

Degradation-Aware Modality Trust Estimation for Camera--LiDAR BEV Fusion in 3D Object Detection

Weifeng Xu

North China Electric Power University

Yaqian Cui

North China Electric Power University

Hongtao Wang

wanght@ncepu.edu.cn

North China Electric Power University

Research Article

Keywords: Sensor fusion, 3D object detection, Camera--LiDAR fusion, Bird's-eye-view perception, Robust perception, Visual degradation

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11104085/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Degradation-Aware Modality Trust Estimation for Camera–LiDAR BEV Fusion in 3D Object Detection

Weifeng Xu¹, Yaqian Cui¹, Hongtao Wang^{1*}

^{1*}North China Electric Power University, 619 Yonghua North Street,
Baoding, Hebei, 071003, China.

*Corresponding author(s). E-mail(s): wanght@ncepu.edu.cn;
Contributing authors: weifengxu@ncepu.edu.cn; yaqiancui@ncepu.edu.cn;

Abstract

Camera–LiDAR bird’s-eye-view (BEV) fusion improves 3D object detection by combining dense visual semantics with accurate geometric structure. This benefit, however, relies on the camera BEV features providing reliable visual evidence. When visual sensing is degraded, corrupted patterns can be encoded into camera BEV features, and features with large activations may come from degradation-induced interference rather than useful scene content. Directly fusing such features with LiDAR BEV features may propagate unreliable visual information into the joint representation. This paper proposes DARFusion, a lightweight Degradation-Aware and Reliability-Guided fusion method for modality trust estimation in camera–LiDAR BEV fusion. DARFusion reformulates BEV fusion as reliability-conditioned modality trust estimation. It first calibrates camera and LiDAR BEV features using learned modality reliability cues, and then estimates how much visual evidence should be fused using the calibrated features together with task-driven auxiliary contexts and compact camera degradation descriptors. We evaluate DARFusion on nuScenes and KITTI using clean validation and controlled camera degradations, with LiDAR left unchanged in the controlled settings. We further evaluate real-world rain and night subsets from nuScenes and assess cross-dataset robustness consistency. DARFusion preserves clean-set performance, achieving 68.4 mean average precision (mAP) and 71.5 nuScenes Detection Score (NDS) on nuScenes, and improves robustness under real-world adverse conditions and most controlled camera degradations. These findings indicate that degradation-aware reliability conditioning is an effective and lightweight way to improve BEV fusion reliability under active camera degradation.

Keywords: Sensor fusion, 3D object detection, Camera–LiDAR fusion, Bird’s-eye-view perception, Robust perception, Visual degradation

1 Introduction

Accurate 3D object detection is essential for autonomous driving perception, as it enables vehicles to localize and recognize surrounding objects in complex traffic scenes [1]. In this task, LiDAR and cameras provide complementary evidence. LiDAR point clouds offer accurate geometric structure, while multi-view images provide dense semantic and appearance cues [2–5]. Building on these complementary cues, existing methods span camera-based, LiDAR-based, and camera–LiDAR designs. Among these designs, camera–LiDAR fusion is particularly attractive because it can jointly exploit accurate geometry and dense visual semantics.

A natural way to exploit this complementarity is to represent both modalities in a bird’s-eye-view (BEV) space, which provides a common top-down view of the road scene. BEV-based camera–LiDAR detectors transform visual and geometric features into this shared representation and perform cross-modal interaction before prediction [6–11]. This unified BEV representation enables spatially aligned cross-modal interaction and has shown a favorable balance between detection accuracy and deployment efficiency in multimodal 3D perception.

Despite this progress, BEV fusion still faces a reliability bottleneck under degraded visual sensing because the strength of a feature response does not necessarily reflect its reliability. Robustness studies have shown that 3D and BEV perception can degrade under adverse weather, blur, noise, low illumination, and other sensor corruptions [12–15]. The same camera BEV response can therefore have different meanings under different imaging conditions. In a clear scene, a strong visual response may indicate useful semantic evidence; under glare, noise, blur, or low illumination, it may instead come from corrupted texture, smeared edges, or sensor artifacts. Such corrupted visual patterns become embedded in the camera features and enter the fused representation as if they were useful visual evidence. Some BEV fusion methods combine camera and LiDAR BEV features through direct concatenation [9] or convolutional fusion [8], while recent variants introduce attention [10] or bidirectional interaction [11] to enhance cross-modal feature exchange. However, these designs do not explicitly address the reliability issue caused by degraded camera sensing. Once degradation-induced visual patterns are lifted into BEV space, a content-driven fusion module may mistake them for task-relevant evidence and retain unreliable camera information in the fused representation, even when LiDAR provides stable geometric information.

To address the reliability ambiguity of feature responses under degraded camera sensing, we propose DARFusion, a Degradation-Aware and Reliability-Guided fusion method that reformulates BEV camera–LiDAR fusion as a modality trust estimation problem. Instead of using feature responses directly as fusion evidence, DARFusion conditions the fusion decision on reliability cues, task-driven auxiliary contexts, and camera BEV feature-derived degradation descriptors. In this design, reliability cues adjust modality features, the auxiliary contexts provide task-driven modality context,

and degradation descriptors make the camera sensing state visible in the same BEV representation where fusion is performed. Together, these cues help separate feature representation from modality trust estimation under degraded sensing, while the resulting module remains lightweight and avoids degradation-label supervision.

The contributions of this paper are summarized as follows:

- We identify and formulate a reliability limitation in content-driven BEV fusion under degraded visual sensing, where feature responses can be entangled with modality trust.
- We propose DARFusion, a lightweight fusion method that calibrates modality features with learned reliability cues and predicts BEV fusion weights using calibrated features, task-driven auxiliary contexts, and camera BEV feature-derived degradation descriptors.
- We conduct a degradation-oriented evaluation on nuScenes [16] and KITTI [17], covering clean validation, real-world rain and night subsets from nuScenes, controlled synthetic camera corruptions, and cross-dataset robustness consistency, to analyze robustness trends under degraded camera sensing.

2 Related Work

2.1 BEV Camera–LiDAR Fusion

Recent camera–LiDAR detectors increasingly fuse heterogeneous features in a unified 3D or BEV representation. Early fusion methods combine image semantics with point-cloud representations through point painting or multimodal voxel features [18, 19]. Transformer-based and BEV-space methods further learn unified representations for cross-modal detection [6, 7, 20]. BEVFusion establishes a simple BEV-space fusion paradigm by lifting image features and combining them with LiDAR BEV features before the detection head [8, 9]. Recent variants improve this paradigm through attention-enhanced fusion [10] and bidirectional complementary interaction [11]. Parallel advances in 3D detection improve different components of the representation pipeline: RFPillarNeXt reallocates sparse convolution modules to better capture fine-grained LiDAR geometry [21], CrossBEV enhances multi-view camera detection through spatiotemporal feature cross-enhancement [22], and GLFF3D combines global and local image cues with point-cloud proposals for indoor detection [23]. Collectively, these methods mainly address spatial alignment, feature exchange, and complementary representation learning. They improve representation quality or cross-modal interaction, but do not explicitly model how much camera evidence should be trusted when the camera stream remains active but degraded. In contrast, we study a different aspect of BEV fusion, namely how much the fusion module should trust camera BEV features when the camera stream remains available but is degraded.

