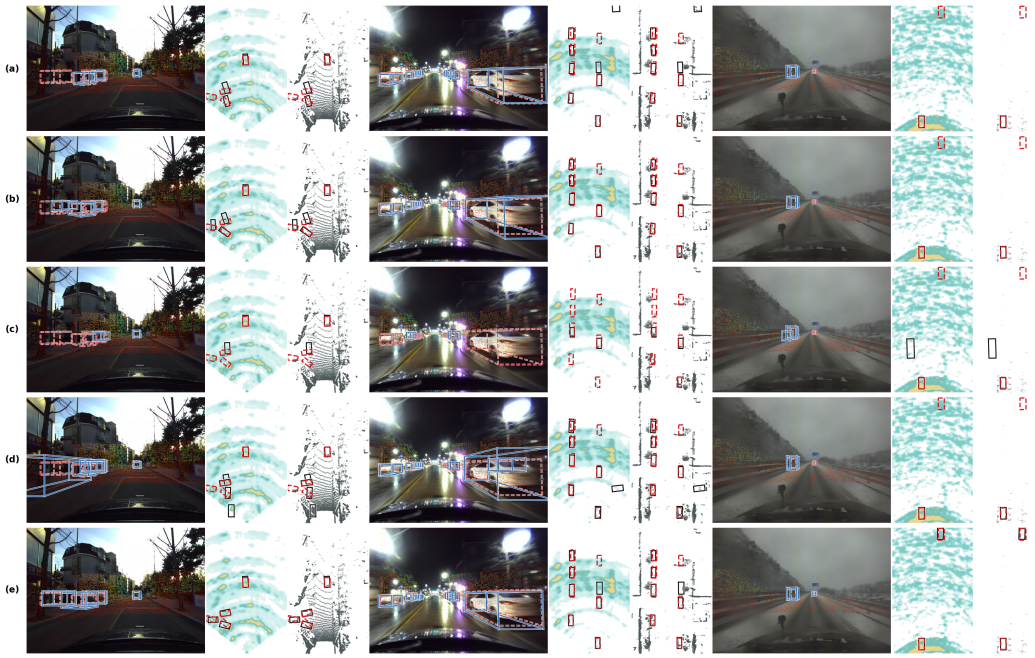

Figure 2: Same-frame qualitative comparison on the K-Radar v2.1 test set. (a) PointPillars, (b) BEVFusion†, (c) ASF, (d) L4DR, and (e) our method. Columns show Normal, Rain, and Heavy snow; each group contains the forward image, radar BEV, and LiDAR BEV. Blue solid cuboids in the images and black solid footprints in BEV denote predictions; red dashed cuboids and footprints denote annotations. All methods use a confidence threshold of 0.3.

4.4 Performance across Weather Conditions

Table 4: Sedan AP by weather on K-Radar v2.1 at IoU 0.5 (%). Each cell gives BEV AP / 3D AP. Best results are in bold and second-best results are underlined; Total is evaluated on the complete evaluation set.

Method	Input	Normal	Overcast	Fog	Rain	Sleet	Light snow	Heavy snow	Total
PointPillars	L	76.5/55.8	83.9/56.2	89.9/79.0	78.9/56.1	65.0/45.7	88.9/62.1	56.5/45.0	77.6/56.5
BEVFusion†	L+R	74.7/55.4	85.8/56.7	94.8/84.1	78.7/56.7	65.6/51.3	90.8/65.6	56.7/43.7	77.8/57.1
L4DR	L+R	75.0/56.2	86.4/59.0	93.8/83.5	79.9/56.9	67.0/54.1	85.4/65.1	57.6/53.1	76.4/58.1
ASF	L+C+R	74.2/56.3	86.5/60.5	94.9/80.4	68.9/52.4	68.1/48.6	88.3/62.3	61.0/51.3	75.2/57.8
Ours	L+R	**81.1/64.5**	**92.1/75.1**	**97.2/88.2**	**85.7/71.8**	**74.7/63.6**	**93.0/69.9**	**67.9/60.8**	**84.4/68.5**

Table 4 reports Sedan performance separately for the seven weather labels. These subsets also differ in road scenes, traffic composition, and target distributions, so their absolute AP values should not be interpreted as a controlled ranking of weather severity. The high AP attained by all methods in Fog and Light snow, relative to Normal, illustrates this distinction. We therefore focus on comparisons among methods within the same subset; the frame and object counts underlying the subsets are reported in Appendix Table A2.

Across all seven weather labels, our method ranks first in both BEV and 3D AP. Its 3D AP margin over the best-performing baseline ranges from 4.1 to 14.9 points. The advantage remains substantial under precipitation, reaching 14.9 points in Rain, 9.5 points in Sleet, and 7.7 points in Heavy snow. The 8.2-point gain in Normal shows that the improvement also extends beyond adverse conditions.

The margins are smaller in Fog and Light snow, where the best-performing baselines already reach 84.1 and 65.6 3D AP, respectively; our method nevertheless raises these values to 88.2 and 69.9. The improvement is therefore consistent across the full range of K-Radar weather labels.

The gains are consistently larger in 3D AP than in BEV AP. The contrast is most pronounced in Rain, where the margins are 14.9 and 5.8 points, and in Overcast, where they are 14.6 and 5.6 points. As discussed in Section 4.2, 3D AP adds constraints on vertical center and height to the ground-plane footprint used by BEV AP. The consistently larger 3D margins show that the advantage is more pronounced when vertical center and height are included in the evaluation. Figure 2 compares representative detection results under Normal, Rain, and Heavy snow. Appendix Table B2 examines whether the trend across weather conditions extends to Bus/Truck, while Appendix Section D provides additional weather examples and failure cases.

5 Conclusion

We presented a consolidate-then-refine framework for LiDAR–radar 3D object detection in adverse weather. Separate modality branches preserve LiDAR and radar representations, while center voting consolidates their candidates into object hypotheses. A center-conditioned decoder retrieves modality-specific evidence and predicts a preliminary 3D box, whose geometry then guides a second retrieval and refinement stage. On the K-Radar v2.1 benchmark, our method achieves the highest BEV and 3D mAP under both IoU criteria, improving the previous best 3D mAP by 9.49 points at IoU 0.5. It also ranks first for Sedan across all seven weather labels. Component ablations support the contributions of object-level consolidation, geometry-conditioned retrieval, and lattice-aware radar encoding.

AI Use Statement

Generative AI tools were used to assist with language editing of this manuscript, including grammar checking, improving readability, suggesting alternative phrasings, and checking terminology and presentation consistency across sections. The authors reviewed and revised all AI-assisted edits and verified the technical statements, numerical results, citations, and scientific conclusions. The authors take full responsibility for the final content of the manuscript.

References

Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1090–1099, 2022.

Yujeong Chae, Hyeonseong Kim, and Kuk-Jin Yoon. Towards robust 3D object detection with LiDAR and 4D Radar fusion in various weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15162–15172, 2024.

