GEOMETRY-CONDITIONED SHARED-SLICE FUSION AND GAUSSIAN SET REFINEMENT FOR ADVERSE-WEATHER 3D DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Gated cameras provide range-selective observations through distinct temporal integration windows, offering complementary cues to LiDAR and RGB for 3D object detection in adverse weather. Existing multimodal fusion methods such as SAMFusion, however, commonly encode the near, middle, and far gated measurements as pseudo-RGB channels, conflating temporally distinct observations and obscuring their range-selective responses. This representation-level limitation is compounded at inference, where conventional suppression determines redundancy from confidence and geometric overlap without accounting for the physical evidence supporting each prediction. We present a representation-to-set framework that addresses both deficiencies. Geometry-Conditioned Shared-Slice Aggregation (GCSA) processes the gated measurements through independent passes of a shared visual encoder and estimates spatial slice contributions from projected depth, projection validity, and cross-view consistency. Slice-Geometry-Conditioned Gaussian Set Refinement (SG-GSR) subsequently derives candidate reliability and pairwise compatibility from gated-slice agreement, local LiDAR support, and depth consistency to modulate same-class Gaussian score decay. SG-GSR preserves box geometry and class labels and requires no additional detector training. Compared with SAMFusion, the complete framework improves condition-specific Overall 3D AP40 by 2.22–4.50 points across daytime, nighttime, snow, and fog, with gains across all three distance intervals.

1 INTRODUCTION

Reliable 3D perception under adverse weather is constrained by modality-dependent failure modes. LiDAR provides metric localization, yet its returns become increasingly sparse with distance and may be attenuated by atmospheric particles. RGB cameras supply dense semantic context but remain vulnerable to low illumination, glare, and scattering. Gated near-infrared (NIR) cameras complement both modalities by actively integrating scene responses over controlled temporal windows, producing range-selective observations that retain structure when passive image contrast deteriorates Bijelic et al. (2020); Julca-Aguilar et al. (2021); Walia et al. (2022). Effective LiDAR–RGB–gated fusion therefore offers a principled route toward robust adverse-weather detection.

Sensor complementarity, however, does not guarantee a physically appropriate representation. As illustrated schematically in Figure 1(a), the gated measurements correspond to distinct temporal windows whose response regions shift with range, while the relative active return generally weakens at longer ranges. Nevertheless, existing tri-modal pipelines commonly inherit the three-channel RGB interface and stack the near, middle, and far measurements as pseudo-RGB input (Figure 1(b)). This convenient encoding conflates temporal gates with chromatic channels and removes their explicit slice identity before feature extraction. The network must therefore recover gate-specific geometry implicitly, while projected depth, projection validity, and cross-view consistency cannot directly govern the spatial contribution of each slice.

A second loss of physical evidence occurs after prediction. Sparse geometric support, projection noise, and weather-dependent gated responses can produce overlapping hypotheses with markedly different sensor support. As shown in Figure 1(c), hard NMS and standard Gaussian decay determine

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.

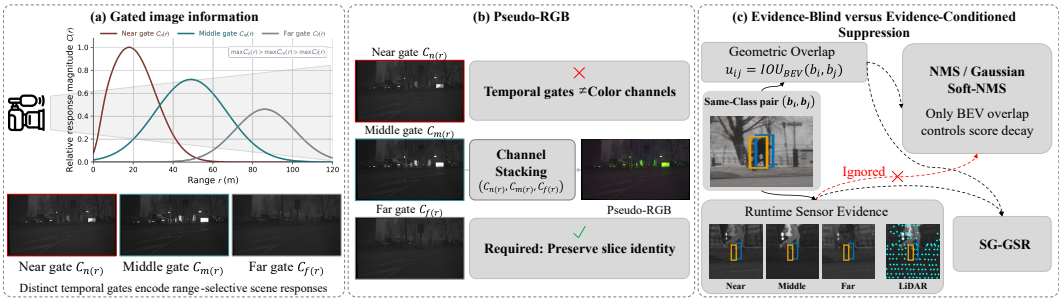

Figure 1: Motivation. (a) Schematic profiles illustrate distinct range-selective gated responses and decreasing return magnitude. (b) Pseudo-RGB stacking conflates temporal gates with color channels, obscuring slice identity. (c) Overlap-based suppression ignores gated and LiDAR support when resolving same-class hypotheses. GCSA preserves slice-specific evidence during representation learning, whereas SG-GSR conditions Gaussian score decay on sensor support at inference.

suppression from confidence and BEV overlap, leaving the available gated and LiDAR evidence outside the decision path. Candidates with comparable IoU therefore receive the same treatment even when they differ in gated agreement, LiDAR support, or depth consistency. Such sensor-agnostic decisions may suppress a valid hypothesis in a weakly observed region or retain a redundant prediction that lacks multimodal support.

Figure 1 exposes two points at which physically meaningful evidence is discarded: premature slice mixing and evidence-blind suppression. Our central premise is that range-selective gated evidence should instead be preserved during representation learning and remain available when resolving the final prediction set. Geometry-Conditioned Shared-Slice Aggregation (GCSA) therefore processes the temporal gates through independent passes of a shared encoder and estimates spatial slice contributions from projected LiDAR geometry. The resulting gated feature pyramid is consumed by the original bidirectional image-to-point (I2P) and point-to-image (P2I) interactions. After detection, Slice-Geometry-Conditioned Gaussian Set Refinement (SG-GSR) derives candidate reliability and pairwise compatibility from gated-slice agreement, local LiDAR support, and depth consistency. These quantities modulate same-class Gaussian score decay without altering box geometry, class labels, or detector parameters. GCSA thus preserves physical evidence in the representation, while SG-GSR retains it for prediction-set reasoning.

Our contributions are threefold:

- We introduce GCSA, which preserves the temporal structure of gated observations through independent passes of a shared visual encoder and uses projected geometry to infer spatially varying slice contributions.
- We propose SG-GSR, a deterministic set-refinement formulation that conditions same-class Gaussian score decay on gated agreement, LiDAR support, and depth consistency without additional detector training.
- Extensive three-class test-set evaluations across distance, weather, and illumination demonstrate complementary representation- and set-level gains. The complete framework achieves 69.72 Overall 3D AP40, exceeding the pseudo-RGB tri-modal baseline by 3.60 points and GCSA by 0.49 points while improving all three distance intervals.

2 Related Work

Multi-modal 3D detection. LiDAR–camera fusion has progressed from point-level augmentation Sindagi et al. (2019) to dense or sparse shared representations Liang et al. (2022); Xie et al. (2023) and object-centric interaction Bai et al. (2022); Yang et al. (2022). Recent studies further improve fusion through depth- and geometry-aware interaction Jiao et al. (2023); Li et al. (2024), object–scene reasoning Li et al. (2023); Yin et al. (2024), compact virtual instances Yu et al. (2025), and robustness to calibration or sensor degradation Song et al. (2024); Huang et al. (2024); Park et al. (2025); Le et al. (2024). Related radar–camera methods also exploit modality-specific geometry

Lin et al. (2024); Bai et al. (2026). These approaches regulate cross-sensor interaction but generally treat each imaging modality as a single observation stream. In contrast, our framework preserves the range-selective temporal structure within a gated camera and carries the resulting sensor evidence from feature aggregation to set-level refinement.

Gated-camera imaging in adverse weather. The Seeing Through Fog benchmark provides synchronized gated NIR and conventional automotive sensors under adverse weather Bijelic et al. (2020). Gated measurements have supported monocular 3D detection and depth estimation Julca-Aguilar et al. (2021); Walia et al. (2022), active stereo Walz et al. (2023); Brucker et al. (2024), and scene reconstruction Ramazzina et al. (2024). SAMFusion further demonstrates their value in multi-sensor 3D detection Palladin et al. (2024). Existing tri-modal detectors commonly expose the gated slices through a pseudo-RGB interface, obscuring their distinct integration windows. GCSA instead encodes the slices independently with shared parameters and uses projected geometry to aggregate their range-selective evidence.

Confidence estimation and set-level refinement. Set-based detectors predict a finite collection of object hypotheses Carion et al. (2020), while heatmap-guided 3D detectors initialize queries from dense object evidence Bai et al. (2022). Quality-aware methods improve the correspondence between classification confidence and localization accuracy through learned ranking or quality estimation Pu et al. (2023); Liu et al. (2024). At inference, NMS removes redundant boxes, whereas Gaussian score decay attenuates overlapping candidates without removing them Bodla et al. (2017). These strategies rely on learned detector features or pairwise geometry, without revisiting the measurements supporting each hypothesis. SG-GSR instead conditions continuous score decay on gated agreement, projected LiDAR support, and depth consistency. It requires no auxiliary quality network, preserves boxes and labels, and reduces to standard Gaussian decay when its conditioning terms are disabled.

3 METHOD

3.1 OVERVIEW

Let P denote a LiDAR point cloud, I an RGB image, and $\{G_s\}_{s\in\mathcal{S}}$ the gated NIR measurements, where $\mathcal{S} = \{n, m, f\}$ indexes the near, middle, and far temporal gates. The base detector extracts LiDAR and image features, exchanges them through bidirectional image-to-point (I2P) and point-to-image (P2I) interaction, and initializes K decoder queries from a dense class heatmap.

Figure 2 summarizes the resulting representation-to-set pipeline. Built upon the base detector, our framework introduces complementary operations at the representation and prediction-set levels. GCSA retains the gated measurements as slice-specific streams and aggregates them under projected geometric guidance before I2P/P2I interaction. After the unchanged detector generates candidates, SG-GSR conditions same-class score suppression on candidate-aligned sensor evidence. Thus, GCSA resolves cross-slice ambiguity during feature construction, whereas SG-GSR resolves residual competition among overlapping predictions without altering their geometry or labels.

3.2 GEOMETRY-CONDITIONED SHARED-SLICE AGGREGATION (GCSA)

As illustrated by the central stage of Figure 2, GCSA constructs a geometry-conditioned gated representation without discarding the range-selective identity of each temporal gate. Let $G_s \in \mathbb{R}^{H_0\times W_0}$ denote a single-channel gated measurement. To retain compatibility with an RGB-pretrained encoder, $\mathrm{Rep}_3(\cdot)$ replicates one slice across three channels. A shared backbone and feature pyramid E_g are then applied independently to each slice:

$$\mathbf{F}_s^\ell = E_g^\ell(\mathrm{Rep}_3(G_s)), \qquad s \in \mathcal{S}, \tag{1}$$

where $\ell \in \{1, \dots, L\}$ indexes an FPN level and $\mathbf{F}_s^\ell \in \mathbb{R}^{C_\ell\times H_\ell\times W_\ell}$. Parameter sharing controls the model size, while independent forward paths retain slice-specific features. Channel replication is used only at the encoder interface; the temporal measurements are not mixed as pseudo-color channels.