2.2 Reliability Modeling in Multimodal Fusion

Robust multimodal fusion often requires adjusting the reliance on different sensors when their observations become unreliable. Existing studies mainly address this problem from the perspective of sensor reliability or sensor failure. For example, Wozniak et

al. [24] study the fusion step under corrupted inputs such as sensor misalignment and missing LiDAR information. Failure-aware BEV detectors further improve robustness to missing or severely corrupted modalities through adaptive BEV queries [25], uniform modality encoders [26], or expert routing [27]. Related uncertainty-guided fusion has also been explored in LiDAR-camera 3D panoptic segmentation to reduce the influence of unreliable features under imperfect sensing [28]. These studies show that multimodal perception should not treat all sensor evidence as equally reliable. They usually assume that a modality is missing, severely corrupted, or accompanied by explicit uncertainty estimates.

Our focus differs from these settings. We study active camera degradation, where the camera stream remains available and continues to produce BEV features, but the visual evidence may be affected by low illumination, blur, or noise. In this case, the key issue is not only whether the camera modality should be trusted, but also how degraded visual sensing affects the reliability of its BEV features. DARFusion therefore incorporates camera BEV feature-derived degradation descriptors into reliability-guided BEV fusion, allowing modality trust estimation to account for the visual sensing state without degradation labels or image-quality supervision.

2.3 Robust 3D Perception Under Visual Degradation

Robustness benchmarks reveal that clean-set accuracy does not necessarily translate to reliable performance under corrupted sensing. Robo3D and related benchmarks evaluate 3D detectors under adverse weather, sensor noise, blur, and low illumination [12, 13]. RoboBEV studies BEV perception robustness under corrupted multi-view images [14], while LiDAR-camera robustness benchmarks and MultiCorrupt analyze multimodal degradation under diverse sensor corruptions [15, 29]. Context-aware fusion further studies adverse lighting and weather by injecting operating-condition cues into multi-sensor fusion [30]. Camera-side failures such as lens soiling highlight visual degradation as a practical failure source in autonomous driving perception [31]. These studies are valuable for diagnosing failure modes, but they mainly report detector-level robustness under corrupted inputs and leave the BEV fusion-side use of camera degradation cues largely open. Different from methods that rely on external operating-condition labels, DARFusion derives compact sensing-state cues from camera BEV features and uses them only as conditioning variables for a learned fusion gate. This motivates exposing cues related to brightness, blur, and noise to BEV fusion, so that the fusion gate does not rely solely on corrupted feature activations.

3 Method

3.1 Overview

Many conventional BEV fusion modules derive camera–LiDAR interactions from aligned feature content. This content-driven design is effective when both sensors are clean, but it has limited access to why a modality may be unreliable. DARFusion changes the conditioning variables of the fusion gate. Instead of learning a mapping from BEV feature content to fusion weights, it learns a mapping from feature content,

modality reliability, task-driven auxiliary contexts, and camera degradation state to modality trust. Figure 1 illustrates the overall pipeline at the BEV fusion stage. Given multi-view images and a LiDAR point cloud, the camera and LiDAR encoders produce camera BEV feature F_c and LiDAR BEV feature F_l respectively.

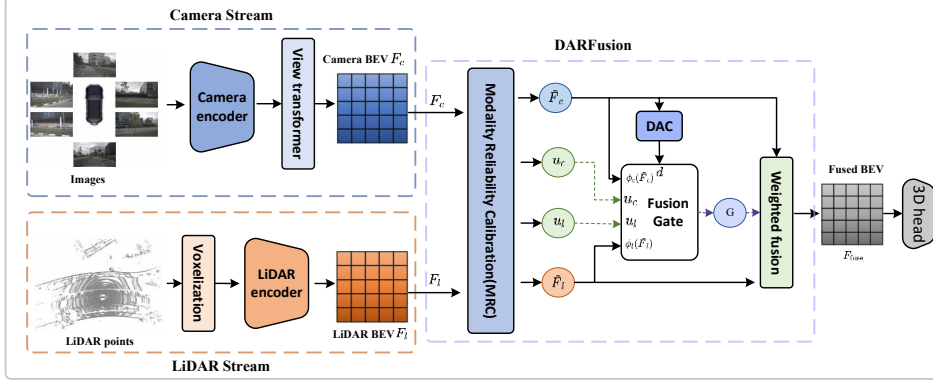


Fig. 1 Overview of DARFusion. Modality Reliability Calibration (MRC) first calibrates camera and LiDAR BEV features and produces task-driven auxiliary contexts. Degradation-Aware Conditioning (DAC) extracts a compact degradation descriptor from the calibrated camera branch. The Fusion Gate then uses the projected calibrated features, task-driven auxiliary contexts, and degradation descriptor to predict the spatial fusion weight for weighted BEV fusion.

DARFusion takes F_c and F_l as input and outputs a fused BEV feature F_{fuse} for the detection head.

In DARFusion, Modality Reliability Calibration (MRC) first models modality reliability and outputs the recalibrated features $\tilde{F}_c$ and $\tilde{F}_l$, together with task-driven auxiliary contexts u_c and u_l ; we denote these contexts jointly as $U = \{u_c, u_l\}$. Degradation-Aware Conditioning (DAC) then extracts a degradation descriptor d from the calibrated camera branch to expose the visual degradation state of the camera feature. The Fusion Gate combines the projected calibrated modality features, the MRC-derived contexts in U , and the DAC descriptor to predict the camera-fusion gate $\mathbf{G}_c$ and the complementary LiDAR-fusion gate $\mathbf{G}_l$. Finally, weighted fusion uses the two gates to fuse the camera and LiDAR branches and obtains the fused BEV feature F_{fuse} .

This design differs from conventional BEVFusion-style layers that concatenate camera and LiDAR BEV features and then apply convolutional fusion. When camera sensing is degraded, such feature-level fusion based only on F_c and F_l is insufficient to indicate whether the visual evidence should be trusted. DARFusion therefore introduces MRC-derived task-driven auxiliary contexts U and DAC-based camera degradation context d before fusion, allowing the Fusion Gate to estimate not only where the two modalities are informative, but also how reliable the camera evidence is under the current sensing condition.