Yujeong Chae, Heejun Park, Hyeonseong Kim, and Kuk-Jin Yoon. Doppler-aware LiDAR-RADAR fusion for weather-robust 3D detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27197–27208, 2025.

Xuanyao Chen, Tianyuan Zhang, Yue Wang, Yilun Wang, and Hang Zhao. FUTR3D: A unified sensor fusion framework for 3D detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 172–181, 2023.

Xinyu Gao, Zhijie Wang, Yang Feng, Lei Ma, Zhenyu Chen, and Baowen Xu. Benchmarking robustness of AI-enabled multi-sensor fusion systems: Challenges and opportunities. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 871–882, 2023.

Yulu Gao, Yifan Sun, Xudong Ding, Chuyang Zhao, and Si Liu. EASE-DETR: Easing the competition among object queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17282–17291, 2024.

Luigi Giuffrida, Guido Masera, and Maurizio Martina. A survey of automotive radar and lidar signal processing and architectures. *Chips*, 2(4):243–261, 2023.

Martin Hahner, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Fog simulation on real LiDAR point clouds for 3D object detection in adverse weather. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15283–15292, 2021.

Martin Hahner, Christos Sakaridis, Mario Bijelic, Felix Heide, Fisher Yu, Dengxin Dai, and Luc Van Gool. LiDAR snowfall simulation for robust 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16364–16374, 2022.

Robin Heinzler, Philipp Schindler, Jürgen Seekircher, Werner Ritter, and Wilhelm Stork. Weather influence and classification with automotive LiDAR sensors. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1527–1534, 2019.

Xun Huang, Ziyu Xu, Hai Wu, Jinlong Wang, Qiming Xia, Yan Xia, Jonathan Li, Kyle Gao, Chenglu Wen, and Cheng Wang. L4DR: LiDAR-4DRadar fusion for weather-robust 3D object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(4):3806–3814, 2025.

Matti Kutila, Pasi Pyykönen, Hanno Holzhüter, Michel Colomb, and Pierre Duthon. Automotive LiDAR performance verification in fog and rain. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1695–1701, 2018.

Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. PointPillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12697–12705, 2019.

Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M. Ni, and Lei Zhang. DN-DETR: Accelerate DETR training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13619–13627, 2022a.

Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V. Le, Alan Yuille, and Mingxing Tan. DeepFusion: Lidar-camera deep fusion for multi-modal 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17182–17191, 2022b.

Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. BEVFusion: A simple and robust LiDAR-camera fusion framework. In *Advances in Neural Information Processing Systems 35*, pp. 10421–10434, 2022.

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 *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2774–2781, 2023.

Shree K. Nayar and Srinivasa G. Narasimhan. Vision in bad weather. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pp. 820–827, 1999.

Duy-Kien Nguyen, Jihong Ju, Olaf Booij, Martin R. Oswald, and Cees G. M. Snoek. BoxeR: Box-attention for 2D and 3D transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4773–4782, 2022.

Dong-Hee Paek and Seung-Hyun Kong. Availability-aware sensor fusion via unified canonical space. In *Advances in Neural Information Processing Systems 38*, pp. 89259–89277, 2025.

Dong-Hee Paek, Seung-Hyun Kong, and Kevin Tirta Wijaya. K-Radar: 4D Radar object detection for autonomous driving in various weather conditions. In *Advances in Neural Information Processing Systems 35*, pp. 3819–3829, 2022.

Sujeet M. Patole, Murat Torlak, Dan Wang, and Murtaza Ali. Automotive radars: A review of signal processing techniques. *IEEE Signal Processing Magazine*, 34(2):22–35, 2017.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10765–10775, 2021.

Tianxu Tong, Xinrun Liu, Hongmin Liu, and Bin Fan. Towards accurate 3D object detection in adverse weather by leveraging 4D radar for LiDAR geometry enhancement. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(12):9548–9556, 2026.

Mathijs R. van Geerenstein, Felicia Ruppel, Klaus Dietmayer, and Dariu M. Gavrila. Multimodal object query initialization for 3D object detection. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 12484–12491, 2024.

Sourabh Vora, Alex H. Lang, Bassam Helou, and Oscar Beijbom. PointPainting: Sequential fusion for 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4604–4612, 2020.

Christian Waldschmidt, Juergen Hasch, and Wolfgang Menzel. Automotive radar—from first efforts to future systems. *IEEE Journal of Microwaves*, 1(1):135–148, 2021.

Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. PointAugmenting: Cross-modal augmentation for 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11794–11803, 2021.

Yingjie Wang, Jiajun Deng, Yao Li, Jinshui Hu, Cong Liu, Yu Zhang, Jianmin Ji, Wanli Ouyang, and Yanyong Zhang. Bi-LRFusion: Bi-directional LiDAR-Radar fusion for 3D dynamic object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13394–13403, 2023.

Yichen Xie, Chenfeng Xu, Marie-Julie Rakotosaona, Patrick Rim, Federico Tombari, Kurt Keutzer, Masayoshi Tomizuka, and Wei Zhan. SparseFusion: Fusing multi-modal sparse representations for multi-sensor 3D object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17591–17602, 2023.

Junjie Yan, Yingfei Liu, Jianjian Sun, Fan Jia, Shuailin Li, Tiancai Wang, and Xiangyu Zhang. Cross modal transformer: Towards fast and robust 3D object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18268–18278, 2023.

Yan Yan, Yuxing Mao, and Bo Li. SECOND: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu, and Li Zhang. DeepInteraction: 3D object detection via modality interaction. In *Advances in Neural Information Processing Systems 35*, pp. 1992–2005, 2022.

Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3D object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11784–11793, 2021.

Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-IoU loss: Faster and better learning for bounding box regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(7):12993–13000, 2020.

Zixiang Zhou, Xiangchen Zhao, Yu Wang, Panqu Wang, and Hassan Foroosh. CenterFormer: Center-based transformer for 3D object detection. In *Computer Vision – ECCV 2022*, pp. 496–513. Springer Nature Switzerland, 2022.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

SUPPLEMENTARY MATERIAL

The supplementary material provides implementation details, additional quantitative analyses, mechanism visualizations, qualitative examples, and a discussion of limitations.

A IMPLEMENTATION DETAILS

A.1 K-RADAR DATASET, ANNOTATIONS, AND EVALUATION SPLIT

K-Radar is a large-scale multisensor dataset and benchmark for 3D object detection. It provides approximately 35K synchronized frames from 58 sequences, including full Doppler–range–azimuth–elevation (DRAE) 4D radar tensors, calibrated high-resolution LiDAR, surround stereo cameras, and RTK-GPS. The sequences cover urban and suburban roads, alleyways, and highways in normal weather and adverse conditions such as fog, rain, sleet, and snow (Paek et al., 2022). The official conditional evaluation groups the data into seven weather labels: Normal, Overcast, Fog, Rain, Sleet, Light snow, and Heavy snow.

The official K-Radar repository provides three annotation revisions, summarized in Table A1.