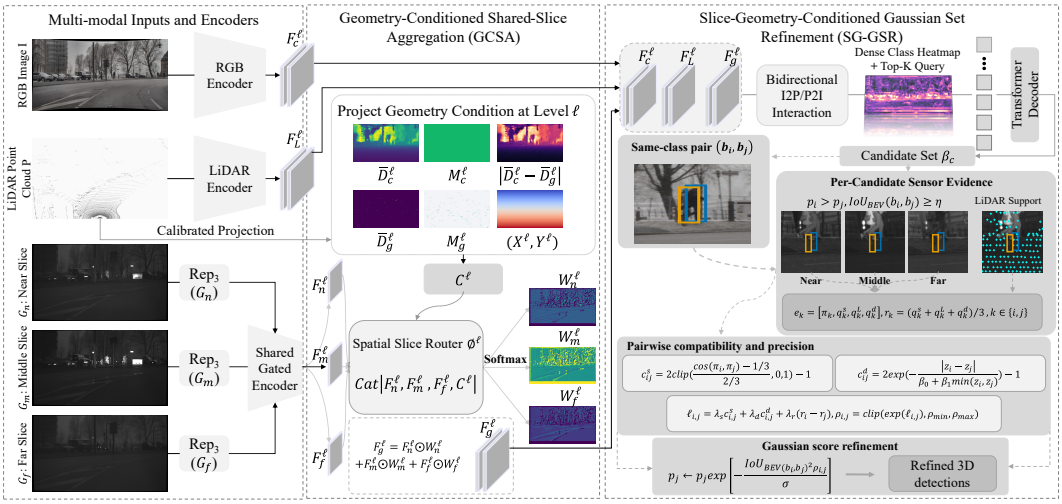

Figure 2: Framework overview. GCSA uses projected geometry to aggregate independently encoded gated slices before tri-modal fusion. SG-GSR then refines same-class candidate scores using slice, depth, and LiDAR evidence while preserving box geometry and class labels.

Slice appearance alone does not identify which temporal gate is geometrically reliable at a given location. At each level, we therefore form the condition $\mathcal{C}^\ell \in \mathbb{R}^{7\times H_\ell \times W_\ell}$:

$$\mathcal{C}^\ell = [\bar{\mathbf{D}}_c^\ell, \bar{\mathbf{D}}_g^\ell, \mathbf{M}_c^\ell, \mathbf{M}_g^\ell, |\bar{\mathbf{D}}_c^\ell - \bar{\mathbf{D}}_g^\ell|, \mathbf{X}^\ell, \mathbf{Y}^\ell], \tag{2}$$

where $\bar{\mathbf{D}}_c^\ell$ and $\bar{\mathbf{D}}_g^\ell$ are normalized LiDAR depth projected into the RGB and gated views; $\mathbf{M}_c^\ell$ and $\mathbf{M}_g^\ell$ are their binary validity masks; and $\mathbf{X}^\ell$ and $\mathbf{Y}^\ell$ are normalized coordinate maps. The absolute depth difference exposes cross-view geometric inconsistency, while the coordinate maps provide an explicit spatial prior.

A lightweight router ϕ^ℓ predicts one spatial logit per slice from the slice features and geometric condition:

$$\begin{aligned} \mathbf{A}^\ell &= \phi^\ell\big(\mathrm{Cat}[\mathbf{F}_n^\ell, \mathbf{F}_m^\ell, \mathbf{F}_f^\ell, \mathcal{C}^\ell]\big), \\ W_s^\ell(u,v) &= \frac{\exp A_s^\ell(u,v)}{\sum_{t\in\mathcal{S}} \exp A_t^\ell(u,v)}, \\ \mathbf{F}_g^\ell &= \sum_{s\in\mathcal{S}} W_s^\ell \odot \mathbf{F}_s^\ell. \end{aligned} \tag{3}$$

Here, $\mathbf{A}^\ell, \mathbf{W}^\ell \in \mathbb{R}^{3\times H_\ell \times W_\ell}$, Cat denotes channel-wise concatenation, and $\odot$ broadcasts a spatial weight across feature channels. The final layer of ϕ^ℓ is initialized to zero, yielding $W_s^\ell(u,v) = 1/3$ at initialization. Learning can thus depart smoothly from uniform aggregation without an initial slice preference. The resulting pyramid $\{\mathbf{F}_g^\ell\}_{\ell=1}^L$ directly replaces the pseudo-RGB gated pyramid in the original I2P/P2I pathway; all subsequent cross-modal operations are unchanged.

3.3 Slice-Geometry-Conditioned Gaussian Set Refinement (SG-GSR)

The rightmost stage of Figure 2 depicts SG-GSR, which makes set-level suppression conditional on the physical evidence supporting each prediction. For class c, let $\mathcal{B}_c = \{(\mathbf{b}_i, p_i)\}_{i=1}^{N_c}$ contain the predicted boxes $\mathbf{b}_i$ and confidence scores p_i. Projecting the center and image-plane extent of $\mathbf{b}_i$ into the gated view defines a local region Ω_i. For slice s, let $\mu_{i,s}$ be the mean response in Ω_i and μ_s^+ the frame-level mean over positive responses. We compute

$$\begin{aligned} a_{i,s} &= \log \frac{\mu_{i,s} + \epsilon}{\mu_s^+ + \epsilon}, \\ \boldsymbol{\pi}_i &= \mathrm{softmax}(\mathbf{a}_i/T), \qquad \mathbf{a}_i = [a_{i,n}, a_{i,m}, a_{i,f}], \\ q_i^{\mathrm{s}} &= 1 - \frac{\mathcal{H}(\boldsymbol{\pi}_i)}{\log 3}, \end{aligned} \tag{4}$$

where $\epsilon > 0$ stabilizes response normalization, T is a temperature, and $\mathcal{H}$ denotes entropy. The posterior $\boldsymbol{\pi}_i$ summarizes the relative gated-slice response, and $q_i^{\mathrm{s}} \in [0,1]$ measures its certainty. An invalid projection is assigned the uniform posterior and zero certainty.

We obtain complementary geometric evidence by projecting LiDAR points into the same region. Let n_i be the number of projected points in Ω_i, $\bar{z}_i$ their mean gated-camera depth, and z_i the predicted box-center depth. LiDAR support and depth agreement are

$$\begin{aligned} q_i^{\mathrm{L}} &= 1 - \exp(-n_i/\kappa), \\ q_i^{\mathrm{d}} &= \mathbb{I}[n_i > 0] \exp\left(-\frac{|\bar{z}_i - z_i|}{\tau_0 + \tau_1 z_i}\right), \\ r_i &= \frac{q_i^{\mathrm{s}} + q_i^{\mathrm{L}} + q_i^{\mathrm{d}}}{3}, \end{aligned} \tag{5}$$

where κ controls count saturation and $\tau_0 + \tau_1 z_i$ models range-dependent depth tolerance. The resulting reliability $r_i \in [0,1]$ combines gated certainty, geometric support, and depth consistency. All terms are computed from runtime predictions and calibrated sensor measurements.

Candidate reliability alone does not determine whether two boxes are duplicates. For a score-ordered pivot i and an overlapping candidate j, we therefore define slice and depth compatibility as

$$\begin{aligned} c_{ij}^{\mathrm{s}} &= 2\,\mathrm{clip}\left(\frac{\cos(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j) - 1/3}{2/3}, 0, 1\right) - 1, \\ c_{ij}^{\mathrm{d}} &= 2\exp\left(-\frac{|z_i - z_j|}{\beta_0 + \beta_1 \min(z_i, z_j)}\right) - 1. \end{aligned} \tag{6}$$

Both terms lie in $[-1,1]$: positive values indicate evidence compatible with a duplicate interpretation, whereas negative values favor retaining distinct hypotheses. We combine them with relative candidate reliability to obtain a bounded pairwise precision:

$$\begin{aligned} \ell_{ij} &= \lambda_{\mathrm{s}} c_{ij}^{\mathrm{s}} + \lambda_{\mathrm{d}} c_{ij}^{\mathrm{d}} + \lambda_{\mathrm{r}}(r_i - r_j), \\ \rho_{ij} &= \mathrm{clip}(\exp(\ell_{ij}), \rho_{\min}, \rho_{\max})\,. \end{aligned} \tag{7}$$

The asymmetric term $r_i - r_j$ strengthens suppression when a well-supported pivot competes with a less reliable candidate. For pairs satisfying $\mathrm{IoU}_{\mathrm{BEV}}(\mathbf{b}_i, \mathbf{b}_j) \geq \eta$, SG-GSR updates only the lower-ranked score:

$$p_j \leftarrow p_j \exp\left(-\frac{\mathrm{IoU}_{\mathrm{BEV}}(\mathbf{b}_i, \mathbf{b}_j)^2 \rho_{ij}}{\sigma}\right). \tag{8}$$

Equation (8) is applied independently within each predicted class; cross-class boxes never suppress one another. Setting $\lambda_{\mathrm{s}} = \lambda_{\mathrm{d}} = \lambda_{\mathrm{r}} = 0$ yields $\rho_{ij} = 1$ and recovers standard Gaussian score decay exactly. SG-GSR is therefore a bounded, sensor-conditioned generalization of Gaussian decay: it preserves every box and class label, requires no detector retraining, and adds only ROI and projection statistics at inference.

4 Experiments

We evaluate the proposed framework through comparisons with representative detectors, qualitative analysis under degraded visibility, and controlled ablations of the proposed components.

4.1 Experimental Setup

Experiments use the SeeingThroughFog dataset and the DENSE protocol adopted by SAMFusion Bijelic et al. (2020); Palladin et al. (2024). Because its frame-level assignment is not available in our setting, we construct a deterministic count- and condition-aligned, but not frame-identical, split: 10,046/1,000/1,941 training/validation/test samples and 1,046/895 daytime/nighttime test samples, leaving 10 of 12,997 samples unused. All methods and variants are retrained on this common split. Models are fitted on training, selected on validation, and evaluated once on test. Following the KITTI-style DENSE protocol, we report 3D and BEV AP_{40} across three classes and distance intervals; complete metric definitions are provided in the supplementary material.

Table 1: Day- and night-time detection performance on the SeeingThroughFog test split. Following SAMFusion, we report 3D and BEV AP40 over three distance intervals for Pedestrian, RidableVehicle, and PassengerCar.