3.2 Modality Reliability Calibration

Reliability calibration answers which modality should be globally strengthened or suppressed before spatial fusion. Here, calibration refers to task-driven feature recalibration rather than statistical uncertainty calibration. For modality $m \in \{c, l\}$, DARFusion applies a global average pooling (GAP) operation to obtain a compact descriptor

$$s_m = \text{GAP}(F_m).$$

This descriptor is passed to two lightweight learnable branches. Both branches use a common 1×1 convolutional embedding, followed by a branch-specific 1×1 output convolution. The first branch applies a sigmoid to $g_m^{\text{rel}}(s_m)$ to obtain a reliability cue $q_m \in [0, 1]$, and the second branch predicts a log-variance value ℓ_m as:

$$q_m = \sigma(g_m^{\text{rel}}(s_m)), \quad \ell_m = g_m^{\text{var}}(s_m),$$

where $\sigma(\cdot)$ denotes the sigmoid function. Here, q_m is not supervised as an absolute sensor-quality score; it is a task-driven cue learned through the detection loss. Each feature is then recalibrated by

$$\tilde{F}_m = q_m \odot F_m, \quad m \in \{c, l\},$$

where $\odot$ denotes broadcast multiplication. The output of the log-variance branch is converted into a positive task-driven auxiliary context:

$$u_m = \exp(\ell_m), \quad m \in \{c, l\}.$$

We use the task-driven auxiliary context u_m as an input to the fusion predictor. Since no likelihood, statistical-calibration, ensemble, or dropout objective is used, u_m is not interpreted as calibrated Bayesian or probabilistic uncertainty.

3.3 Degradation-Aware Conditioning

Reliability cues and task-driven auxiliary contexts provide modality-level context, but they do not directly expose the sensing state that makes a camera response suspicious. This branch targets the active-camera case, where the camera feature is still available but its response may be affected by imaging degradation. DARFusion therefore extracts a compact degradation descriptor from the calibrated camera BEV feature $\tilde{F}_c$.

We first project the camera BEV feature to a single-channel response map

$$X = \psi(\tilde{F}_c),$$

where $\psi(\cdot)$ is a 1×1 convolution. From X , we compute three response-statistics proxies: an activation-level proxy r_b , an edge-response proxy r_s , and a high-frequency residual proxy r_n :

$$r_b = \text{Mean}(X),$$

$$r_s = \frac{1}{1 + \text{Mean}(|\Delta(X)|)},$$

$$r_n = \text{Std}(X - \text{AvgPool}(X)).$$

Here, $\text{Mean}(\cdot)$ and $\text{Std}(\cdot)$ compute the spatial mean and standard deviation over the BEV dimensions for each sample. $\Delta(\cdot)$ denotes a spatial Laplacian filter applied to X , and $\text{AvgPool}(\cdot)$ denotes local average pooling on the BEV response map. The activation-level proxy r_b summarizes the global activation level of the projected camera BEV response, the edge-response proxy r_s grows as edge responses weaken, and the high-frequency residual proxy r_n captures high-frequency perturbations in local residuals. These proxies are derived from BEV feature responses rather than direct image-quality measurements.

Each raw statistic is mapped to $[0, 1]$ by a learnable affine transform with parameters a_i and b_i , followed by a sigmoid:

$$d_i = \sigma(a_i r_i + b_i), \quad i \in \{b, s, n\}.$$

We define the degradation descriptor as $d = [d_b, d_s, d_n]$, a compact summary derived from camera BEV features to expose response patterns associated with illumination, blur, and noise. It is provided to the fusion gate as camera-state context, helping the gate judge whether camera BEV responses are likely to represent reliable visual evidence under the current sensing condition. The descriptors are task-driven conditioning cues rather than hand-crafted physical-quality measurements. By conditioning fusion on d , DARFusion allows the camera degradation state to influence modality trust estimation and the final spatial fusion weight.

3.4 Degradation-Conditioned Fusion

At the fusion stage, DARFusion uses a lightweight gate to estimate how much camera evidence should be used at each BEV location. The gate takes the projected camera and LiDAR BEV features $\bar{F}_c, \bar{F}_l$ together with MRC-derived task-driven auxiliary contexts u_c, u_l and the DAC degradation descriptor d , allowing the fusion weight to depend on both local BEV evidence and reliability context. The resulting camera and LiDAR gate matrices, $\mathbf{G}_c$ and $\mathbf{G}_l$, are applied through weighted fusion, which provides a continuous and differentiable way to balance camera semantics and LiDAR geometry while preserving their complementary information.

After reliability calibration, the camera and LiDAR features are projected into a common channel dimension:

$$\bar{F}_c = \phi_c(\tilde{F}_c), \quad \bar{F}_l = \phi_l(\tilde{F}_l),$$

where $\tilde{F}_c$ and $\tilde{F}_l$ are reliability-calibrated features, $\phi_c(\cdot)$ and $\phi_l(\cdot)$ are lightweight projection layers.

The lightweight gate takes the projected features $\bar{F}_c, \bar{F}_l$, task-driven auxiliary contexts u_c, u_l , and degradation descriptor d as input. We denote the gate input by

$$Z = [\bar{F}_c, \bar{F}_l, u_c, u_l, d],$$

where $[\cdot]$ denotes channel-wise concatenation. Before concatenation, the sample-level contexts u_c, u_l and the three components of d are expanded to the $H \times W$ BEV grid as additional channels. The gate is implemented as two 1×1 convolutional layers with ReLU between them, followed by a sigmoid activation:

$$\mathbf{G}_c = \sigma\left(\text{Conv}_{1 \times 1}^{(2)}\left(\text{ReLU}\left(\text{Conv}_{1 \times 1}^{(1)}(Z)\right)\right)\right),$$

Compared with feature-only gated fusion, this predictor receives not only local BEV content but also task-driven auxiliary context and the visual degradation state of the camera branch. Reliability calibration and the task-driven auxiliary contexts provide sample-level modality context, while spatial variation is modeled by the gating network through local BEV features. The final fused feature is then computed with the weighted sum as

$$F_{\text{fuse}} = \mathbf{G}_c \odot \bar{F}_c + \mathbf{G}_l \odot \bar{F}_l,$$

where $\mathbf{G}_c, \mathbf{G}_l \in [0, 1]^{H \times W}$, $\mathbf{G}_c + \mathbf{G}_l = \mathbf{1}$, and H and W denote the height and width of the BEV feature map. Here, $\mathbf{1}$ denotes the all-ones BEV matrix.

4 Experiments

4.1 Datasets and Evaluation Metrics

The experiments are designed to answer four questions: **(Q1)** whether DARFusion preserves clean-set detection accuracy; **(Q2)** whether its robustness improvements are generally consistent across datasets and sensor configurations; **(Q3)** how the individual conditioning components affect detection; and **(Q4)** whether the learned descriptors and fusion behavior respond consistently to increasing camera degradation.