Table A1: Official K-Radar annotation revisions.

Version	Revision
v1.0	Original 3D bounding-box annotations.
v2.0	Improved box dimensions, added missing annotations, and revised track IDs.
v2.1	Added object-level LiDAR and 4D radar visibility labels.

In the released files used here, v2.0 and v2.1 contain the same box geometry and yaw; v2.1 adds the visibility field. The original K-Radar benchmark evaluates Sedan with v1.0 labels over $x \in [0, 72]$ m, $y \in [-6.4, 6.4]$ m, and $z \in [-2, 6]$ m. The official wide benchmark expands the lateral and vertical ranges to $y \in [-16, 16]$ m and $z \in [-2, 7.6]$ m, evaluates Sedan and Bus/Truck, and uses v2.0 labels. We use the official v2.1 annotations and train–test split while retaining the wide benchmark's region and two evaluated classes.

Applying the two-class mapping, wide region of interest, and nonempty-frame filter to the unaugmented splits produces 14,386 training frames with 39,634 ground-truth boxes and 13,727 test frames with 35,456 ground-truth boxes. Table A2 gives their weather and class distributions; each entry reports train/test counts.

Table A2: K-Radar v2.1 train/test distribution under our data filtering.

Weather	Frames	Sedan boxes	Bus/Truck boxes
Normal	5,488 / 5,415	17,604 / 15,988	2,814 / 2,763
Overcast	411 / 414	1,146 / 1,107	106 / 151
Fog	1,511 / 1,445	2,103 / 1,947	0 / 0
Rain	1,835 / 1,424	7,502 / 5,978	179 / 120
Sleet	2,297 / 2,206	1,890 / 1,622	1,357 / 1,132
Light snow	1,115 / 1,103	1,542 / 1,360	660 / 540
Heavy snow	1,729 / 1,720	1,571 / 1,611	1,160 / 1,137
Total	14,386 / 13,727	33,358 / 29,613	6,276 / 5,843

Sedan boxes outnumber Bus/Truck boxes by approximately 5.31:1 in training and 5.07:1 in testing. Neither split contains Bus/Truck boxes in the Fog subset.

A.2 Sensor Inputs and Preprocessing

Each raw K-Radar frame contains a dense Doppler–range–azimuth–elevation (DRAE) radar tensor with a nominal shape of $64 \times 256 \times 107 \times 37$, occupying approximately 260 MB in 32-bit floating-point format. The complete multisensor release is approximately 15 TB. To make full-dataset training tractable within our storage and I/O budget, we use the preprocessed sparse radar representation distributed by K-Radar.

The official preprocessing aggregates the Doppler dimension and retains the top 1% of measurements by power on a regular 0.4-m Cartesian lattice. We keep the resulting (x, y, z, p) records, where p denotes aggregated radar power, and crop them to the wide region of interest.

For LiDAR, we use the synchronized OS2-64 long-range point clouds with coordinates and return intensity. The preprocessed point-cloud product aligns LiDAR with radar using the dataset's translation parameters. The calibration files provide the horizontal components, while the official loader fixes the vertical component to 0.7 m. During training, LiDAR points, radar measurements, and annotations receive the same rigid augmentation before the final region-of-interest crop.

A.3 Evaluation Protocol and Coordinate Conversion

We represent a 3D box as $\mathbf{b} = (x, y, z, l, w, h, \theta)$, where x, y, and z are the forward, lateral, and vertical coordinates, respectively, and θ is the physical heading in the horizontal plane.

The original K-Radar evaluator contains two inconsistencies that affect AP and 3D IoU. It constructs a precision array at 41 recall positions but averages every fourth value, thereby mixing 41-point sampling with 11-point averaging. It also computes vertical overlap with $z_{\text{center}} = 1.0$, although the exported vertical coordinate represents the box center. The official repository subsequently added a revised evaluator that averages all 41 precision values and uses $z_{\text{center}} = 0.5$, giving the vertical extent $[z - h/2, z + h/2]$. We adopt both corrections.

Our protocol audit identified a separate heading-conversion issue in the KITTI-style serialization. Under the center mapping $(x_{\text{cam}}, y_{\text{cam}}, z_{\text{cam}}) = (y, z, x)$ and the evaluator's camera-plane rotation convention, the physically consistent KITTI heading r_y is

$$r_y = \theta - \frac{\pi}{2}. \tag{7}$$

Directly copying the K-Radar heading θ into the KITTI heading field rotates the serialized box by 90° relative to its physical orientation. Table A3 distinguishes the official evaluator revisions from this additional conversion.

Table A3: Evaluation conventions and their provenance.

Aspect	Applied convention	Provenance
AP averaging	All 41 recall positions	Official K-Radar revision
Vertical overlap	$z_{\text{center}} = 0.5$	Official K-Radar revision
Heading export	$r_y = \theta - \pi/2$	Our protocol audit

The validation routine provided with the official K-Radar evaluator constructs predictions directly from the ground-truth boxes. Because each pair has coincident centers, rotating both boxes by the same erroneous 90° preserves unit IoU and does not expose the heading mismatch. We instead verify the conversion by matching the transformed physical corners for boxes with nonzero translation and arbitrary heading.

Unless otherwise stated, we retain predictions with confidence strictly above 0.3 and accumulate each weather subset independently. Predictions from all methods are evaluated using the same 41-point AP calculation, center-based vertical overlap, and physically consistent yaw conversion.

A.4 Modality-Specific Encoders

The LiDAR branch uses a voxel size of $(0.075, 0.08, 0.2)$ m, and its sparse 3D backbone reduces the planar resolution by a factor of eight. The outputs of the two BEV stages used as decoder memories

have 128 and 256 channels at resolutions of 0.60×0.64 m and 1.20×1.28 m, respectively. For center prediction, both stage outputs are brought to the finer resolution, projected to 256 channels each, concatenated, and reduced to 128 channels before the heatmap and offset heads.

The radar encoder retains the native 0.4-m lattice and the native vertical bins within the input region of interest. Two one-dimensional convolutions encode each vertical profile, after which two-head attention pooling fuses it with the column descriptor. The resulting 128-dimensional tokens are scattered to their planar cells and processed by a 2D encoder–decoder. Its 0.4-m and 0.8-m outputs, both with 128 channels, form the radar memories. The radar heatmap and offset heads operate directly on the 0.4-m output.

A.5 Center Voting

Candidate extraction first takes the maximum response over classes at each location, applies 3×3 local-maximum filtering, and retains at most 64 candidates from each modality. For every candidate, the voting module samples both feature levels from each modality branch, together with the local center responses and measurement support. A source-specific voting head first predicts a center and spatial scale from the candidate's native representation. A joint correction head then uses the bimodal representation to correct that center and scale and to predict objectness. The source-specific vote and joint correction have maximum displacements of 8 m and 4 m, respectively.