Class	Method	Mod.	Day						Night					
			3D object detection			BEV detection			3D object detection			BEV detection		
			0–30m	30–50m	50–80m	0–30m	30–50m	50–80m	0–30m	30–50m	50–80m	0–30m	30–50m	50–80m
Pedestrian	SECOND Yan et al. (2018)	L	58.6548	49.0505	13.3108	59.2173	49.8362	13.6585	52.0743	38.7426	14.0917	52.1452	38.9415	14.0953
	MVXNet Sindagi et al. (2019)	CL	57.6775	43.0645	14.3680	57.7019	43.2793	14.3858	54.7019	35.6519	11.3952	54.9618	35.7306	11.3879
	M3D-RPN Brazil & Liu (2019)	C	13.1316	0.5586	0.0285	13.3132	0.5706	0.0286	11.2329	0.3727	0.0695	11.6681	0.3909	0.0699
	TransFusion Bai et al. (2022)	CL	60.3882	49.9678	20.4300	60.5866	50.3917	20.5259	51.7476	41.4083	19.2090	52.7242	41.6148	19.4264
	BEVFusion Liang et al. (2022)	CL	60.9121	48.6461	15.8978	61.0843	49.2358	16.1540	52.4318	41.9569	19.2479	52.6091	42.4648	19.7781
	DeepInteraction Yang et al. (2022)	CL	54.0299	53.4252	22.0758	54.3834	53.6796	22.3389	30.0572	32.7995	17.6820	30.0651	32.8060	17.6840
	SparseFusion Xie et al. (2023)	CL	61.9872	53.2837	20.9618	62.1349	53.6394	21.5903	54.1087	44.8377	20.9843	54.1743	45.1368	21.0462
	DAL Huang et al. (2024)	CL	28.0425	19.6563	3.6108	28.2253	19.9214	3.8646	19.1690	19.3445	2.3380	19.3672	19.7817	2.6546
	SAMFusion Palladin et al. (2024)	CGLR	71.3104	63.4373	28.2576	71.4837	63.5147	28.3201	61.3998	54.9874	25.5618	61.4727	55.1260	25.5882
	MoME Park et al. (2025)	CL	39.9579	32.7310	12.9712	40.2153	33.4600	13.2039	35.1356	27.2031	5.5567	35.6302	27.3006	5.6045
	CMF-IoU Ning et al. (2026)	CL	72.4314	55.3462	13.9078	72.4314	55.2985	13.9078	62.8509	45.0910	11.9702	62.8855	45.1369	12.2758
	BEVDilation Zhang et al. (2026)	CL	53.0604	43.7456	14.3934	58.3351	51.5646	18.9982	39.5521	28.0128	7.3779	45.3899	33.7874	9.9072
	Ours	CGL	**76.4666**	**66.5211**	**31.8763**	**77.3539**	**67.0815**	**32.5718**	**72.8731**	**60.4996**	**28.3885**	**73.5318**	**60.9528**	**28.7006**
RidableVehicle	SECOND Yan et al. (2018)	L	84.2550	51.8811	22.0556	84.2550	51.8811	22.0556	76.6243	42.4423	13.6213	76.6243	42.4423	13.6213
	MVXNet Sindagi et al. (2019)	CL	74.5179	43.7529	12.3636	74.5179	43.7529	12.3636	74.3115	47.6690	12.1193	74.3115	47.6690	12.1193
	M3D-RPN Brazil & Liu (2019)	C	51.7017	1.9284	0.0000	51.7017	1.9284	0.0000	43.4766	4.1622	0.0685	43.4766	4.1622	0.0685
	TransFusion Bai et al. (2022)	CL	88.8219	48.5838	18.8946	88.8219	48.5838	18.8946	83.3335	58.4850	20.8561	83.3335	58.4850	20.8561
	BEVFusion Liang et al. (2022)	CL	88.1592	47.9373	19.2832	88.1592	47.9373	19.2832	85.4567	59.4457	20.7194	85.4567	59.4457	20.7194
	DeepInteraction Yang et al. (2022)	CL	78.4620	50.3891	30.5221	80.3952	50.4943	30.6251	36.6942	26.0686	10.7892	36.6976	26.1101	10.7892
	SparseFusion Xie et al. (2023)	CL	89.5062	51.1663	23.1678	89.5062	51.1663	23.1678	80.7205	55.6890	18.6186	80.7205	55.6890	18.6186
	DAL Huang et al. (2024)	CL	27.5338	20.6507	8.1285	27.5660	20.6508	8.1285	31.9265	26.2406	5.1665	31.9265	26.3232	5.1665
	SAMFusion Palladin et al. (2024)	CGLR	85.7730	37.9519	**35.6514**	85.9943	38.9261	**35.6794**	85.4458	63.0045	23.2240	85.4458	63.0045	23.2240
	MoME Park et al. (2025)	CL	38.9741	24.7664	6.1662	38.9741	24.7666	6.1662	31.6599	28.0348	2.7578	31.7937	28.0844	2.7617
	CMF-IoU Ning et al. (2026)	CL	51.7842	36.0855	9.8333	51.7842	36.0855	9.8333	64.7656	44.2081	14.0909	64.7656	44.2081	14.0909
	BEVDilation Zhang et al. (2026)	CL	83.3700	35.4865	12.0349	89.2106	41.1910	14.0240	68.5486	38.1372	11.2460	68.8484	38.1885	12.0555
	Ours	CGL	**92.5961**	**64.6423**	30.5295	**92.5961**	**64.6423**	30.5295	**86.1469**	**68.5958**	**26.4634**	**86.1469**	**68.5958**	**26.4634**
PassengerCar	SECOND Yan et al. (2018)	L	92.2266	81.2454	60.8664	92.2299	81.4121	61.0092	92.5407	80.5039	49.6593	92.5561	80.5579	50.2254
	MVXNet Sindagi et al. (2019)	CL	94.3397	86.6675	74.6006	94.4021	86.8024	74.8713	93.6763	85.2176	62.6063	93.6802	84.8594	62.6787
	M3D-RPN Brazil & Liu (2019)	C	86.1342	45.4124	7.3166	86.8233	47.5979	7.7601	81.3868	36.1150	3.5856	82.8742	39.1198	4.3237
	TransFusion Bai et al. (2022)	CL	94.0027	84.1751	67.8424	93.8034	84.2462	67.9960	92.7823	84.0451	58.6932	92.8006	84.0224	59.2269
	BEVFusion Liang et al. (2022)	CL	94.2387	83.9360	66.5609	94.2846	84.1460	67.2388	92.9341	83.1511	57.2567	92.9739	83.2712	58.1548
	DeepInteraction Yang et al. (2022)	CL	92.5915	86.6016	70.9945	92.6877	86.7550	71.2915	85.9124	82.8958	62.4844	85.9596	82.9769	62.6127
	SparseFusion Xie et al. (2023)	CL	95.2465	86.8622	71.8087	95.2637	87.2898	72.0130	94.2502	87.1630	62.0528	94.2580	87.1745	62.3437
	DAL Huang et al. (2024)	CL	61.1201	50.3982	21.8578	61.5268	50.8243	22.4366	55.8713	51.6888	18.8557	56.0493	52.0532	19.5589
	SAMFusion Palladin et al. (2024)	CGLR	95.5542	**89.2058**	78.3229	95.6505	**90.0829**	78.7006	93.5649	**89.7709**	66.7454	93.6129	**89.8671**	66.9537
	MoME Park et al. (2025)	CL	51.1614	40.4594	24.4694	51.2485	40.5640	24.7429	48.1823	39.9007	15.8154	48.2359	39.9610	15.9675
	CMF-IoU Ning et al. (2026)	CL	**95.7688**	88.7520	66.5559	95.7720	89.8595	66.5658	**94.4645**	86.6591	53.8984	**94.4708**	86.6653	54.0075
	BEVDilation Zhang et al. (2026)	CL	88.4567	66.1025	49.3453	**95.9334**	83.1636	73.4729	86.7278	66.6262	42.6431	93.9359	81.5494	62.4984
	Ours	CGL	94.4674	89.1414	**79.7385**	94.5946	89.5832	**80.5947**	94.3468	89.3150	**69.5516**	94.3536	89.3458	**69.9970**

Our implementation follows DeepInteraction Yang et al. (2022) with ResNet-50–FPN visual encoders He et al. (2016); Cai & Vasconcelos (2021) and $K = 200$ heatmap proposals. Training proceeds progressively from LiDAR to LiDAR–RGB and then LiDAR–RGB–gated detection. The final stage uses AdamW for 30 epochs with an initial learning rate of 5×10^{-5}. SG-GSR is applied to the fixed GCSA detector without additional training and modifies only confidence scores. Complete preprocessing, initialization, freezing, optimization, and SG-GSR settings are provided in the supplementary material.

4.2 Comparison with State-of-the-Art Methods

Tables 1 and 2 compare our method with representative camera-, LiDAR-, and fusion-based detectors. The former reports 3D and BEV AP40 under daytime and nighttime illumination; the latter isolates snow and fog to assess robustness to atmospheric degradation. We follow the modality notation of SAMFusion: C, G, S, R, and L denote RGB camera, gated NIR camera, stereo RGB, radar, and LiDAR, respectively.

The gains are strongest and most consistent for vulnerable road users. Relative to SAMFusion, our method improves Pedestrian 3D AP_{40} in every day/night range bin, with the largest margin at night and short range. RidableVehicle also improves in five of six bins. PassengerCar is already near saturation at short range, and its gains emerge primarily at 50–80 m. This class- and range-dependent behavior rules out a uniform score shift: gated evidence is most beneficial where passive appearance or LiDAR support is unreliable.

The weather-specific Overall scores improve for all three classes under both snow and fog, despite isolated bin-level trade-offs. Relative to SAMFusion, the gains under snow/fog are 5.63/8.49 points for Pedestrian, 0.27/2.30 points for RidableVehicle, and 0.77/0.22 points for PassengerCar. Pedestrian AP also increases throughout the evaluated range in both conditions. The complete framework

Table 2: Adverse-weather 3D AP_{40} on SeeingThroughFog. Overall denotes the arithmetic mean across the three distance intervals.