Our main experiments are conducted on nuScenes [16], which provides large-scale autonomous-driving scenes with multi-view camera images and LiDAR point clouds. We report mean average precision (mAP), nuScenes Detection Score (NDS), and class-wise average precision (AP). To assess cross-dataset consistency, we additionally evaluate KITTI [17]. KITTI results use the official 3D AP_{R40} for Car, Pedestrian, and Cyclist under the Easy, Moderate, and Hard difficulty levels; unless stated otherwise, KITTI robustness is summarized by the mean Moderate 3D AP_{R40} over the three classes. We additionally report parameter count and throughput.

4.2 Implementation Details

The base detector follows a BEVFusion-style architecture. In our reproduced nuScenes experiments, all fusion variants use an ImageNet-pretrained Swin-Tiny image backbone and a SECOND LiDAR encoder with a voxel size of $0.075 \times 0.075 \times 0.2$ m. The detection range is $[-54, 54]$ m along both BEV axes, the BEV grid resolution is 0.3 m, and the multi-view camera input is cropped and scaled to 256×704 . Training follows a two-stage protocol: the LiDAR branch is first trained, and the camera-LiDAR fusion model is then initialized from the same LiDAR checkpoint. All models are trained on two RTX 4090 GPUs with AdamW, learning rate 2×10^{-4} , weight decay 0.01, and identical data augmentations.

For KITTI, both the reproduced BEVFusion baseline and DARFusion are initialized from the same LiDAR-only checkpoint and use a Swin-Tiny image backbone, a SECOND-based LiDAR encoder, a front-view input resolution of 720×1280 , a voxel size of $0.075 \times 0.075 \times 0.2$ m, and a six-epoch fusion-stage schedule. Except for the fusion layer under study, all training, inference, and checkpoint-selection settings are identical. Models are trained only on clean data; camera corruptions are applied at evaluation time while LiDAR and calibration parameters remain unchanged.

Statistical protocol. The two principal models—BEVFusion and DARFusion—are evaluated on both nuScenes and KITTI over three random seeds. Principal comparisons are reported as means over the three seeds; the real-world rain and night table additionally reports seed-level standard deviation. Hyperparameters and corruption definitions are fixed before inspecting the final results, and the same checkpoint-selection rule is used for every method.

4.3 Comparative Experiments

4.3.1 Results on nuScenes

Table 1 reports clean validation performance together with representative 3D detection methods for context. Under the same reproduced BEVFusion setting, DARFusion improves the baseline from 67.8 to 68.4 mAP and from 71.0 to 71.5 NDS. DARFusion also achieves higher clean-set NDS than several representative camera–LiDAR detectors listed in the table, including AttBEV and KeyGeoFusion.

Table 1 Performance comparison with representative 3D detection methods on the nuScenes validation set. mAP and NDS are reported on the 0-100 scale; the reproduced BEVFusion and DARFusion results are means over three seeds. Published methods may use different backbones, input resolutions, voxelization settings, and training schedules; this table is intended to contextualize clean-set validation performance. The highest mAP and NDS among the listed values are shown in bold.

Method	Modality	Image Enc.	LiDAR Enc.	mAP	NDS
SECOND [32]	LiDAR	–	SECOND	32.2	45.3
PointPillars [2]	LiDAR	–	PointPillars	30.5	45.3
CenterPoint [3]	LiDAR	–	VoxelNet	58.0	65.5
BEVFormer [4]	Camera	ResNet-101	–	41.6	51.7
TransFusion [6]	LiDAR+Cam	ResNet-50	VoxelNet	67.5	71.3
AttBEV [10]	LiDAR+Cam	Swin-T	DynamicVFE+SECOND	64.3	68.0
KeyGeoFusion [33]	LiDAR+Cam	HRNet/SegFormer-B2	SECOND	62.1	68.5
BEVFusion (Base) [8]	LiDAR+Cam	Swin-T	SECOND	67.8	71.0
DARFusion (Ours)	LiDAR+Cam	Swin-T	SECOND	68.4	71.5

Table 2 further breaks down the clean validation result by object category. It compares BEVFusion and DARFusion under the same backbone, LiDAR encoder, and training protocol, reporting class-wise AP together with mAP and NDS. The gain is not concentrated in a single category: DARFusion improves most categories, including Truck, C.V., Bus, Trailer, Bicycle, Pedestrian, and T.C., while only Barrier and Motorcycle are lower than the baseline. For Car, the two methods achieve the same AP after rounding. The largest margins appear on large rigid objects, especially Truck (+1.7 AP), Bus (+1.6 AP), and Trailer (+0.9 AP), suggesting better detection performance on large-object categories.

Table 2 Per-class detection performance on the nuScenes validation set. Class AP, mAP, and NDS are reported on the 0-100 scale and averaged over three seeds. The better value in each column is shown in bold.

Method	Car	Truck	C.V.	Bus	Trailer	Barrier	Motor.	Bicycle	Ped.	T.C.	mAP	NDS
BEVFusion	89.2	62.8	29.5	76.8	43.5	70.6	76.5	63.0	87.9	78.1	67.8	71.0
DARFusion	89.2	64.5	29.9	78.4	44.4	70.5	76.4	63.4	88.4	78.7	68.4	71.5

4.3.2 Results on KITTI

Table 3 reports the clean-set KITTI results for the reproduced BEVFusion baseline and DARFusion, including the complete Easy/Moderate/Hard breakdown and the mean Moderate 3D AP_{R40} used for cross-dataset comparison.

Table 3 Clean-set 3D detection performance on the KITTI validation split. AP_{R40} is reported on the 0-100 scale as the mean over three seeds. “Mean Mod.” is the mean Moderate AP over Car, Pedestrian, and Cyclist, computed from unrounded per-class averages before rounding. The better value in each column is shown in bold.

Method	Car E	Car M	Car H	Ped. E	Ped. M	Ped. H	Cyc. E	Cyc. M	Cyc. H	Mean Mod.
BEVFusion	91.8	81.9	78.8	66.3	60.9	56.7	93.6	85.9	80.8	76.2
DARFusion (Ours)	92.1	82.6	79.5	67.5	62.0	57.7	94.0	86.9	81.8	77.1

Result interpretation. DARFusion improves the mean Moderate 3D AP_{R40} from 76.2 to 77.1 (+0.9). The 3D gains are consistent across the reported Easy, Moderate, and Hard settings for Car, Pedestrian, and Cyclist; the largest Moderate gain occurs on Pedestrian (+1.1), followed by Cyclist (+1.0) and Car (+0.7). These results indicate that DARFusion preserves and slightly improves clean-set accuracy relative to the reproduced BEVFusion baseline.

4.4 Robustness Evaluation

We first evaluate robustness using real-world rain and night data. Specifically, we select non-overlapping rain and night subsets from the nuScenes validation set, containing 1088 and 602 samples, respectively. These subsets provide real-world evidence under adverse weather or low illumination while preserving the original nuScenes sensor streams. Because these subsets provide limited coverage of degradation types and severities, we complement them with controlled camera corruptions while keeping LiDAR and calibration parameters unchanged.