The voting field covers the input region of interest on a 0.4-m lattice. Its anisotropic Gaussian scales are constrained to $[0.2, 3.2]$ m, and each vote is evaluated within three standard deviations of its predicted center. We retain the 64 highest 3×3 local maxima as object modes and mask unused query positions when fewer modes are available. During training, a candidate is assigned to a ground-truth object when its offset-corrected location lies inside the oriented BEV box. The source-specific vote and joint correction are optimized with Gaussian negative log-likelihoods based on their predicted scales. Objectness uses a balanced binary loss, and the voting field uses Gaussian focal loss. Their weights are given in Section A.7.

A.6 Progressive Decoder

The decoder uses 64 queries, a hidden dimension of 128, eight attention heads, a 512-dimensional feed-forward network, and a dropout rate of 0.1. Each modality-specific deformable-attention module reads two feature levels and uses five sampling locations per head and level. Before sampling, each BEV level is projected to the decoder dimension and split across attention heads; F_s^m in Eq. (4) denotes the resulting per-head value feature. For the object queries initialized by voting, all five first-layer sampling locations share the consolidated center and receive learned metric offsets with a scale of 0.5 m.

The second layer places five fixed sampling anchors on the preliminary box footprint, one at its center and four at its corners. Learned corrections are applied in the local box frame before rotation by the predicted yaw.

A.7 Training and Inference

We train the complete network end to end for 20 epochs with a batch size of 4 on one NVIDIA GeForce RTX 4090 GPU. Training uses FP32, PyTorch 2.1.2, and CUDA 12.1. We use AdamW with an initial learning rate of 2×10^{-4}, a minimum learning rate of 2×10^{-7}, $(\beta_1, \beta_2) = (0.9, 0.999)$, and a weight decay of 0.01. Bias and normalization parameters are excluded from weight decay. The learning rate is linearly warmed up for 500 iterations from one third of its initial value and then follows cosine annealing. Gradients are clipped to a maximum norm of 35. We use a random lateral flip with probability 0.5, a global rotation sampled from $[-10^\circ, 10^\circ]$, point shuffling, and a training seed of 2023.

Hungarian matching is performed between the valid object queries and ground-truth boxes. Table A4 lists the matching-cost and detection-loss weights. Classification uses sigmoid focal loss, while the quality branch is supervised by the 3D IoU of the matched box. The box objectives include center, height, log-scale dimension, directional yaw, BEV-corner, vertical-IoU, and 3D-IoU terms. Within the dimension loss, length, width, and height are weighted by $(1, 1, 2)$. The first decoder layer

receives auxiliary supervision with a weight of 0.5. We additionally use three groups of denoising queries; their combined loss has a weight of 1.0, and negative denoising queries are weighted by 0.5. The heatmap and offset losses of both modality-specific center heads have unit weights, as do the source-specific vote, joint correction, objectness, and voting-field objectives.

Table A4: Weights used for Hungarian matching and query detection losses.

Component	Matching cost	Detection loss	Component	Matching cost	Detection loss
Classification	1.0	1.0	Quality	–	1.0
BEV center	2.0	2.0	BEV corners	–	0.5
Height	1.0	1.0	Vertical IoU	–	1.0
Dimensions	1.0	1.0	3D IoU	2.0	1.0
Yaw	0.5	1.0			

At inference, the confidence for each class is the geometric mean of its classification probability and the predicted quality. We retain at most 64 predictions per frame. All results for our model use the final training checkpoint and the evaluation threshold specified in Section A.3.

B Additional Quantitative Analysis

B.1 Center-Voting Diagnostics

We inspect center voting with the final checkpoint on a reproducible sample of 2,746 frames, approximately 20% of the 13,727-frame evaluation set. We generate a random permutation of the official test indices with seed 2023 and take the first 2,746 indices without replacement. This sample is used only for the internal center-voting diagnostics below; the reported AP results use the complete evaluation set. A positive candidate is an offset-corrected heatmap candidate located inside an oriented ground-truth BEV box, and its localization error is the Euclidean distance to the corresponding box center. The audit contains 17,828 positive candidates.

Table B1: Localization error of positive center candidates on the reproducible 2,746-frame test sample. P90 is the error below which 90% of the positive candidates fall.

Center state	Mean (m)	Median (m)	P90 (m)
Input candidate	1.014	0.519	2.330
After source-specific vote	0.665	0.375	1.595
After joint correction	0.447	0.264	0.945

The source-specific vote reduces the mean error from 1.014 to 0.665 m, and the joint correction further reduces it to 0.447 m. The median error falls from 0.519 to 0.264 m, while P90 falls from 2.330 to 0.945 m, showing consistent reductions in typical and high-error cases. After both steps, 73.1% of the positive candidates are closer to their assigned centers than they were at input. By source, the mean errors at input, after the source-specific vote, and after joint correction are 0.767/0.439/0.432 m for LiDAR and 1.235/0.868/0.460 m for radar. The joint correction therefore produces a substantially larger correction for the initially coarser radar candidates.

Across the same sample, the voting field produces 115,630 valid modes, or 42.1 per frame. The effective number of contributing candidates, $(\sum_j r_{ij}^2)^{-1}$, averages 2.24 per mode, so the responsibility mass of each mode is concentrated on a few candidates. The mean responsibility masses across all modes are 0.460 for LiDAR and 0.540 for radar.

B.2 Bus/Truck Results by Weather

Table B2 complements the Sedan results in Table 4 with Bus/Truck AP under the same protocol. Fog has no Bus/Truck annotations, so its AP is undefined. The Total column reports AP evaluated on the complete evaluation set.

Table B2: Bus/Truck AP by weather on K-Radar v2.1 at IoU 0.5 (%). Each cell gives BEV AP / 3D AP. Best results are in bold and second-best results are underlined; Total is evaluated on the complete evaluation set.

Method	Input	Normal	Overcast	Fog	Rain	Sleet	Light snow	Heavy snow	Total
PointPillars	L	67.0/48.9	82.9/76.2	–/–	0.1/0.1	82.4/76.5	89.9/74.3	50.6/24.6	67.2/49.5
BEVFusion†	L+R	**71.1**/54.6	81.6/71.3	–/–	1.0/1.0	82.6/76.6	93.9/**80.4**	69.0/19.3	73.1/51.4
L4DR	L+R	69.8/58.1	**85.2/85.2**	–/–	11.6/4.9	70.3/65.1	87.5/67.8	70.6/**42.1**	69.3/54.9
ASF	L+C+R	64.7/48.8	81.3/73.9	–/–	7.6/3.0	60.4/48.5	**97.0**/77.1	84.6/39.2	71.4/47.4
Ours	L+R	68.3/**62.5**	84.8/81.4	–/–	**12.5/5.2**	**91.6/88.3**	89.9/78.2	**91.2**/38.9	**79.5/63.4**

At the aggregate level, the advantage observed for Sedan extends to the less frequent Bus/Truck category. Our method attains 79.5 BEV AP and 63.4 3D AP on the complete evaluation set, improving over the best-performing baseline by 6.4 and 8.5 points, respectively. Across weather subsets, the method achieves the highest 3D AP in Normal, Rain, and Sleet and ranks second in Overcast and Light snow. The largest margin occurs in Sleet, where it exceeds the best-performing baseline by 11.7 points.