Class	Method	Mod.	3D object detection							
			Snow				Fog			
			0–30m	30–50m	50–80m	**Overall**	0–30m	30–50m	50–80m	**Overall**
Pedestrian	SECOND Yan et al. (2018)	L	70.1382	57.0977	17.9248	48.3869	67.6844	49.7480	11.1749	42.8691
	MVXNet Sindagi et al. (2019)	CL	69.6751	53.3083	18.8544	47.2793	65.4217	38.6131	10.1090	38.0479
	M3D-RPN Brazil & Liu (2019)	C	14.6729	0.5419	0.0689	5.0946	14.4797	0.5298	0.1091	5.0395
	TransFusion Bai et al. (2022)	CL	71.8087	60.4673	25.0215	52.4325	76.6186	52.7070	16.3288	48.5515
	BEVFusion Liang et al. (2022)	CL	72.5760	59.2782	21.1419	50.9987	74.6344	56.0913	13.4501	48.0586
	DeepInteraction Yang et al. (2022)	CL	67.0316	61.9353	26.5607	51.8425	64.8656	53.2242	9.6853	42.5917
	SparseFusion Xie et al. (2023)	CL	73.5777	63.4111	25.8122	54.2670	78.2077	57.5577	15.1556	50.3070
	DAL Huang et al. (2024)	CL	32.4100	26.9158	4.6849	21.3369	28.5457	28.2381	7.4178	21.4005
	SAMFusion Palladin et al. (2024)	CGLR	76.5027	71.6951	36.8243	61.6740	79.4474	65.0768	20.9137	55.1460
	MoME Park et al. (2025)	CL	49.0319	36.7214	9.7638	31.8390	46.5525	29.9752	5.4479	27.3252
	CMF-IoU Ning et al. (2026)	CL	80.6932	61.9300	16.9176	53.1803	77.7606	44.2030	7.0138	42.9925
	BEVDilation Zhang et al. (2026)	CL	66.2931	49.0933	15.0730	43.4865	51.4788	38.1860	6.7647	32.1432
	Ours	CGL	**83.5149**	**76.7121**	**41.6805**	**67.3025**	**88.7625**	**71.4607**	**30.6755**	**63.6329**
RidableVehicle	SECOND Yan et al. (2018)	L	56.4256	26.0833	4.3398	28.9496	20.3752	11.8950	0.1087	10.7930
	MVXNet Sindagi et al. (2019)	CL	49.3965	28.2277	3.2871	26.9704	19.0347	7.4038	0.1213	8.8533
	M3D-RPN Brazil & Liu (2019)	C	44.7989	2.6197	0.0893	15.8360	6.1111	0.8333	0.0000	2.3148
	TransFusion Bai et al. (2022)	CL	61.3423	37.9220	7.4476	35.5706	22.6656	8.6814	0.2632	10.5367
	BEVFusion Liang et al. (2022)	CL	61.4811	36.2609	6.1970	34.6463	22.0550	8.3300	0.5000	10.2950
	DeepInteraction Yang et al. (2022)	CL	56.2987	38.0589	4.7145	33.0240	21.0119	8.3024	0.0781	9.7975
	SparseFusion Xie et al. (2023)	CL	61.7708	37.7835	8.1378	35.8974	22.4685	8.3352	0.3690	10.3909
	DAL Huang et al. (2024)	CL	17.5085	21.4256	2.8260	13.9200	12.1057	5.8443	0.0000	5.9833
	SAMFusion Palladin et al. (2024)	CGLR	**65.6831**	40.4339	10.0857	38.7342	21.4840	10.7241	0.0790	10.7624
	MoME Park et al. (2025)	CL	19.5382	15.5136	1.1441	12.0653	8.6509	8.4964	0.0450	5.7308
	CMF-IoU Ning et al. (2026)	CL	47.0227	26.9560	5.0000	26.3262	22.5000	9.5833	0.0000	10.6944
	BEVDilation Zhang et al. (2026)	CL	58.8922	28.1193	8.7467	31.9194	**22.8102**	6.9559	0.9286	10.2316
	Ours	CGL	62.6033	**42.2925**	**12.1229**	**39.0062**	22.5575	**14.6875**	**1.9318**	**13.0589**
PassengerCar	SECOND Yan et al. (2018)	L	95.7554	90.2757	59.2912	81.7741	95.0822	81.8359	53.6160	76.8447
	MVXNet Sindagi et al. (2019)	CL	96.7463	93.1196	74.8312	88.2324	96.1365	90.1106	69.1639	85.1370
	M3D-RPN Brazil & Liu (2019)	C	87.5362	47.0690	7.4082	47.3378	92.9470	58.0187	7.2208	52.7288
	TransFusion Bai et al. (2022)	CL	96.2879	92.0471	65.8208	84.7186	99.4236	83.8371	55.0481	79.4363
	BEVFusion Liang et al. (2022)	CL	96.2228	90.0526	62.5953	82.9569	**99.5818**	84.3331	53.6070	79.1740
	DeepInteraction Yang et al. (2022)	CL	95.2247	93.2086	69.3532	85.9288	95.6783	83.9643	53.4087	77.6838
	SparseFusion Xie et al. (2023)	CL	97.2262	93.9570	69.2902	86.8245	97.3175	85.5698	57.9960	80.2944
	DAL Huang et al. (2024)	CL	62.8289	52.2157	19.8455	44.9634	54.4033	54.2698	20.3933	43.0221
	SAMFusion Palladin et al. (2024)	CGLR	96.4710	**94.2524**	77.5742	89.4325	96.8544	**94.6194**	74.7776	88.7505
	MoME Park et al. (2025)	CL	52.4110	43.6559	18.9456	38.3375	56.5864	45.8803	18.6733	40.3800
	CMF-IoU Ning et al. (2026)	CL	**97.4560**	92.5852	61.2461	83.7624	98.1308	92.2431	60.3793	83.5844
	BEVDilation Zhang et al. (2026)	CL	86.4712	69.5004	45.0640	67.0119	91.6126	71.9505	41.7021	68.4217
	Ours	CGL	97.0032	94.0677	**79.5429**	**90.2046**	97.2411	93.2234	**76.4568**	**88.9738**

reaches 69.7193 Overall AP_{40} and improves upon GCSA without sacrificing its strong short-range predictions. Consistency across illumination and weather indicates that the gain follows sensor support rather than a condition-specific bias.

4.3 Qualitative Results

Figure 3 compares the complete model with three reference detectors under identical display settings. In the snow/day scene, our method retains annotated objects missed by the baselines. In fog/night, its predictions follow the annotated set with fewer extraneous boxes. In clear/night, it improves object coverage over TransFusion and SAMFusion while avoiding the numerous extra predictions of MVXNet. The paired BEV views allow the spatial placement of these detections to be inspected; further condition-specific and set-refinement examples are provided in the supplementary material.

4.4 Ablation Experiments

We first disentangle gains due to additional sensing modalities from those due to the proposed representation and refinement mechanisms. Table 3 traces the complete progression from LiDAR-only detection to RGB fusion, pseudo-RGB gated fusion, GCSA, and SG-GSR. This ordering exposes whether each component contributes complementary evidence or merely redistributes accuracy across distance ranges.

RGB raises the LiDAR baseline by 9.56 Overall AP_{40} through dense semantic context. Naively adding pseudo-RGB gated input then reduces Overall AP_{40} by 0.48 points: its 2.09-point far-range gain is offset by losses of 2.14 and 1.39 points at near and middle range. This reversal exposes the

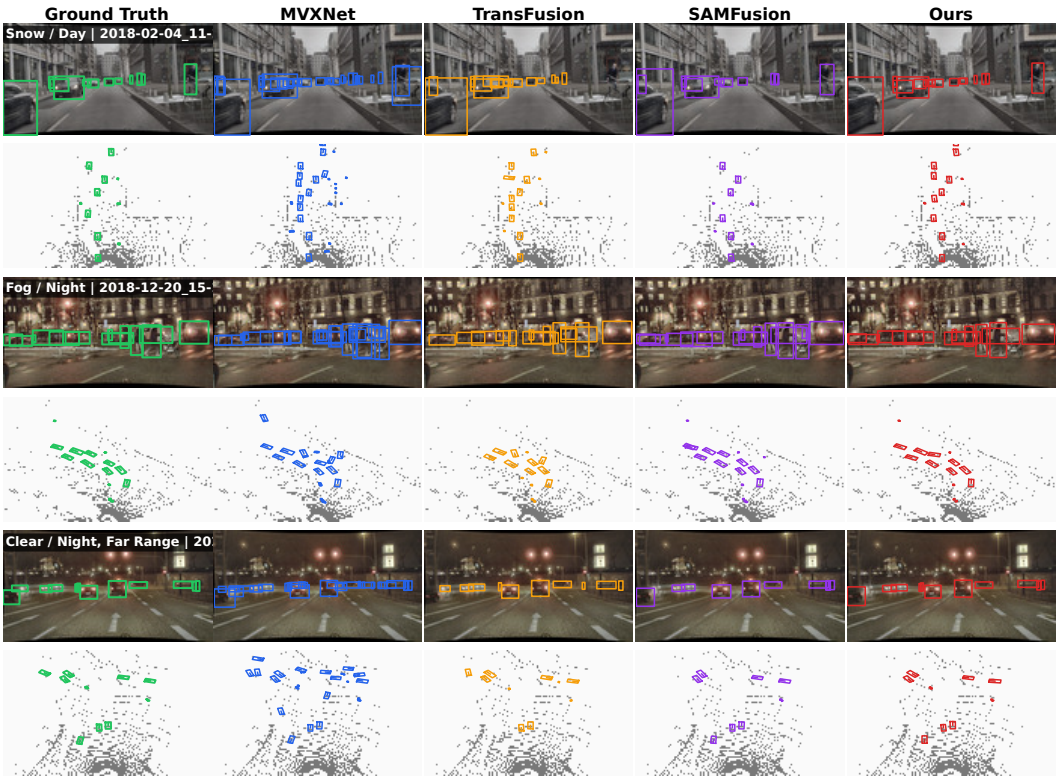

Figure 3: Qualitative comparison in snow/day, fog/night, and clear/night test scenes. Columns show ground truth (green), MVXNet (blue), TransFusion (orange), SAMFusion (purple), and ours (red). Paired RGB views (top) and LiDAR BEV views (bottom) reveal missed and extraneous detections under a common confidence threshold of 0.10.

Table 3: Progressive modality and component ablation. GCSA and SG-GSR denote geometry-conditioned shared-slice aggregation and slice-geometry-conditioned Gaussian set refinement, respectively.

Variant	Input modality			Proposed component		3D AP_{40}			
	L	C	G	GCSA	SG-GSR	0–30m	30–50m	50–80m	Overall
LiDAR baseline	✓					77.5759	58.5557	34.9889	57.0402
+ RGB	✓	✓				**85.9068**	70.3561	43.5330	66.5986
+ pseudo-RGB gated	✓	✓	✓			83.7713	68.9688	45.6187	66.1196
+ geometry-conditioned shared slices (GCSA)	✓	✓	✓	✓		85.5487	72.3462	49.7799	69.2249
+ all-class SG-GSR (full method)	✓	✓	✓	✓	✓	85.8725	**72.9520**	**50.3333**	**69.7193**

cost of collapsing range-selective slices before fusion. GCSA resolves the trade-off, improving the pseudo-RGB model by 3.11 points overall and by 1.78–4.16 points in every range. SG-GSR adds a further 0.49 points after decoding. The progression therefore separates two complementary effects: GCSA improves candidate formation, whereas SG-GSR improves confidence ordering within the final candidate set. Table 4 isolates slice preservation and appearance- and geometry-conditioned routing under identical initialization and training.

Preserving slice identity is beneficial even under uniform averaging, improving the pseudo-RGB baseline by 1.02 points overall and 2.66 points at 50–80 m. Appearance- and geometry-conditioned routing add 1.51 and 1.94 points, respectively, over this shared-slice baseline. Geometry-only routing is strongest at 50–80 m, whereas their combination gives the best Overall and near-to-middle-range results. Thus, geometry provides the dominant routing cue under sparse support, while appearance resolves ambiguities when observations are denser. Table 5 ablates the SG-GSR evidence terms on fixed GCSA predictions and Gaussian kernels; RidableVehicle retains its fixed recipe.

Table 4: Component ablation of GCSA. Slice denotes slice-preserving shared encoding; App. and Geo. denote appearance- and geometry-conditioned routing, respectively.