Synthetic night uses brightness/gamma/read-noise settings (0.75, 1.2, 4), (0.60, 1.5, 8), and (0.45, 1.8, 12) for low, mid, and high levels, with Poisson scale 30. Motion blur is generated by a camera-aware spatial exposure model in which k controls the base image-space blur length, and Gaussian noise uses zero-mean perturbations with standard deviation σ . The detailed nuScenes evaluation uses $k = 1, 5, 9, 13, 17, 21$ and $\sigma = 5, 10, 15, 20, 25, 30$. The cost-controlled KITTI study uses $k = 5, 13, 21$, $\sigma = 10, 20, 30$, and the same three synthetic-night levels. Random perturbations are generated once with fixed seeds and shared by all methods and training seeds.

4.4.1 Real-World Degraded Scenes

Table 4 evaluates DARFusion across real rain and night scenes. DARFusion improves mAP on the full validation set, rain subset, and night subset, with the largest gain under rain (+1.41 mAP) followed by night (+1.31 mAP). These two subset gains are larger than the overall validation gain, suggesting that the improvement is more pronounced under adverse weather or low illumination. This indicates that DARFusion is particularly useful for robustness-oriented fusion under degraded camera sensing, rather than merely increasing clean-set accuracy.

Table 4 Robustness on real rain and night scenes. mAP is reported on the 0-100 scale as mean $\pm$ standard deviation over three seeds. The rain and night subsets are non-overlapping. Sample counts are shown in parentheses, and Δ is the DARFusion gain over BEVFusion.

Stratum	BEVFusion	DARFusion	Δ DAR
All (6019)	67.82 $\pm$ 0.02	68.41 $\pm$ 0.06	+0.59
Rain (1088)	68.68 $\pm$ 0.18	70.09 $\pm$ 0.14	+1.41
Night (602)	40.26 $\pm$ 0.20	41.57 $\pm$ 0.04	+1.31

4.4.2 Controlled Camera Degradations

To more systematically test robustness under degraded camera sensing, we construct three simulated visual degradations on the validation images, including synthetic night, motion blur, and Gaussian noise. Figure 2 provides a visual comparison of these simulated degradations under the same camera view.

Synthetic Night. Table 5 shows how the two methods perform as the synthetic night degradation becomes stronger. DARFusion remains higher across all night-degradation settings and gives the largest gain under *night-mid* (+1.1 mAP). Under the strongest *night-high* setting, both methods are heavily affected because the camera images contain much less usable visual evidence, so the relative gain becomes smaller.

Table 5 Synthetic night-time degradation. mAP is reported on the 0-100 scale as the mean over three seeds; all standard deviations are below 0.052.

Method	Clean	Night-low	Night-mid	Night-high
BEVFusion	67.8	66.1	64.3	62.5
DARFusion	68.4	66.9	65.4	62.7
Δ DAR	+0.6	+0.8	+1.1	+0.2

Motion Blur. Table 6 shows the results under increasing motion-blur strength. With the proposed camera-aware spatial blur model, DARFusion keeps higher mAP than BEVFusion across all tested blur strengths. The gain remains visible from moderate

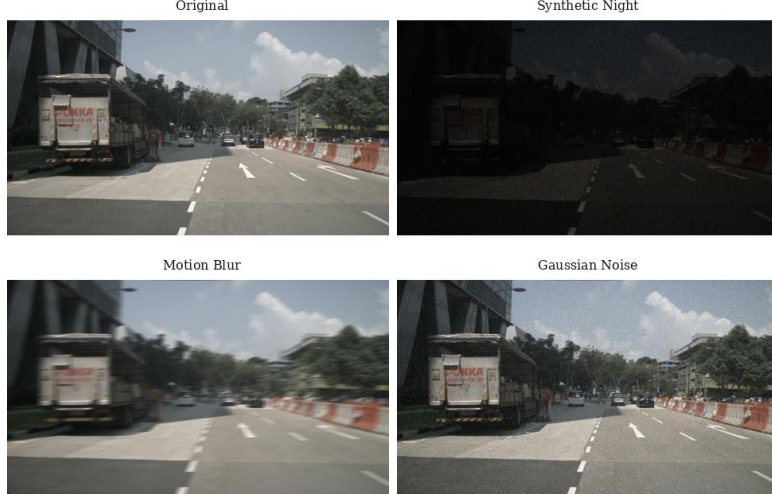


Fig. 2 Examples of controlled camera degradations used in the robustness analysis. The same nuScenes front-view image is shown under synthetic night, camera-aware spatial motion blur, and additive Gaussian noise, corresponding to the degradation types evaluated in this subsection.

blur ($k = 9$ and $k = 13$) to stronger blur ($k = 17$ and $k = 21$), suggesting better robustness when camera appearance is degraded by motion-induced blur.

Table 6 Motion-blur robustness under increasing blur strength. mAP is the mean over three seeds on the 0-100 scale, and k denotes the base image-space blur length.

Method	$k = 1$	$k = 5$	$k = 9$	$k = 13$	$k = 17$	$k = 21$
BEVFusion	67.8	67.6	67.1	66.3	64.1	62.3
DARFusion	68.4	68.1	67.9	67.1	65.3	62.9
Δ DAR	+0.6	+0.5	+0.8	+0.8	+1.2	+0.6

Gaussian Noise. Figure 3 reports mAP under additive Gaussian noise with increasing noise strength. Both methods show a gradual performance drop as σ increases. DARFusion performs better than BEVFusion under low-to-moderate noise levels, with the largest gain reaching +0.6 mAP. As the noise becomes stronger, the margin becomes smaller and the two methods are nearly tied at $\sigma = 30$. This suggests that DARFusion can improve robustness when the noise level is moderate, but very strong dense pixel-level perturbations remain difficult because they corrupt visual evidence across the whole image.

4.4.3 Cross-Dataset Robustness

Table 7 applies the clean-trained KITTI models to controlled camera degradation on KITTI. To limit experimental cost while retaining a severity trend, three representative levels are used for each corruption. Each entry is the mean Moderate 3D AP_{R40} over

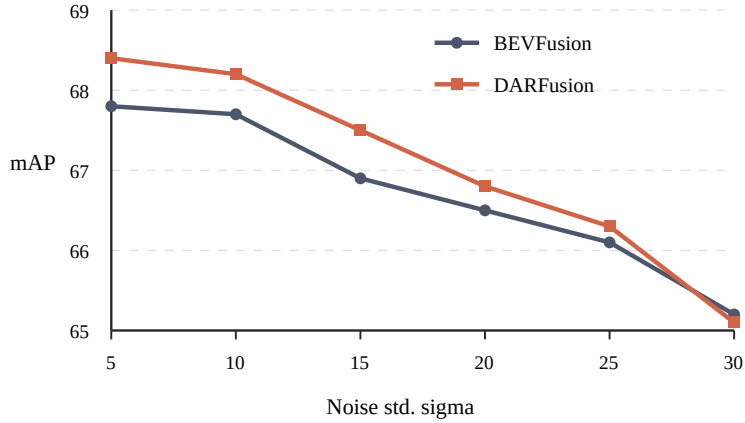


Fig. 3 Gaussian image noise robustness under increasing noise strength. mAP is the mean over three seeds on the 0-100 scale, and σ denotes the standard deviation of additive Gaussian noise.

the three KITTI classes. The clean result and the nine corruption-severity settings are reported in separate rows.