Bus/Truck has 5,843 test instances, compared with 29,613 for Sedan, and its distribution varies substantially across weather conditions (Appendix Table A2). Fog contains no Bus/Truck annotations, while Rain contains only 120; the corresponding conditional AP is therefore more sensitive to individual predictions and scene composition. The method achieves the highest BEV AP in Heavy snow at 91.2, while its 38.9 3D AP trails L4DR by 3.2 points. The large BEV–3D gap points to errors in vertical center and height estimation. Representative classification and localization failures are discussed in Appendix Section D.

C Mechanism Visualizations

C.1 Full-Scene and Modality-Specific Center Evidence

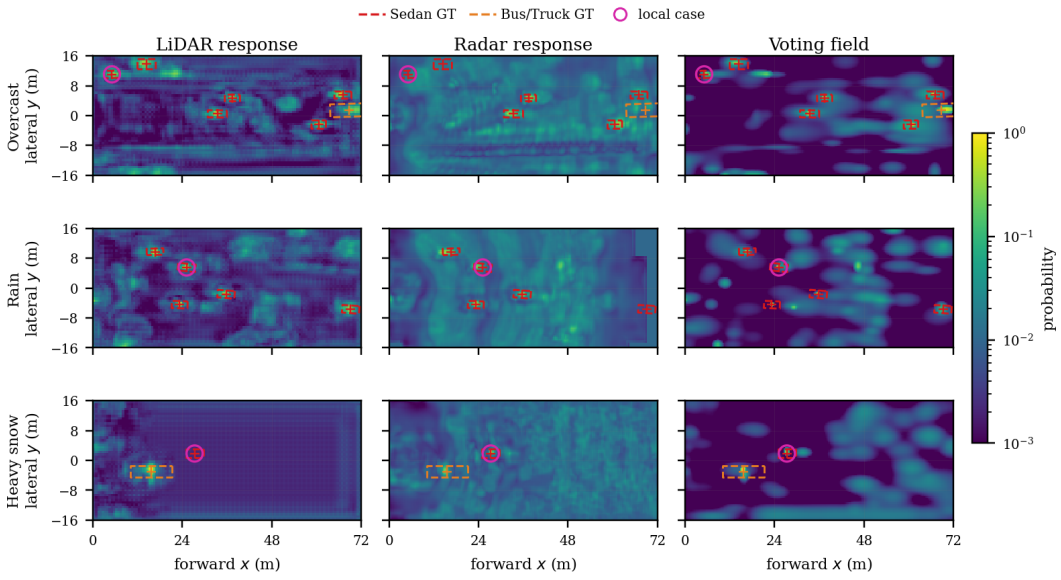

Figure C1: Full-scene center evidence under Overcast, Rain, and Heavy snow (top to bottom). Columns show the maximum LiDAR and radar class responses and the class-agnostic voting field. Red and orange boxes mark Sedan and Bus/Truck annotations, respectively; magenta circles mark the local cases in Figure C2. All panels share axes and a logarithmic probability scale.

At the scene level, the modality-specific maps differ in response strength and spatial concentration around annotated objects, while center voting forms compact object modes from their candidates (Figure C1). The three circled cases illustrate three representative evidence patterns (Figure C2). In Overcast, both branches produce nearby candidates, with a much stronger LiDAR score. In Heavy snow, only the radar branch produces a same-class candidate within 2 m of the annotation. In Rain, the LiDAR candidate cell lies close to the annotated center while the radar candidate cell is spatially displaced; the voting field forms a local maximum near the annotation. Together, the examples show how voting carries forward clear single-modality evidence and consolidates displaced dual-modality evidence. Replacing separate candidate generation and voting with fused-heatmap initialization reduces 3D mAP by 4.38 points in Table 3.