Variant	Slice	App.	Geo.	0–30m	30–50m	50–80m	Overall
Pseudo-RGB				83.7713	68.9688	45.6187	66.1196
Shared slices + uniform averaging	✓			83.7372	69.4111	48.2785	67.1423
Appearance-only routing	✓	✓		85.1165	71.0825	49.7646	68.6545
Geometry-only routing	✓		✓	84.5550	71.6452	51.0599	69.0867
Full GCSA	✓	✓	✓	85.5487	72.3462	49.7799	69.2249

Table 5: Evidence-term ablation of SG-GSR. Slice, depth, and reliability denote the precision terms weighted by λ_s, λ_d, and λ_r in Eq. (7).

Refinement	Slice	Depth	Reliability	0–30m	30–50m	50–80m	Overall
Original NMS				85.5487	72.3462	49.7799	69.2249
Gaussian decay				85.8415	72.8822	50.2527	69.6588
+ slice consistency	✓			85.8621	72.9315	50.3096	69.7011
+ depth consistency		✓		85.8573	72.9170	50.2912	69.6885
+ reliability difference			✓	85.8420	72.8837	50.2535	69.6597
+ slice and depth	✓	✓		85.8722	72.9515	50.3327	69.7188
Full SG-GSR	✓	✓	✓	**85.8725**	**72.9520**	**50.3333**	**69.7193**

Unconditioned Gaussian decay contributes 0.43 Overall AP_{40} over the original suppression rule. Conditioning the decay on sensor evidence adds another 0.06 points, with slice agreement and depth consistency accounting for nearly all of the conditional gain. The reliability difference provides a smaller, conservative correction, while the complete composition is best in every distance interval. Because SG-GSR changes only scores, these gains isolate evidence-aware ranking from box regression or class relabeling. Figure 4 complements these component analyses by characterizing the accuracy–efficiency trade-off of slice-preserving gated encoding. Distance-conditioned routing profiles and classwise analyses are reported in the supplementary material.

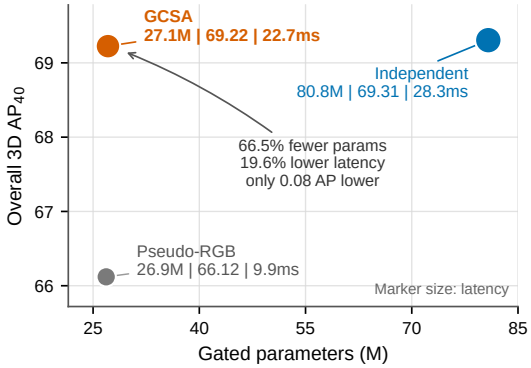

Figure 4: Accuracy–efficiency trade-off among gated encoders. The axes report gated-branch parameters and test Overall 3D AP_{40}; marker area denotes gated-path latency. Measurements use FP32 inference with batch size 1 and 448×800 inputs on an RTX 3090.

Pseudo-RGB packing is economical but discards slice identity, whereas independent encoders preserve it through substantial model replication. GCSA is only 0.08 AP below independent encoders while using 66.5% fewer gated-branch parameters and 19.6% lower latency. This operating point indicates that shared encoding serves as a structural prior for range-selective observations rather than merely a compression device.

5 Conclusion

This work addresses two sources of information loss in adverse-weather LiDAR–RGB–gated 3D detection: the loss of gated slice structure during fusion and overlap-only suppression after decoding. GCSA preserves range-selective gated observations through shared encoding and geometry-conditioned aggregation. SG-GSR uses slice agreement, depth consistency, and LiDAR support to refine confidence scores without altering box geometry or class labels. Experiments on SeeingThroughFog show consistent improvements across distance, illumination, and adverse weather. Future work will incorporate temporal evidence and explicit projection uncertainty.

AI USE STATEMENT

Generative AI tools were used to aid or polish writing and for retrieval and discovery, including language refinement, literature organization, and identification of relevant prior work. All AI-assisted text and suggested references were independently reviewed and verified by the authors. The authors retained full control over the research design, experimental execution, result interpretation, and manuscript content, and take full responsibility for the final content of this work.

REFERENCES

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

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*, 2022.

Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11682–11692, 2020.

Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. Soft-NMS: Improving object detection with one line of code. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5561–5569, 2017.

Garrick Brazil and Xiaoming Liu. M3D-RPN: Monocular 3d region proposal network for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9287–9296, 2019.

Samuel Brucker, Stefanie Walz, Mario Bijelic, and Felix Heide. Cross-spectral gated-RGB stereo depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21654–21665, 2024.

Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1483–1498, 2021.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pp. 213–229, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Junjie Huang, Yun Ye, Zhujin Liang, Yi Shan, and Dalong Du. Detecting as labeling: Rethinking LiDAR–camera fusion in 3d object detection. In *European Conference on Computer Vision*, pp. 439–455, 2024.

Yang Jiao, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. MSMDFusion: Fusing LiDAR and camera at multiple scales with multi-depth seeds for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21643–21652, 2023.

Frank Julca-Aguilar, Jason Taylor, Mario Bijelic, Fahim Mannan, Ethan Tseng, and Felix Heide. Gated3D: Monocular 3d object detection from temporal illumination cues. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Duy-Tho Le, Hengcan Shi, Jianfei Cai, and Hamid Rezatofighi. Diffusion model for robust multi-sensor fusion in 3d object detection and BEV segmentation. In *European Conference on Computer Vision*, pp. 232–249, 2024.

Xiaotian Li, Baojie Fan, Jiandong Tian, and Huijie Fan. GAFusion: Adaptive fusing LiDAR and camera with multiple guidance for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21209–21218, 2024.

Xin Li, Tao Ma, Yuenan Hou, Botian Shi, Yuchen Yang, Youquan Liu, Xingjiao Wu, Qin Chen, Yikang Li, Yu Qiao, and Liang He. LoGoNet: Towards accurate 3d object detection with local-to-global cross-modal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17524–17534, 2023.

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*, 2022.

Zhiwei Lin, Zhe Liu, Zhongyu Xia, Xinhao Wang, Yongtao Wang, Shengxiang Qi, Yang Dong, Nan Dong, Le Zhang, and Ce Zhu. RCBEVDet: Radar-camera fusion in bird's eye view for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14928–14937, 2024.

Zhe Liu, Jinghua Hou, Xiaoqing Ye, Tong Wang, Jingdong Wang, and Xiang Bai. SEED: A simple and effective 3d DETR in point clouds. In *European Conference on Computer Vision*, pp. 110–126, 2024.

Zhiwei Ning, Zhaojiang Liu, Xuanang Gao, Yifan Zuo, Jie Yang, Yuming Fang, and Wei Liu. CMF-IoU: Multi-stage cross-modal fusion 3d object detection with iou joint prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(2):2177–2190, 2026.

Edoardo Palladin, Roland Dietze, Praveen Narayanan, Mario Bijelic, and Felix Heide. SAMFusion: Sensor-adaptive multimodal fusion for 3d object detection in adverse weather. In *European Conference on Computer Vision*, 2024.

Konyul Park, Yecheol Kim, Daehun Kim, and Jun Won Choi. Resilient sensor fusion under adverse sensor failures via multi-modal expert fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6720–6729, 2025.

Yifan Pu, Weicong Liang, Yiduo Hao, Yuhui Yuan, Yukang Yang, Chao Zhang, Han Hu, and Gao Huang. Rank-DETR for high quality object detection. In *Advances in Neural Information Processing Systems*, pp. 16100–16113, 2023.

Andrea Ramazzina, Stefanie Walz, Pragyan Dahal, Mario Bijelic, and Felix Heide. Gated fields: Learning scene reconstruction from gated videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10530–10541, 2024.

Vishwanath A. Sindagi, Yin Zhou, and Oncel Tuzel. MVX-Net: Multimodal voxelnet for 3d object detection. In *IEEE International Conference on Robotics and Automation*, pp. 7276–7282, 2019.

Ziying Song, Lei Yang, Shaoqing Xu, Lin Liu, Dongyang Xu, Caiyan Jia, Feiyang Jia, and Li Wang. GraphBEV: Towards robust BEV feature alignment for multi-modal 3d object detection. In *European Conference on Computer Vision*, pp. 347–366, 2024.

Amanpreet Walia, Stefanie Walz, Mario Bijelic, Fahim Mannan, Frank Julca-Aguilar, Michael Langer, Werner Ritter, and Felix Heide. Gated2Gated: Self-supervised depth estimation from gated images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Stefanie Walz, Mario Bijelic, Andrea Ramazzina, Amanpreet Walia, Fahim Mannan, and Felix Heide. Gated stereo: Joint depth estimation from gated and wide-baseline active stereo cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13252–13262, 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.

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*, 2022.

Junbo Yin, Jianbing Shen, Runnan Chen, Wei Li, Ruigang Yang, Pascal Frossard, and Wenguan Wang. IS-Fusion: Instance-scene collaborative fusion for multimodal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14905–14915, 2024.

Zhuochen Yu, Bijie Qiu, and Andy W. H. Khong. ViKIENet: Towards efficient 3d object detection with virtual key instance enhanced network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11844–11853, 2025.

Guowen Zhang, Chenhang He, Liyi Chen, and Lei Zhang. BEVDilation: Lidar-centric multi-modal fusion for 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

SUPPLEMENTARY MATERIAL: GEOMETRY-CONDITIONED SHARED-SLICE FUSION AND GAUSSIAN SET REFINEMENT FOR ADVERSE-WEATHER 3D DETECTION

Anonymous authors
Paper under double-blind review

This supplementary material provides the experimental and visual evidence that supports the main paper. We first specify the complete training and evaluation protocol, then analyze the quantitative behavior of the two proposed components. We subsequently expose runtime GCSA tensors, proposal responses, and candidate-level SG-GSR evidence, followed by enlarged qualitative examples across illumination and weather conditions. We conclude with classwise refinement results and a discussion of applicability and limitations.

A COMPLETE EXPERIMENTAL PROTOCOL

A.1 DATASET AND EVALUATION

Experiments are conducted on the SeeingThroughFog dataset used by SAMFusion Bijelic et al. (2020); Palladin et al. (2024). It contains 12,997 synchronized samples acquired under diverse illumination and adverse weather, including daytime, nighttime, fog, snow, and rain. SAMFusion reports the training, validation, and test sample counts and the day/night composition of its test set, but the original frame-level assignment is not available in our setting. We therefore construct a deterministic reconstructed split that matches these reported statistics: 10,046 samples for training, 1,000 for validation, and 1,941 for testing, with 1,046 daytime and 895 nighttime test samples. Because these counts sum to 12,987, the remaining 10 samples are left unused. The resulting split is count- and condition-aligned with the SAMFusion protocol, but should not be interpreted as an exact frame-level copy of its original split. All compared methods and proposed variants are retrained and evaluated using this common reconstructed split; published baseline scores are not directly reused. Each sample provides LiDAR, a left-view RGB image, and near-, middle-, and far-range gated NIR measurements. After the partition is fixed, training uses the training split, all model and hyperparameter selection uses validation, and the test partition is reserved for final evaluation.