Table 7 Controlled camera-degradation robustness on KITTI. The clean result and each corruption-severity setting are reported in separate rows. Values are the mean Moderate 3D AP_{R40} over Car, Pedestrian, and Cyclist, reported as means over three seeds. The better value in each row is shown in bold.

Corruption type	Severity level	BEVFusion	DARFusion (Ours)	Δ
Clean	—	76.2	77.1	+0.9
Synthetic night	Low	75.7	76.6	+0.9
Synthetic night	Medium	74.6	75.9	+1.3
Synthetic night	High	72.8	74.0	+1.2
Motion blur	Low	75.9	76.8	+0.9
Motion blur	Medium	75.0	76.2	+1.2
Motion blur	High	73.0	74.6	+1.6
Gaussian noise	Low	75.8	76.6	+0.8
Gaussian noise	Medium	74.4	75.4	+1.0
Gaussian noise	High	72.0	72.9	+0.9

Robustness interpretation. Separate clean-trained models are used for each dataset, following the same clean-training and corruption-generation procedure. On nuScenes, DARFusion improves clean mAP from 67.8 to 68.4, remains higher under synthetic night and motion blur, and is comparable under the strongest Gaussian noise. On KITTI, Table 7 shows that DARFusion improves the clean mean Moderate 3D AP_{R40} from 76.2 to 77.1 and outperforms BEVFusion in all nine degraded settings. The largest margin is +1.6 under high motion blur, and the smallest is +0.8 under low

Gaussian noise. These results across two datasets and sensor configurations support overall cross-dataset robustness consistency.

4.5 Ablation Study

The ablation study isolates the contribution of each conditioning component. Starting from full DARFusion, the w/o DAC variant removes degradation-aware conditioning while keeping MRC and the MRC-derived task-driven auxiliary contexts u_c, u_l ; w/o U further removes these contexts from the fusion gate and keeps only MRC; and w/o MRC removes modality reliability calibration, reducing the model to the original BEVFusion-style fusion baseline.

Component Ablation. We isolate the contributions of Degradation-Aware Conditioning (DAC), the MRC-derived task-driven auxiliary contexts u_c, u_l , and Modality Reliability Calibration (MRC) under the same detector backbone, training protocol, and evaluation setting. Table 8 shows that each component contributes to the final model. Removing DAC reduces both mAP and NDS, further removing u_c, u_l lowers the clean-set operating point, and removing MRC gives the BEVFusion baseline.

Inference Cost Analysis. Deployment in autonomous driving requires low overhead. Table 8 shows that DARFusion adds 0.04M parameters (about 0.1% over BEVFusion) and reduces frames per second (FPS) from 4.57 to 4.46 on one RTX 4090. The overhead is small relative to the robustness gains under degraded sensing.

Table 8 Core module ablation and algorithmic efficiency. mAP and NDS are reported on the 0-100 scale; Parameter counts and FPS retain their original units. FPS denotes detector forward throughput in frames per second, measured on a single RTX 4090. The best value for each metric is shown in bold.

Method	MRC	U	DAC	mAP	NDS	Params	FPS
DARFusion (Ours)	✓	✓	✓	68.4	71.5	40.84M	4.46
w/o DAC	✓	✓	-	68.1	71.2	40.82M	4.49
w/o U	✓	-	-	68.0	71.0	40.81M	4.54
w/o MRC	-	-	-	67.8	71.0	40.80M	4.57

Robustness Ablation. Figure 4 further compares the same ablation variants under controlled camera degradations. The full DARFusion model achieves the best performance in most settings, with clear gains under synthetic night and motion blur. Under Gaussian noise, the margin becomes smaller, and the w/o DAC variant is slightly higher at $\sigma = 20$ and tied with the full model at $\sigma = 25$. This suggests that the proposed components improve robustness in general, while dense pixel-level noise remains a more challenging condition.

4.6 Mechanism Analysis

Figure 5 analyzes how DARFusion adjusts camera usage under controlled degradations. Since the camera gate $\mathbf{G}_c$ alone does not directly measure how much the camera

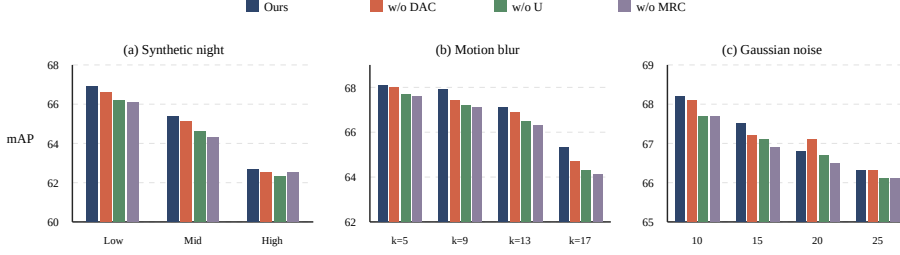


Fig. 4 Robustness ablation under controlled camera degradations. mAP is reported on the 0-100 scale. Each panel uses a truncated y-axis to show relative differences among ablation variants.

branch contributes after reliability calibration and feature projection, we focus on the magnitude-based relative camera contribution

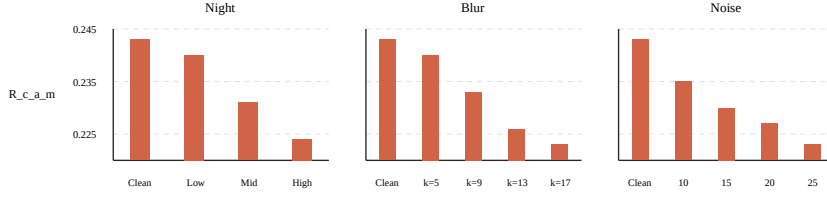


Fig. 5 Fusion behavior analysis of DARFusion under controlled camera degradations. Each panel shows the magnitude-based relative camera contribution ratio R_{cam} for clean images and increasing degradation severity. Values are averaged over 128 held-out nuScenes samples, and the y-axis is zoomed to highlight small but systematic changes.

$$R_{\text{cam}} = \frac{\|\mathbf{G}_c \odot \bar{F}_c\|_1}{\|\mathbf{G}_c \odot \bar{F}_c\|_1 + \|\mathbf{G}_l \odot \bar{F}_l\|_1}.$$

Across all three degradation types, R_{cam} decreases as the degradation severity increases. This trend indicates that DARFusion progressively reduces the effective contribution of the camera branch when the visual input becomes less reliable. The fusion module adjusts camera reliance according to the degraded sensing state through reliability-conditioned fusion and projected feature magnitudes.