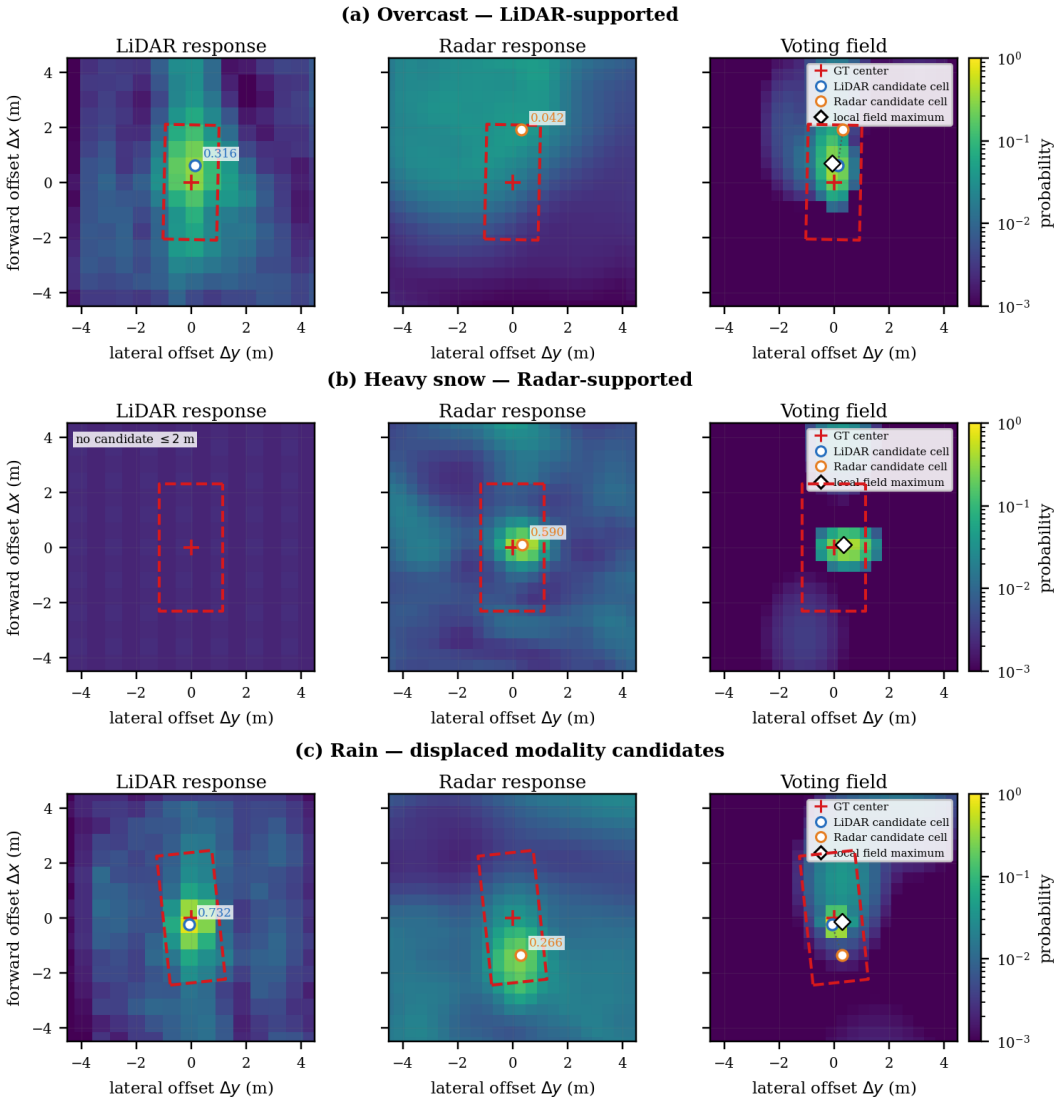

Figure C2: Local modality-specific center evidence and the resulting voting field. Row headings identify the weather condition and evidence pattern. Columns give the class-specific LiDAR response, class-specific radar response, and class-agnostic voting field. The horizontal axis is lateral offset and the vertical axis is forward offset. Red crosses mark annotated geometric centers. Blue and orange circles identify the native heatmap cells that generated the displayed LiDAR and radar candidates; adjacent numbers give their raw probabilities, and an absent circle indicates no same-class candidate within 2 m of the annotation. White diamonds mark the nearest local maxima of the voting field. The dashed rectangle denotes the ground-truth BEV footprint. All maps use the same logarithmic probability scale.

C.2 Progressive Geometry-Conditioned Sampling

Progressive geometry conditioning changes the spatial reference used for modality-specific feature retrieval. D1 samples around the voted center and predicts a preliminary box; D2 uses that box geometry to update the sampling locations for refinement. Figure C3 visualizes this transition for one query in each of three scenes under Normal, Rain, and Heavy snow. Each panel overlays the sampling pattern for one decoder stage and modality on the corresponding raw sensor observations in a target-centered BEV. In the Heavy-snow Bus/Truck example, the D2 samples span the annotated longitudinal extent in both modality branches.

LiDAR and radar retain distinct sampling patterns over their respective observations at both stages, with the evolving object hypothesis providing a shared geometric reference. The ablations in Table 3 support these roles: full-resolution global attention lowers 3D mAP by 2.48 points, while replacing box-conditioned sampling in D2 with center-conditioned sampling lowers it by 0.58 points.

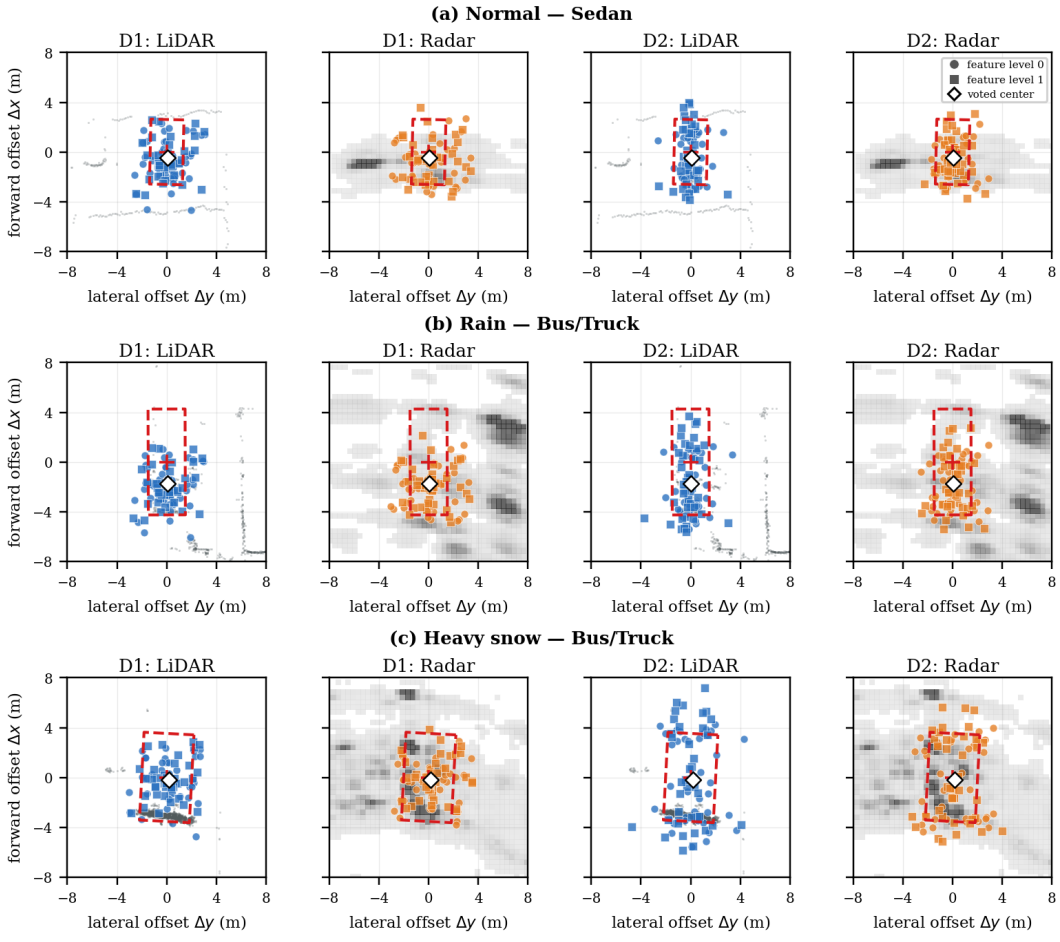

Figure C3: Progressive geometry-conditioned sampling under Normal, Rain, and Heavy snow, shown from top to bottom. Columns show the LiDAR and radar sampling patterns for D1 and D2. Gray backgrounds show raw LiDAR returns or sparse radar power projected into the target-centered BEV. Blue and orange markers denote LiDAR and radar samples; circles and squares distinguish the two feature levels. White diamonds mark the voted query center, and red dashed rectangles denote ground-truth BEV footprints.



Figure D1: Qualitative results under Normal, Overcast, Fog, Rain, Sleet, Light snow, and Heavy snow. Each row contains two scenes from the official test set. Each scene shows the forward image as environmental context, followed by the radar and LiDAR BEV observations used by the detector. Every annotation has a same-class prediction at 3D IoU ≥ 0.5; displayed predictions use a confidence threshold of 0.3. Blue solid cuboids in the images and black solid footprints in BEV denote predictions; red dashed cuboids and footprints denote annotations.

D Qualitative Examples

D.1 Representative Weather Conditions

The forward images provide environmental context for the seven weather conditions, while the paired radar and LiDAR BEV panels show the observations used for detection (Figure D1). Under Normal and Overcast conditions, LiDAR preserves clear object geometry across multiple lanes and depths, complemented by radar responses around the same traffic participants. Under Sleet and Heavy snow, the LiDAR panels show noisier or incomplete returns, while radar retains discernible responses near annotated vehicles. Fog and Heavy snow produce the strongest loss of visual contrast, and Rain adds wet-road reflections and glare. These adverse-weather scenes include matched targets beyond 60 m.