Following the DENSE protocol adopted by SAMFusion, detection accuracy is measured with KITTI-style average precision over 40 recall positions (AP_{40}). We evaluate Pedestrian, RidableVehicle, and PassengerCar. The primary metric is 3D AP_{40}; BEV AP_{40} is additionally reported in the main comparison. Results are stratified into 0–30 m, 30–50 m, and 50–80 m intervals. Overall AP_{40} denotes the arithmetic mean of the corresponding distance-stratified 3D results.

A.2 ARCHITECTURE AND INITIALIZATION

The framework is implemented in PyTorch and MMDetection3D, with DeepInteraction Yang et al. (2022) providing the underlying detector. We retain its bidirectional image-to-point (I2P) and point-to-image (P2I) interactions, heatmap-guided query initialization, transformer decoder, and detection head. Following SAMFusion Palladin et al. (2024), the LiDAR stream consists of a sparse voxel encoder and a BEV backbone, whereas the RGB stream employs a ResNet-50 He et al. (2016) with an FPN. The visual branch is initialized from pretrained Cascade Mask R-CNN weights Cai & Vasconcelos (2021). The gated stream uses the same backbone family but processes the three temporal slices independently with shared parameters. GCSA is instantiated at each of the five FPN levels.

The point-cloud extent is $[0, 100] \times [-40, 40] \times [-4, 4]$ m, and $K = 200$ heatmap proposals initialize the decoder queries. Voxelization, image preprocessing, decoder architecture, and detection post-processing otherwise follow the base detector.

A.3 Progressive Optimization

We proceed from LiDAR-only to LiDAR–RGB and finally to LiDAR–RGB–gated detection. The LiDAR–RGB model is initialized from the selected LiDAR-only checkpoint, and the tri-modal model is initialized from the selected LiDAR–RGB checkpoint. The gated ResNet-50 and FPN are initialized with the corresponding weights from the same pretrained Cascade Mask R-CNN model. During tri-modal optimization, the LiDAR and RGB feature extractors are frozen; the gated encoder, GCSA routers, cross-modal interaction layers, decoder, and detection heads remain trainable.

Optimization uses AdamW for 30 epochs with an initial learning rate of 5×10^{-5}, weight decay 0.01, first-epoch warmup, cosine annealing, and a total batch size of 6. We evaluate every epoch on the validation split and retain the checkpoint with the highest validation Overall AP_{40}.

A.4 SG-GSR Configuration

SG-GSR is applied to the selected GCSA checkpoint without further detector optimization. Candidate evidence is computed with $T = 0.35$, $\kappa = 3$, $\tau_0 = \beta_0 = 1$, and $\tau_1 = \beta_1 = 0.05$. We set $(\lambda_s, \lambda_d, \lambda_r) = (0.35, 0.25, 0.25)$ and constrain the adaptive precision to $[0.5, 2.0]$. The Gaussian kernel uses $\sigma = 0.5$ and $\eta = 0.05$; a more conservative kernel is used for RidableVehicle in the classwise study below. At inference, SG-GSR accesses only detector predictions, gated measurements, LiDAR points, and sensor calibration. It does not use ground-truth annotations and modifies only confidence scores.

B Extended Quantitative Discussion

B.1 Illumination and Weather

The day/night results in Table 1 of the main paper are most pronounced for vulnerable road users. Relative to SAMFusion, our method improves Pedestrian 3D AP_{40} by $+5.16$, $+3.08$, and $+3.62$ points from near to far range during the day, and by $+11.47$, $+5.51$, and $+2.83$ points at night. RidableVehicle improves in five of the six illumination–range settings, with the largest gain at 30–50 m during the day. PassengerCar accuracy is already near saturation at short range, but its 50–80 m AP improves by 1.42 during the day and 2.81 at night.

Weather-specific results exhibit the same behavior. Under snow, Pedestrian AP improves over SAMFusion by 7.01, 5.02, and 4.86 points across increasing ranges; under fog, the corresponding gains are 9.32, 6.38, and 9.76 points. Improvements for the other categories are more conservative and remain concentrated at middle and long range. These results indicate that range-selective gated representation principally benefits geometrically sparse or visually ambiguous objects.

The 3D and BEV results follow the same ordering across illumination conditions. This agreement is important because a gain confined to BEV could arise from improved horizontal localization without accurate 3D extent. Instead, the consistent behavior of both metrics indicates that the proposed components improve candidate support without introducing a corresponding vertical or size-estimation penalty. The smaller PassengerCar gains also define a useful boundary: when dense appearance and LiDAR geometry already provide reliable evidence, gated conditioning has less room to alter the ranking.

B.2 Progressive Modality Behavior

The progressive results in Table 3 of the main paper show that adding a sensor does not guarantee monotonic improvement. RGB supplies dense semantics and substantially strengthens the LiDAR detector, whereas pseudo-RGB gated packing improves the far interval but weakens near- and middle-range performance. This contrast attributes the degradation to representation rather than

to an absence of useful gated signal: the additional modality helps where LiDAR is sparse, yet its premature channel mixing interferes with already reliable features elsewhere.

GCSA restores monotonic gains over the pseudo-RGB model in all three intervals without replicating the gated encoder. SG-GSR then improves the same checkpoint without changing boxes or labels. The two stages therefore address different failure modes under a controlled progression: GCSA changes which candidates are formed from multimodal evidence, while SG-GSR changes how overlapping candidates are ordered after decoding.

B.3 GCSA Component Analysis

Preserving temporal slice identity is beneficial even without learned routing. Replacing pseudo-RGB packing with shared slice encoding and uniform averaging improves Overall AP from 66.1196 to 67.1423. The gain is concentrated at 50–80 m ($+2.6598$ AP), whereas the 0–30 m result remains essentially unchanged (-0.0341 AP). This result indicates that early channel packing primarily obscures range-selective information under sparse long-range observations.

Relative to uniform averaging, appearance-only routing gains 1.5122 Overall AP and improves all three intervals. Geometry-only routing yields a larger 1.9444-point gain, including improvements of 2.2341 and 2.7814 AP at 30–50 m and 50–80 m. Combining appearance and geometry produces the best Overall AP (69.2249) and the strongest near- and middle-range results. Geometry-only routing retains the highest far-range AP (51.0599), suggesting that appearance cues can partially dilute geometric preference under the sparsest observations. Their joint use nevertheless provides the best aggregate trade-off.

This non-monotonic far-range result clarifies the role of the two conditioning sources. Geometry provides the more stable prior when projected support is sparse, while appearance better resolves local structures in denser near- and middle-range regions. Full GCSA does not force either cue to dominate globally; it favors the combination that maximizes aggregate performance across the operating range. The component study therefore supports conditional routing rather than a fixed slice preference or a single-cue router.

B.4 SG-GSR Evidence Analysis

Applying unconditioned Gaussian decay to the fixed class policy improves the original prediction set by 0.4339 Overall AP. Slice and depth consistency provide additional gains of 0.0423 and 0.0297 AP, respectively, demonstrating that both gated response agreement and geometric proximity refine the ordering of overlapping hypotheses. Their joint use reaches 69.7188 AP, a 0.0600-point improvement over isotropic Gaussian decay. The reliability-difference term is deliberately conservative: it yields only 0.0009 AP in isolation and adds 0.0005 AP to the slice–depth combination. The complete formulation achieves the best result in every distance interval.

The modest magnitude of the refinement gain is consistent with its restricted scope. SG-GSR acts only on overlapping same-class candidates already produced by GCSA; it cannot recover a missed object, correct box geometry, or change a class label. Its ablation therefore isolates ranking quality from detector retraining. Slice and depth consistency account for most of the improvement, whereas reliability difference adjusts only asymmetric pairs with unequal support. This behavior is preferable to an aggressive global rescaling that could perturb isolated high-confidence detections.

B.5 Statistical Stability and Mechanism Audit

We further compare SG-GSR with a matched Gaussian-decay baseline on the same frozen detector predictions. The matched baseline preserves the candidate set, classwise policy, σ, and η, but disables sensor conditioning by setting $\lambda_{\mathrm{s}} = \lambda_{\mathrm{d}} = \lambda_{\mathrm{r}} = 0$. This isolates the contribution of candidate-aligned evidence from the effect of Gaussian score decay itself. The following analyses are post-hoc diagnostics: ground-truth annotations are used only to recompute evaluation statistics and define diagnostic matches, never by SG-GSR at inference.

Sequence-cluster bootstrap. Adjacent frames from the same acquisition sequence are correlated, so we group the 1,941 test frames into 484 sequence clusters using their sequence identifiers and

resample clusters rather than individual frames. For each of 2,000 paired bootstrap replicates, we recompute the AP_{40} difference between SG-GSR and matched Gaussian decay. The bootstrap evaluator reproduces the official point estimates to numerical precision. Table 1 reports the bootstrap mean difference, percentile interval, and empirical fraction of replicates with positive improvement; the latter is a descriptive bootstrap quantity rather than a hypothesis-test p-value.

Table 1: Paired sequence-cluster bootstrap for SG-GSR relative to matched Gaussian decay. Differences are in 3D AP_{40} points. The 0–30, 30–50, and 50–80 m rows average the classwise differences within the corresponding distance interval.

Metric	Mean Δ	95% interval	Positive replicates
Overall	+0.0582	[0.0359, 0.0796]	1.000
0–30 m	+0.0145	[−0.0312, 0.0519]	0.756
30–50 m	+0.0730	[0.0405, 0.1029]	1.000
50–80 m	+0.0871	[0.0585, 0.1155]	1.000
Pedestrian mean	+0.1669	[0.0989, 0.2306]	1.000
PassengerCar mean	+0.0077	[0.0022, 0.0123]	0.994

The Overall interval lies entirely above zero. The range-stratified results localize this stability to the middle and far intervals; the near-range interval includes zero and should not be interpreted as a consistent gain. Pedestrian provides the dominant classwise contribution, while the much smaller PassengerCar effect remains positive in 99.4% of bootstrap replicates.

Evidence-alignment counterfactual. To test whether the gain depends on associating the correct sensor evidence with each candidate, we randomly permute all SG-GSR evidence fields among candidates of the same predicted class within each frame. This preserves the predictions, the number of candidates, and the marginal evidence distribution, while breaking candidate–evidence alignment. Table 2 summarizes 20 permutations.

Table 2: Within-frame, within-class evidence-shuffling counterfactual. The shuffled result is the mean $\pm$ standard deviation over 20 permutations.

Evidence assignment	Overall 3D AP_{40}
Matched Gaussian decay	69.6588
Shuffled SG-GSR evidence	69.6677 ± 0.0026
Correctly aligned SG-GSR evidence	**69.7193**

Correct alignment exceeds the shuffled mean by 0.0516 points, whereas shuffling retains only a 0.0089-point gain over matched Gaussian decay. None of the 20 shuffled runs reaches the aligned result. Thus, most of the conditional gain cannot be explained by an arbitrary score perturbation with the same evidence distribution; it depends on candidate-specific sensor alignment.