Detection scores alone do not establish that the proposed conditioning responds to the intended camera states. We therefore complement the existing nuScenes camera-contribution curves in Figure 5 with the quantitative tests in Table 9. For every corruption family, severity is encoded by its ordered level and Spearman’s rank correlation is computed over paired samples. The descriptor column uses the corresponding DAC component: d_b for synthetic night, d_s for motion blur, and d_n for Gaussian noise. Confidence intervals are obtained by bootstrapping scenes rather than individual camera views.

Table 9 Mechanism analysis under increasing corruption severity on the 128 held-out nuScenes samples used in Figure 5. $\rho(d_i, s)$ is the Spearman correlation between the relevant degradation descriptor and severity rank s ; $\rho(R_{\text{cam}}, s)$ measures the response of the magnitude-based camera contribution; and ΔR_{cam} is the change from clean to the highest severity. All estimates are reported with 95% scene-bootstrap confidence intervals (CIs).

Corruption	Descriptor	$\rho(d_i, s)$ [95% CI]	$\rho(R_{\text{cam}}, s)$ [95% CI]	ΔR_{cam} [95% CI]
Synthetic night	d_b	0.70 [0.64, 0.76]	-0.62 [-0.69, -0.54]	-0.019 [-0.020, -0.017]
Motion blur	d_s	0.65 [0.58, 0.71]	-0.57 [-0.64, -0.49]	-0.020 [-0.022, -0.017]
Gaussian noise	d_n	0.55 [0.47, 0.62]	-0.42 [-0.50, -0.33]	-0.020 [-0.023, -0.016]

Mechanism interpretation. The dedicated descriptor correlations with severity are positive, whereas the camera-contribution correlations are negative, and all reported 95% confidence intervals exclude zero. From clean to the highest severity, R_{cam} decreases by 0.019–0.020 in absolute magnitude across the three dedicated corruption families. These results describe task-driven cues and magnitude-based statistical associations; they are not calibrated physical-quality measurements or causal claims.

5 Limitations and Future Work

The current study focuses on active camera degradation in camera–LiDAR BEV fusion. The controlled protocol makes it possible to isolate how visual degradation affects fusion behavior, but the overall evaluation also defines several boundaries of the present work.

- **Controlled camera-degradation setting.** The controlled degradation protocol keeps LiDAR and calibration parameters unchanged, which isolates the effect of camera degradation on fusion behavior. This design supports controlled mechanism analysis, but it does not cover coupled LiDAR–camera failures or cases in which LiDAR itself is degraded. Future work should evaluate joint sensor degradation and broader adverse-weather conditions.
- **Synthetic degradation coverage.** Synthetic night, motion blur, and Gaussian noise provide controllable severity levels, which is useful for analyzing severity-dependent behavior, but they are proxies rather than exhaustive models of real imaging degradation. The non-overlapping rain and night subsets offer complementary real-world evidence, yet they still cover a limited range of weather, illumination, and lens conditions. Future work should include more diverse real-world corruptions and sensor-specific failure modes.
- **Task-driven descriptors.** The DAC descriptors are derived from camera BEV feature response statistics, and their learnable mapping parameters are optimized with the original detection objective. This avoids image-quality supervision and keeps the module lightweight, but it also means that they are task-driven cues rather than calibrated physical-quality measurements. Future work could combine such cues with explicit image-quality or physical sensor models to improve interpretability.
- **Strong dense noise.** At the highest Gaussian noise level, the margin over BEVFusion becomes small and the two methods are comparable. This suggests that dense pixel-level corruption can mask useful visual evidence and motivates noise-aware or spatially adaptive fusion strategies.

These observations define the scope of the present study and motivate the future directions above.

6 Conclusion

This paper presented DARFusion, a lightweight Degradation-Aware and Reliability-Guided BEV fusion method for camera–LiDAR 3D object detection. DARFusion makes the camera sensing state visible to BEV fusion through modality reliability cues, task-driven auxiliary contexts, and compact camera degradation descriptors, without requiring degradation labels or auxiliary supervision. Experiments on nuScenes and KITTI show that this conditioning improves robustness in most degraded camera settings while preserving clean-set performance and adding little computational overhead. These results suggest that explicitly modeling camera reliability is a practical way to make BEV fusion less dependent on unreliable visual evidence.

Statements and Declarations

Data availability. No new datasets were generated in this study. The nuScenes and KITTI datasets analyzed in this work are publicly available from their official repositories at <https://www.nuscenes.org/nuscenes> and <https://www.cvlibs.net/datasets/kitti/>, respectively.

Code availability. The implementation and configuration files used in this study are available from the corresponding author on reasonable request.

Funding. The authors received no funding for this work.

Competing interests. The authors have no relevant financial or non-financial interests to disclose.

Author contributions. Weifeng Xu contributed to the conceptualization, methodology, software implementation, and review and editing of the manuscript. Yaqian Cui contributed to the experimental design, investigation, data curation, validation, experimental evaluation, data analysis, visualization, and preparation of the original manuscript. Hongtao Wang contributed to supervision, methodology, project administration, interpretation of the results, and review and editing of the manuscript. All authors read and approved the final manuscript.

References

- [1] Wang, H., Chen, X., Yuan, Q., Liu, P.: A review of 3d object detection based on autonomous driving. *The Visual Computer* **41**(3), 1757–1775 (2025) <https://doi.org/10.1007/s00371-024-03480-6>
- [2] Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12697–12705 (2019). <https://doi.org/10.1109/CVPR.2019.01298>