These examples reflect the sensing motivation introduced in Section 1. LiDAR contributes precise geometry from reliable returns; radar contributes stable object evidence as LiDAR quality varies. By preserving modality-specific representations, the detector consolidates the available center evidence and progressively refines each object hypothesis. The resulting predictions maintain aligned BEV footprints across near and distant Sedan and Bus/Truck instances, including targets approaching 72 m. These examples complement the AP results across weather conditions in Table 4, while Section C visualizes the corresponding center evidence and sampling behavior.

D.2 Representative Failure Cases

We group representative errors into four categories: class confusion, missed detections, inaccurate boxes, and false positives. The forward images in Figures D2–D5 provide scene context, while the radar and LiDAR BEV panels show the detector inputs around the analyzed instance. Each row shows only the relevant prediction, annotation, or both. Blue solid cuboids in the images and black solid footprints in BEV denote predictions; red dashed cuboids and footprints denote annotations. All displayed predictions are retained at the 0.3 confidence threshold used for the official test evaluation.

Class assignment can remain incorrect even when localization is accurate. In Figure D2, both predictions reach 3D IoU above 0.5 but assign the wrong class. The upper row confuses a Sedan with Bus/Truck at 0.893 BEV IoU and 0.862 3D IoU, whereas the lower row makes the reverse error at 0.719 BEV IoU and 0.608 3D IoU. The lower scene contains a plow-equipped utility truck, whose atypical geometry makes it a useful corner case for the Bus/Truck–Sedan class boundary.

Weather-degraded observations can leave a target unsupported by precise geometry. The two missed targets in Figure D3 lie at about 33–35 m, and their LiDAR footprints contain almost no usable returns. Radar produces a local response in each scene, yet neither response is well aligned with the target footprint. The Heavy-snow response is diffuse across the surrounding cells, and the strongest Light-snow response is displaced to the right of the target. This spatial mismatch weakens the consolidated object evidence, and neither target produces a detection above the 0.3 confidence threshold.

Radar background ambiguity can produce spurious hypotheses when LiDAR geometry is weak. Figure D4 contains Sedan predictions in Rain and Heavy snow with confidence scores of 0.837 and 0.780; both have zero BEV overlap with every annotation. The Rain example contains sparse, unstructured LiDAR returns and a weak, dispersed radar response near the predicted footprint. The Heavy-snow example contains no local LiDAR returns, while a broad high-power radar band crosses the lower part of the predicted footprint. These cases expose the background clutter and coarse spatial structure of sparse radar power. Diffuse or extended responses can resemble partial object support and generate false center evidence.

Geometric regression can fail along different spatial dimensions even when the class is predicted correctly. The two examples in Figure D5 separate vertical and planar errors. In Heavy snow, the Bus/Truck footprint reaches 0.784 BEV IoU, while a 1.29 m vertical-center error reduces 3D IoU to 0.326. In Sleet, a 1.60 m planar-center error reduces Sedan BEV IoU to 0.458 and 3D IoU to 0.422; its vertical-center error is 0.13 m. Together, the cases distinguish two sources of low 3D IoU. The first preserves planar overlap but misestimates vertical position, whereas the second shifts the footprint in BEV.

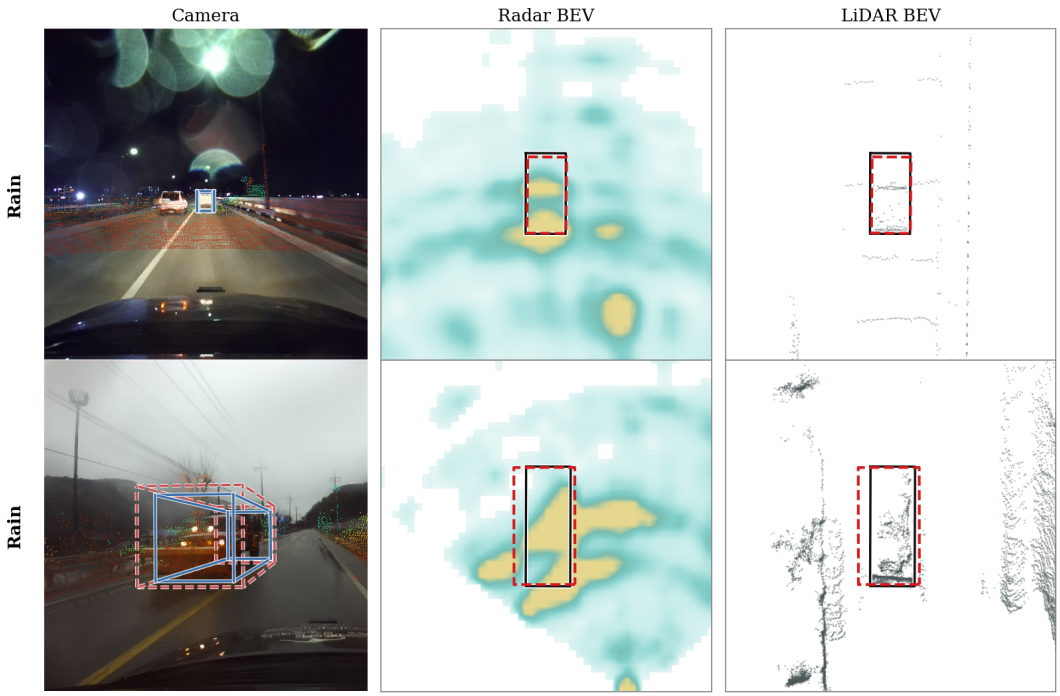

Figure D2: Classification errors in Rain. The upper row shows a ground-truth Sedan predicted as Bus/Truck, and the lower row shows a ground-truth Bus/Truck predicted as Sedan. The respective BEV/3D IoUs are 0.893/0.862 and 0.719/0.608.