Duplicate diagnostics at matched recall. Finally, we examine whether the score changes reduce redundant hypotheses at a fixed operating recall. For this diagnostic, a Pedestrian prediction is assigned to its best valid ground-truth instance when their 3D IoU exceeds 0.1; multiple predictions assigned to the same instance are counted as duplicates. Valid instances are unoccluded, untruncated, and within 80 m. As shown in Table 3, SG-GSR reduces duplicate predictions by 25.0% at 70% recall and 40.7% at 90% recall relative to matched Gaussian decay.

Table 3: Pedestrian duplicate predictions at matched recall. Values report duplicates per 100 covered ground-truth instances.

Recall	Gaussian decay	SG-GSR	Relative reduction
0.70	0.431	0.323	25.0%
0.90	22.264	13.208	40.7%

This diagnostic is threshold dependent and is not a replacement for AP. At the common score threshold of 0.10, SG-GSR reduces Pedestrian duplicates per 100 covered instances from 4.525 to 2.621,

while ground-truth coverage changes slightly from 85.06% to 84.94%. Together with the matched-recall and bootstrap results, this indicates that SG-GSR primarily improves confidence ordering by removing redundant high-recall hypotheses, with a small operating-point trade-off at a fixed threshold.

C Runtime Visualization of the Proposed Components

The following visual audit uses test sample `2018-02-04_11-26-31_00200`, a snowy daytime scene containing objects across all three evaluation ranges. All feature, weight, proposal, and query maps are captured from one evaluation forward of the selected GCSA checkpoint. Ground truth is used only for the explicitly labeled reference visualization; it is not involved in producing the runtime tensors.

C.1 Slice-Resolved GCSA Behavior

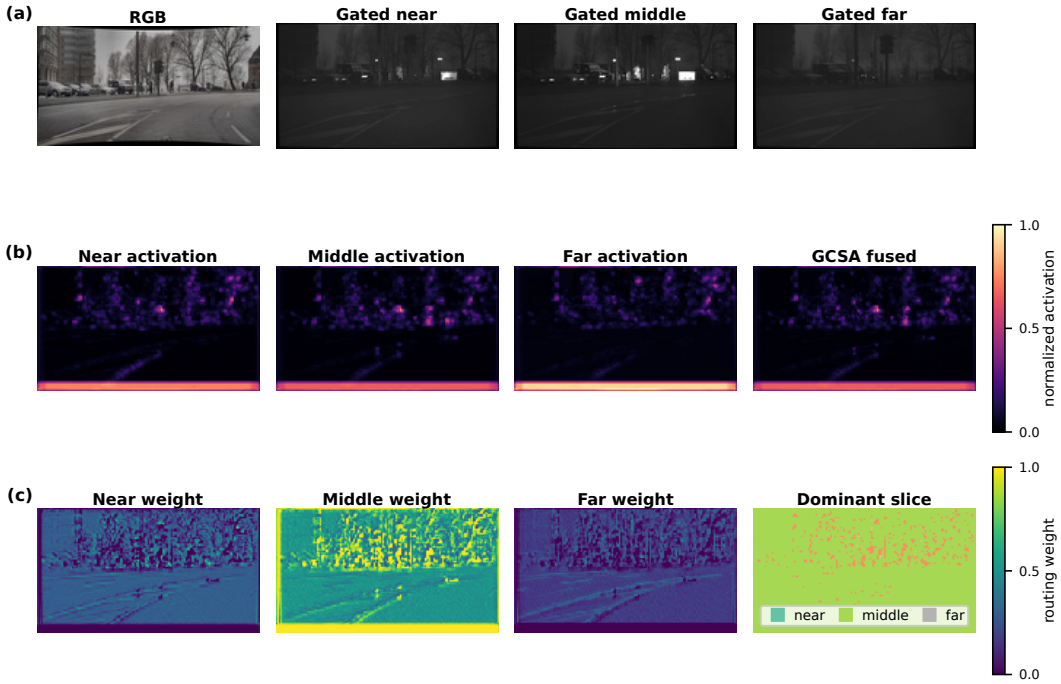

Figure 1: Runtime visualization of GCSA. (a) The synchronized RGB image and raw near-, middle-, and far-range gated slices. (b) Mean absolute FPN-level-0 activations for the three shared-encoder outputs and their GCSA aggregation. (c) Normalized spatial routing weights and the dominant-slice assignment. Each activation map is independently normalized to $[0, 1]$ for spatial comparison, whereas the routing-weight maps share their native interval $[0, 1]$ and are directly comparable.

Figure 1 exposes the representation retained by the slice-preserving design. The three gated measurements exhibit visibly different responses despite observing the same scene, and these differences persist in the shared FPN features. The router does not collapse them into a globally fixed mixture: although the middle slice receives broad support in this frame, the near and far weights vary around scene structures and object-bearing regions. The fused feature consequently preserves strong common responses while selectively incorporating complementary slice evidence. This runtime behavior provides a mechanism-level counterpart to the component ablation in the main paper.

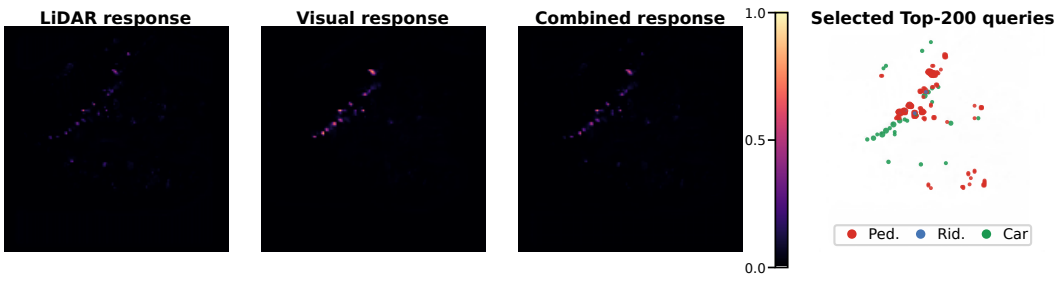

Figure 2: Proposal formation for the same test sample. From left to right, we show the maximum class response of the LiDAR proposal heatmap, the visual-fused response, their combined response, and the selected Top-K queries. Heatmaps are pre-Top-K sigmoid responses and should not be interpreted as semantic segmentation maps.

C.2 Multimodal Proposal Formation

The visual-fused response in Figure 2 is localized around object-bearing regions that are weak or spatially sparse in the LiDAR heatmap. Their combination retains the geometric anchor supplied by LiDAR while increasing proposal support at visually corroborated locations. The selected queries remain concentrated on high-response structures rather than spreading uniformly over the BEV plane. This trace clarifies where the routed gated representation affects detection: it contributes before query selection, not only through a late confidence adjustment.

C.3 Candidate-Level SG-GSR Evidence

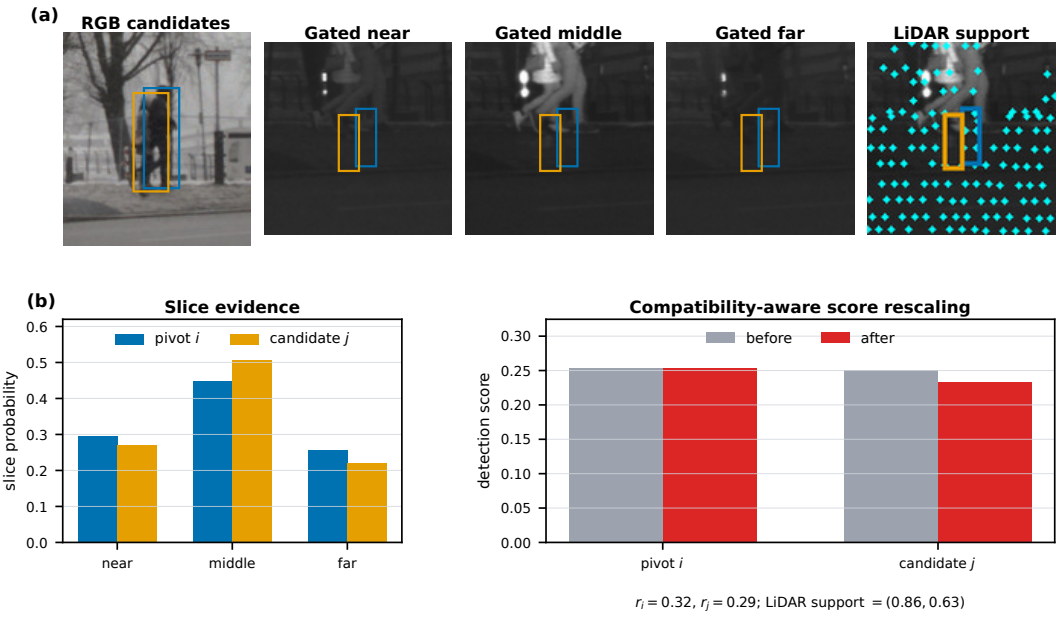

Figure 3: Candidate-level SG-GSR evidence. (a) A competing Pedestrian pair in RGB, the three gated slices, and projected LiDAR support. Blue denotes the higher-ranked pivot and orange the competing candidate. (b) Detector scores before and after SG-GSR for the predictions in the same frame. SG-GSR changes confidence ordering only; class labels and box geometry are preserved.

The pair in Figure 3 occupies nearly the same image region, so overlap alone is insufficient to determine which candidate should be suppressed more strongly. The gated slices provide distinct local

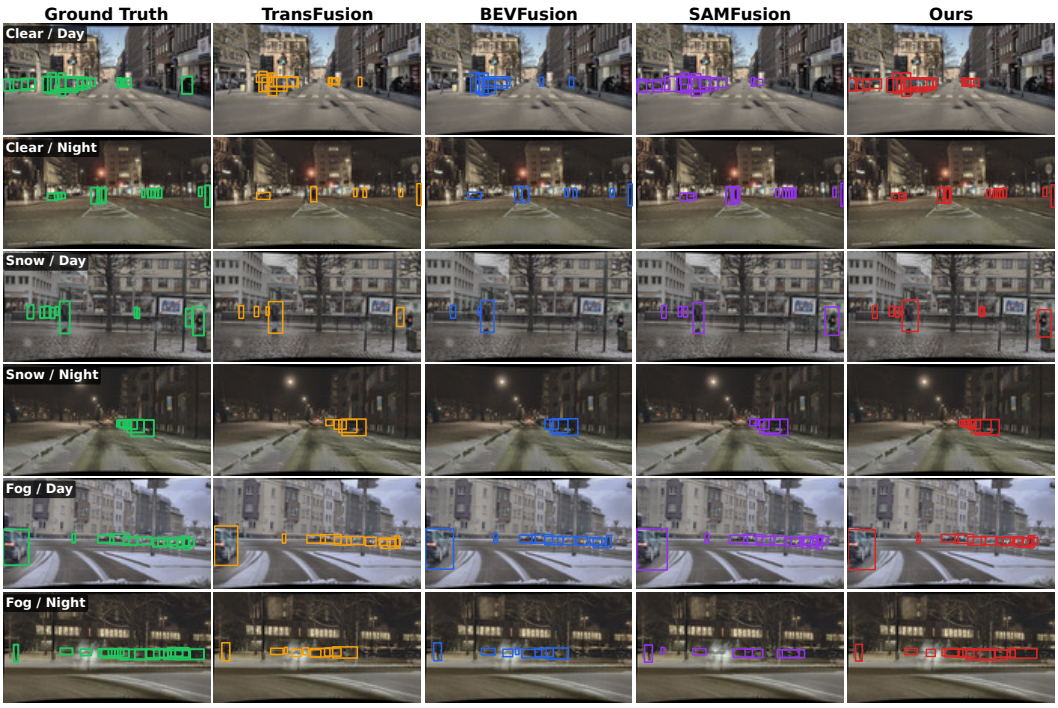

Figure 4: Cross-condition RGB comparison. Rows cover clear, snow, and fog under day and night illumination. Columns show ground truth, TransFusion, BEVFusion, SAMFusion, and ours on matched frames at a common confidence threshold of 0.10.

responses, while projected LiDAR support supplies an independent geometric cue. For this pair, the pivot has higher fused reliability (0.3211 versus 0.2929), and the competing score is reduced from 0.2508 to 0.2329 while the pivot remains at 0.2536. Across the frame, isolated or well-supported detections remain nearly unchanged, whereas selected ambiguous candidates receive stronger decay. This case illustrates the intended conservative role of SG-GSR rather than a global recalibration of all predictions.