- [3] Yin, T., Zhou, X., Krähenbühl, P.: Center-based 3d object detection and tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11784–11793 (2021). <https://doi.org/10.1109/CVPR46437.2021.01161>
- [4] Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 1–18 (2022). https://doi.org/10.1007/978-3-031-20077-9_1
- [5] Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., Li, Z.: Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 1477–1485 (2023). <https://doi.org/10.1609/aaai.v37i2.25233>
- [6] Bai, X., Hu, Z., Zhu, X., Huang, Q., Chen, Y., Fu, H., Tai, C.-L.: Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1090–1099 (2022). <https://doi.org/10.1109/CVPR52688.2022.00116>
- [7] Yang, Z., Chen, J., Miao, Z., Li, W., Zhu, X., Zhang, L.: Deepinteraction: 3d object detection via modality interaction. In: Advances in Neural Information Processing Systems, vol. 35, pp. 1992–2005 (2022)
- [8] Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D.L., Han, S.: Bevfusion: Multi-task multi-sensor fusion with unified bird’s-eye view representation. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), pp. 2774–2781 (2023). <https://doi.org/10.1109/ICRA48891.2023.10160968>
- [9] Liang, T., Xie, H., Yu, K., Xia, Z., Lin, Z., Wang, Y., Tang, T., Wang, B., Tang, Z.: Bevfusion: A simple and robust lidar-camera fusion framework. In: Advances in Neural Information Processing Systems, vol. 35, pp. 10421–10434 (2022)
- [10] Zhang, N., Guerra, E., Grau, A.: Attbev: Enhancing multi-modal 3d object detection with cbam attention in bevfusion for autonomous driving. *IEEE Robotics and Automation Letters* **11**(2), 1322–1329 (2026) <https://doi.org/10.1109/LRA.2025.3641130>
- [11] Song, Y., Wang, L.: Bico-fusion: Bidirectional complementary lidar-camera fusion for semantic- and spatial-aware 3d object detection. *IEEE Robotics and Automation Letters* **10**(2), 1457–1464 (2025) <https://doi.org/10.1109/LRA.2024.3518845>
- [12] Dong, Y., Kang, C., Zhang, J., Zhu, Z., Wang, Y., Yang, X., Su, H., Wei, X., Zhu, J.: Benchmarking robustness of 3d object detection to common corruptions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

Recognition (CVPR), pp. 1022–1032 (2023)

- [13] Kong, L., Liu, Y., Li, X., Chen, R., Zhang, W., Ren, J., Pan, L., Chen, K., Liu, Z.: Robo3d: Towards robust and reliable 3d perception against corruptions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 19994–20006 (2023)
- [14] Xie, S., Kong, L., Zhang, W., Ren, J., Pan, L., Chen, K., Liu, Z.: Robobev: Towards robust bird’s eye view perception under corruptions. arXiv preprint arXiv:2304.06719 (2023) <https://doi.org/10.48550/arXiv.2304.06719>
- [15] Beemelmans, T., Zhang, Q., Geller, C., Eckstein, L.: Multicorrupt: A multi-modal robustness dataset and benchmark of lidar-camera fusion for 3d object detection. In: Proceedings of the IEEE Intelligent Vehicles Symposium (IV), pp. 3255–3261 (2024). <https://doi.org/10.1109/IV55156.2024.10588664>
- [16] Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11621–11631 (2020). <https://doi.org/10.1109/CVPR42600.2020.01164>
- [17] Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the KITTI vision benchmark suite. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3354–3361 (2012). <https://doi.org/10.1109/CVPR.2012.6248074>
- [18] Vora, S., Lang, A.H., Helou, B., Beijbom, O.: Pointpainting: Sequential fusion for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4604–4612 (2020). <https://doi.org/10.1109/CVPR42600.2020.00466>
- [19] Sindagi, V.A., Zhou, Y., Tuzel, O.: Mvx-net: Multimodal voxelnet for 3d object detection. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), pp. 7276–7282 (2019). <https://doi.org/10.1109/ICRA.2019.8794195>
- [20] Li, Y., Chen, Y., Qi, X., Li, Z., Sun, J., Jia, J.: Unifying voxel-based representation with transformer for 3d object detection. In: Advances in Neural Information Processing Systems, vol. 35, pp. 18442–18455 (2022)
- [21] Chen, W., Yin, S., Jiang, X., Ji, Z., Liu, H., Zhu, G., Armendariz-Inigo, J.E.: Rfpillarnext: Geometry-aware 3d object detection for dynamic lidar point clouds. *The Visual Computer* **41**, 12025–12036 (2025) <https://doi.org/10.1007/s00371-025-04142-x>
- [22] Yu, J., Wang, J., Zhu, W., Liu, H., Yang, W.: Crossbev: Enhancing multi-view

- 3d object detection via spatiotemporal feature cross-enhancement. *The Visual Computer* **42**, 225 (2026) <https://doi.org/10.1007/s00371-026-04431-z>
- [23] Du, H., Zhang, Z., Zhang, W., Cao, L.: Global–local feature fusion network for accurate 3d object detection in point clouds and images. *The Visual Computer* **42**, 12 (2026) <https://doi.org/10.1007/s00371-025-04220-0>
 - [24] Wozniak, M.K., Kårefjård, V., Thiel, M., Jensfelt, P.: Toward a robust sensor fusion step for 3d object detection on corrupted data. *IEEE Robotics and Automation Letters* **8**(11), 7018–7025 (2023) <https://doi.org/10.1109/LRA.2023.3313924>
 - [25] Ge, C., Chen, J., Xie, E., Wang, Z., Hong, L., Lu, H., Li, Z., Luo, P.: Metabev: Solving sensor failures for 3d detection and map segmentation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8721–8731 (2023)
 - [26] Wang, S., Caesar, H., Nan, L., Kooij, J.F.P.: Unibev: Multi-modal 3d object detection with uniform bev encoders for robustness against missing sensor modalities. In: *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, pp. 2776–2783 (2024). <https://doi.org/10.1109/IV55156.2024.10588783>
 - [27] Park, K., Kim, Y., Kim, D., Choi, J.W.: 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 (CVPR)*, pp. 6720–6729 (2025). <https://doi.org/10.1109/CVPR52734.2025.00630>
 - [28] Mohan, R., Drews, F., Miron, Y., Cattaneo, D., Valada, A.: Up-fuse: Uncertainty-guided lidar-camera fusion for 3d panoptic segmentation. *arXiv preprint arXiv:2602.19349* (2026) <https://doi.org/10.48550/arXiv.2602.19349>
 - [29] Yu, K., Tao, T., Xie, H., Lin, Z., Liang, T., Wang, B., Chen, P., Hao, D., Wang, Y., Liang, X.: Benchmarking the robustness of lidar-camera fusion for 3d object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3188–3198 (2023)
 - [30] Sural, S., Sahu, N., Rajkumar, R.: Contextualfusion: Context-based multi-sensor fusion for 3d object detection in adverse operating conditions. *arXiv preprint arXiv:2404.14780* (2024) <https://doi.org/10.48550/arXiv.2404.14780>
 - [31] Uricar, M., Sistu, G., Rashed, H., Vobecky, A., Ravi Kumar, V., Krizek, P., Burger, F., Yogamani, S.: Let’s get dirty: Gan based data augmentation for camera lens soiling detection in autonomous driving. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 766–775 (2021)
 - [32] Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. *Sensors* **18**(10), 3337 (2018) <https://doi.org/10.3390/s18103337>

- [33] Ju, C., Xie, Y., Duan, J., Chang, Y., Wu, M.: Keygeofusion: A multi-modal keypoint and geometry-aware framework for small and distant 3d object detection in sparse point clouds. *Neurocomputing* **672**, 132842 (2026) <https://doi.org/10.1016/j.neucom.2026.132842>