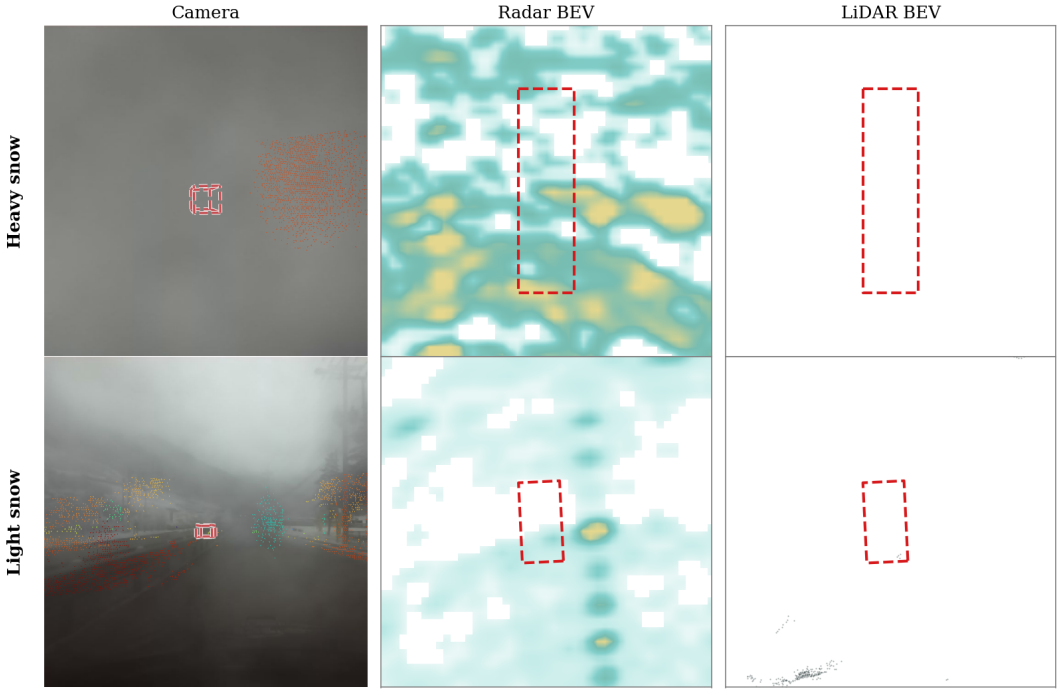

Figure D3: Missed detections at the 0.3 confidence threshold. The upper row shows a Bus/Truck in Heavy snow, and the lower row shows a Sedan in Light snow.

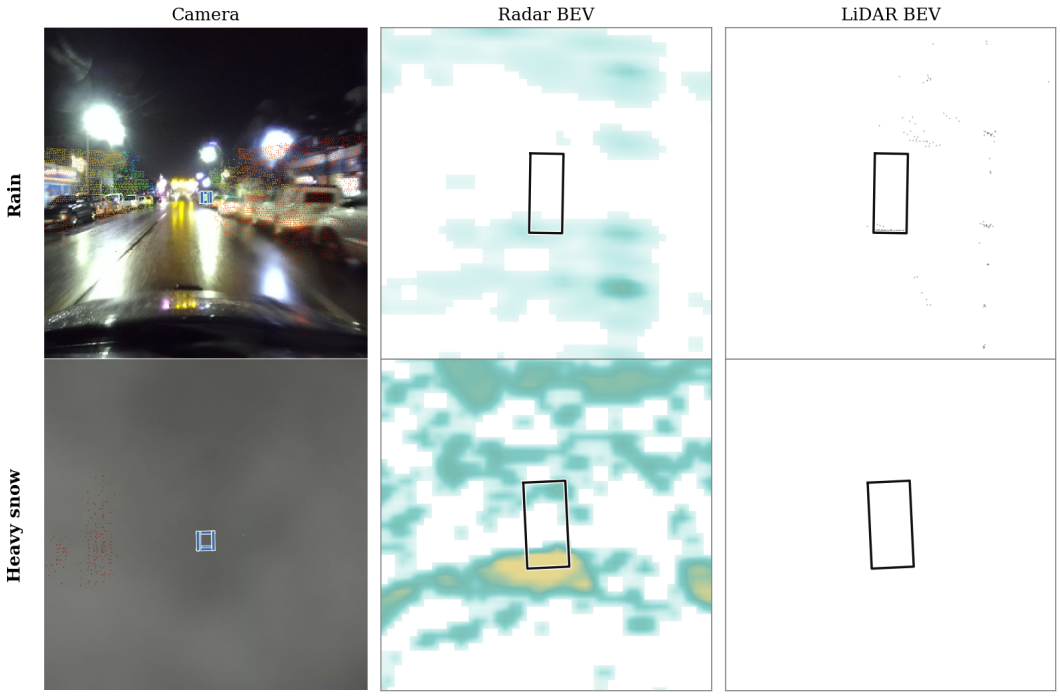

Figure D4: False positives in Rain and Heavy snow. The upper and lower rows show Sedan predictions with zero BEV overlap with any annotation.

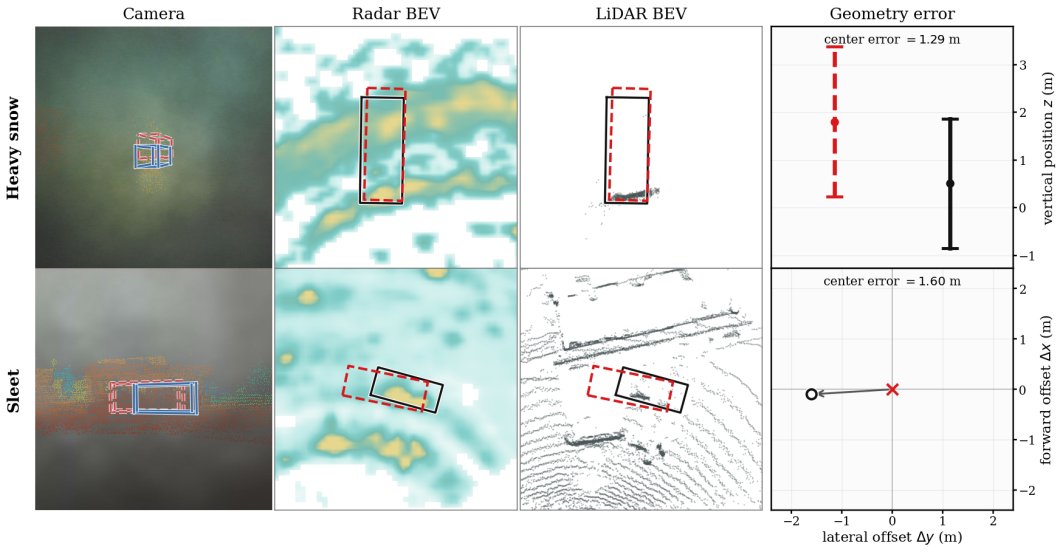

Figure D5: Box errors for a Bus/Truck in Heavy snow and a Sedan in Sleet. The upper row shows vertical displacement, and the lower row shows planar displacement. The right column compares the vertical extents in the upper row and the planar centers in the lower row.

E LIMITATIONS

Our framework relies on spatially coherent evidence from at least one modality. When adverse weather leaves LiDAR returns extremely sparse and radar power is diffuse or dominated by clutter, the center heads may miss an object or produce a spurious candidate, leaving center voting without a reliable hypothesis for refinement. Figures D3 and D4 show these two failure modes. Center consolidation is performed in BEV, whereas vertical position and height are recovered during decoding. Ambiguous elevation evidence can therefore preserve a plausible BEV footprint while yielding poor 3D overlap, as illustrated in Figure D5.

The detector operates on a single synchronized frame and assumes accurate LiDAR–radar calibration in a shared coordinate system. It cannot use evidence from adjacent frames to compensate for transient observation loss, while timing offsets or extrinsic drift can separate corresponding center candidates and displace the feature samples used during refinement.