D Additional Qualitative Results

Figure 4 complements the main paper's day/night and weather comparisons with six matched RGB examples. In the clear/day, snow/day, snow/night, and fog/night rows, ours retains boxes at several annotated locations missed by at least one baseline. The clear/night and fog/day rows reveal another pattern: methods also differ in extra or overlapping boxes. Together, these examples illustrate both missing and redundant predictions, rather than a change in box count alone. Because each row uses the same frame and display threshold, the local differences are directly comparable.

The compact grid surveys six conditions but does not resolve individual matches or 3D placement reliably. Figure 5 therefore examines related error types in separate, enlarged RGB and LiDAR BEV examples. The Figure 4 frames were selected from predefined condition groups using a recorded prediction/annotation audit at the same display threshold. This illustrative selection did not alter predictions or quantitative evaluation, and the six examples do not estimate the frequency of these errors. Condition-level trends are established by the quantitative comparisons in the main paper.

In Figure 5(a), our model detects an annotated PassengerCar for which SAMFusion has no overlapping prediction at the display threshold. In (b), BEVFusion predicts a RidableVehicle on a roadside marker with no nearby matching annotation, whereas our model makes no corresponding prediction. In (c), for an annotated PassengerCar at approximately 68 m, the ground-plane center error is 0.09 m for ours versus 1.60 m for TransFusion. The RGB crops locate the inspected region, while

the paired BEV crops expose the spatial discrepancy. These annotation-assisted examples illustrate distinct errors of complete detectors; they neither estimate error frequency nor isolate SG-GSR. Its candidate-level effect is examined separately in Figure 3.

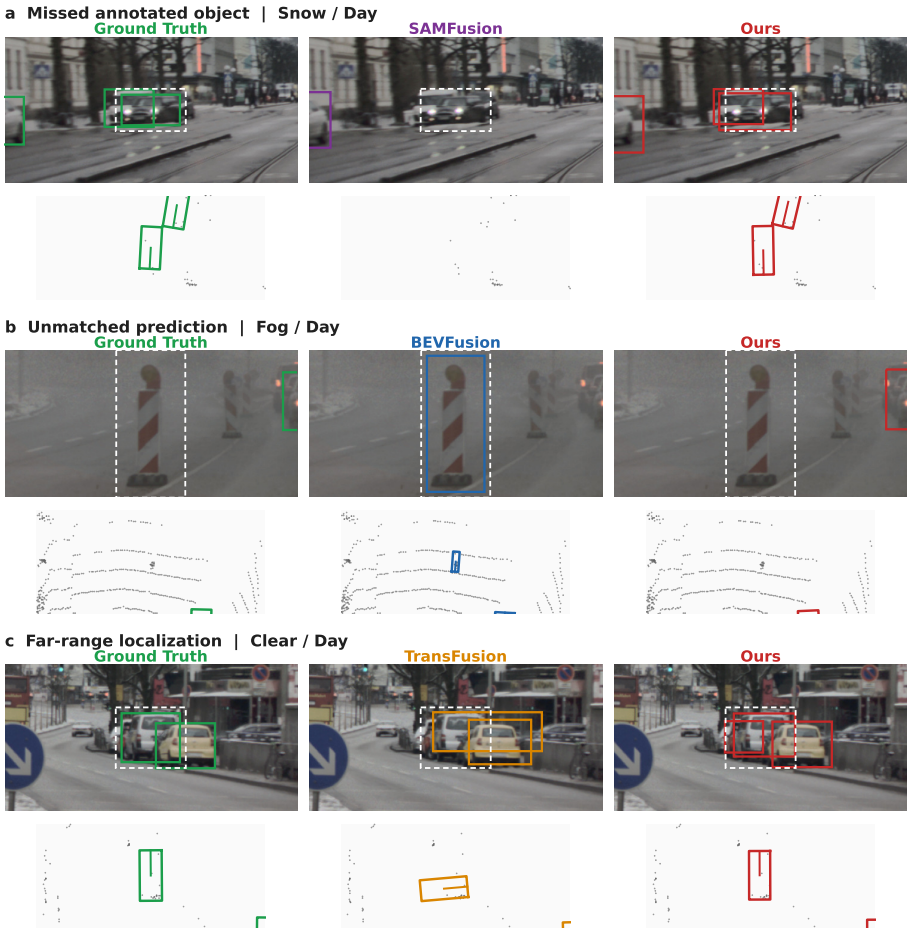

Figure 5: Focused RGB (top) and LiDAR BEV (bottom) error cases: (a) an annotated car missed by SAMFusion in snow/day, (b) a BEVFusion prediction without a nearby annotation in fog/day, and (c) a distant car localized more closely to its annotation than by TransFusion in clear/day. Each case compares ground truth, the indicated baseline, and ours at a common score threshold of 0.10. Dashed white boxes mark the inspected image region, not detections.

E Distance-Conditioned Slice Routing

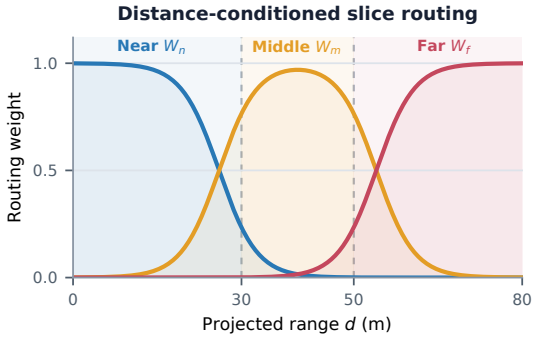

Figure 6: Distance-conditioned routing in GCSA. The normalized weights $W_n(d)$, $W_m(d)$, and $W_f(d)$ encode overlapping near-, middle-, and far-range slice preferences.

Figure 6 visualizes the routing behavior as a function of range. The learned profiles follow the expected acquisition geometry: near-, middle-, and far-range slices dominate progressively with distance. Their overlapping transitions avoid abrupt feature changes around gate boundaries, allowing GCSA to preserve range selectivity while maintaining continuous spatial aggregation.

The overlap between adjacent profiles is equally important as their dominant regions. A hard assignment would make the representation discontinuous near a gate boundary and amplify small projection errors. Soft transitions instead permit neighboring slices to contribute when range estimates are uncertain. The routing curves thus encode a smooth acquisition prior rather than a rigid partition of the image plane.

F Classwise SG-GSR Analysis

Pedestrian-only SG-GSR accounts for 0.4850 of the complete 0.4944-point gain, whereas PassengerCar contributes 0.0095 and RidableVehicle remains unchanged. Full SG-GSR raises Pedestrian AP by 0.9758, 1.7850, and 1.6037 points from near to far range. PassengerCar changes by only -0.0043, $+0.0325$, and $+0.0565$ AP, while RidableVehicle is preserved by its fixed Gaussian recipe. The classwise policy therefore concentrates refinement on ambiguous Pedestrian rankings without perturbing categories whose confidence ordering is already reliable.

Table 4: Classwise analysis of SG-GSR on the test split. Each class-specific row activates refinement only for the indicated category, while the all-class variant composes the fixed class recipes without cross-class suppression. All entries report distance-stratified 3D AP_{40}.

Refinement	Scope	Pedestrian			RidableVehicle			PassengerCar			Overall
		0–30m	30–50m	50–80m	0–30m	30–50m	50–80m	0–30m	30–50m	50–80m	
Original NMS	All	73.9240	61.9881	28.7036	88.6058	66.1630	45.5244	94.1163	88.8873	75.1118	69.2249
Gaussian decay	Ped.	74.8060	63.5716	30.0809	88.6058	66.1630	45.5244	94.1163	88.8873	75.1118	69.6519
SG-GSR	Ped.	**74.8998**	**63.7731**	**30.3073**	88.6058	66.1630	45.5244	94.1163	88.8873	75.1118	69.7099
SG-GSR	Rid.	73.9240	61.9881	28.7036	88.6058	66.1630	45.5244	94.1163	88.8873	75.1118	69.2249
SG-GSR	Car	73.9240	61.9881	28.7036	88.6058	66.1630	45.5244	**94.1120**	**88.9198**	**75.1683**	69.2344
All-class SG-GSR	All	**74.8998**	**63.7731**	**30.3073**	88.6058	66.1630	45.5244	**94.1120**	**88.9198**	**75.1683**	**69.7193**

G Applicability and Limitations

GCSA assumes synchronized and calibrated LiDAR, RGB, and gated observations. Its geometric conditions become less informative when projected LiDAR support is extremely sparse or calibration error displaces valid correspondences. The shared gated encoder also prioritizes parameter efficiency over complete slice-specific specialization. As shown by the efficiency comparison in

the main paper, independent encoders retain a small accuracy advantage, but require substantially greater gated-branch capacity.

SG-GSR is intentionally restricted to confidence refinement. It cannot recover an object absent from the detector candidate set, correct an inaccurate box, or repair a class error. Its gains are consequently largest for categories with ambiguous overlapping predictions and smaller for already saturated categories. The current evaluation is conducted on SeeingThroughFog with a fixed sensor configuration. Evaluating transfer across alternative gated timing profiles, calibration regimes, and adverse-weather datasets remains an important next step.

REFERENCES

Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11682–11692, 2020.

Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1483–1498, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Edoardo Palladin, Roland Dietze, Praveen Narayanan, Mario Bijelic, and Felix Heide. SAMFusion: Sensor-adaptive multimodal fusion for 3d object detection in adverse weather. In *European Conference on Computer Vision*, 2024.

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*, 2022.