SCIENCE CHINA
Information Sciences

• Supplementary File •

Ro3DOP: robust 3D object perception via multi-modal feature reconstruction and adaptive fusion for autonomous vehicles

Yichi Zhang[1,2†], Haobing Pang[1,2†], Jianshan Zhou[1,2], Daxin Tian[1,2], Xuting Duan[1,2], Kaige Qu[1,2], Zhengguo Sheng[3], Dongpu Cao[4], Ruizhong Du[5] & Xiaoyun Guang[5]

[1]*State Key Lab of Intelligent Transportation System, Beihang University, Beijing 100191, China*
[2] *Beijing Key Laboratory for Cooperative Vehicle Infrastructure Systems and Safety Control, School of Transportation Science and Engineering, Beihang University, Beijing 100191, China*
[3]*Department of Engineering and Design, University of Sussex, BN1 9RH Brighton, U.K.*
[4]*School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China*
[5]*Hebei Province Key Laboratory of High Confidence Information System, School of Cyber Security and Computer, Hebei University, Baoding Hebei 071000, China*

Appendix A Literature review

Appendix A.1 3D object detection with single-modal sensors

3D object detection is a crucial task in autonomous driving perception [1]. Unlike 2D detection, 3D detection provides richer spatial information, which is important for understanding surrounding objects and the environment. Cameras are widely used in perception systems because they can provide dense and high-resolution semantic information while maintaining a low cost. To connect multi-perspective images with 3D spatial inference, Bird's-Eye-View (BEV) representation learning has gradually become a mainstream solution in this field and early studies mainly relied on dense projection strategies, while later works moved toward more powerful and more deployment-friendly frameworks, such as Fast-BEV [2], which established a solid baseline for real-time perception. More recently, the field has also converted to sparse detection paradigms, where 3D bounding boxes are predicted through learnable queries, as in BEVFormer [3] and SparseBEV [4]. Although camera-based models are attractive in terms of semantic richness and computational efficiency, they still suffer from an inherent geometric limitation: monocular images do not directly provide depth information, so 3D inference often depends on depth estimation from the visual texture. However, this step is unstable in real driving environments such as under adverse lighting, intense glare, or sensor noise. Once the camera vision is degraded by occlusion, bandwidth constraints, or other failures, the visual semantic information may no longer be anchored reliably in space and the detection performance of vision-only systems can drop sharply as a result, which is especially problematic in safety-oriented autonomous driving scenarios.

In contrast, LiDAR provides direct geometric measurements of the environment through point clouds and offers stronger structural information. In order to handle the sparsity and irregularity of raw points, existing methods develops along two main lines, including voxel-based discretization [5, 6] and sparse point processing [7], which have enabled accurate localization and structural inference in many real driving scenes. Recent methods such as FocalFormer3D [8] further improve recall in complex and occluded cases by using systematic hard-sample mining. However, LiDAR-based perception also has its limitations: although the geometric accuracy is higher, the semantic information carried by sparse point returns is often insufficient, especially for distant targets or small objects. As a result, LiDAR alone may not provide enough category information for reliable recognition, and external disturbances such as heavy rain or snow, as well as internal hardware failures, can damage the geometric integrity of the point cloud itself. Without complementary semantic information or a stronger fusion mechanism, LiDAR-only systems also suffer significant performance degradation in real driving environments, particularly when needing to recognize distant or small-scale objects.

Appendix A.2 Multi-modal fusion: from precision to resilience

By combining the complementary strengths of LiDAR and cameras, multi-modal fusion has become one of the main approaches for high-precision 3D perception. Representative methods such as BEVFusion [9] and TransFusion [10] report state-of-the-art results on standard benchmark datasets. However, most of these accuracy-oriented frameworks are built on a strong assumption: the sensor inputs are reliable and the calibration remains valid, however, such assumption is often difficult to satisfy in real-world deployments, where hardware corruption, data loss, and other anomalous sensor conditions are not uncommon. Once one modality is degraded, the corrupted features may still continue to pass through the fusion pipeline and disturb the shared representation. Because most existing fusion strategies do not explicitly model feature corruption or isolate its effect, degraded visual features can interfere with the geometric information from LiDAR and impair the reliability of the overall perception process and vice versa.

* Corresponding author (email: jianshanzhou@foxmail.com)
† Yichi Zhang and Haobing Pang contributed equally to this work.

Recent studies have started to pay more attention to robustness under missing or abnormal modalities. Methods such as MetaBEV [11] and MoME [12] reduce sensor dependency through modality-arbitrary training or Mixture-of-Experts designs, while ProFusion3D [13] uses a progressive fusion strategy to reweight inputs according to their estimated quality. Although these methods are useful when an entire modality is missing, their basic treatment is still largely bypass-oriented: once a sensor is regarded as corrupted, its features mainly treated as useless and excluded from further combination pipeline. This choice is not always ideal, especially for small objects such as pedestrians and bicycles. In such cases, the pipeline degrades as LiDAR-only, and the point cloud returns are often too sparse to support reliable classification on its own, and removing the visual branch may also remove semantic information that is still partially useful. This point suggests that, under severe degradation, the problem is not only how to avoid corrupted inputs, but also how to recover useful structure from them. Ro3DOP is motivated by this observation. Instead of simply bypassing degraded inputs, it attempts to reconstruct corrupted feature tensors by exploiting intrinsic cross-modal correlations. In this way, the method restores the structural completeness of the high-dimensional feature manifold and provides more reliable semantic support for adaptive fusion under severe sensor degradation.

Appendix B High dimensional feature reconstruction

Incomplete feature tensors caused by sensor failure are addressed here through a high-dimensional feature tensor reconstruction method, named High Dimensional Feature Reconstruction (HD-FR). Instead of treating feature loss caused by sensors as simple pixel-wise missing data, we regard it as manifold damage in the high-dimensional feature space.By utilizing the intrinsic relationship among multi-modal feature streams, HD-FR aims to recover the structural integrity of the feature manifold while preserving local details. Firstly, we let $\mathcal{X} \in \mathbb{R}^{C\times W\times H}$ denote the corrupted feature tensor that encodes multi-dimensional structural and semantic features and $\mathcal{Y} \in \mathbb{R}^{C\times W\times H}$ represent its corresponding reconstructed part, where C, W, and H represent the number of channels, width, and height, respectively. Then we decompose the high-dimensional tensor into multiple low-dimensional factors, and introduce a mask tensor to indicate the missing regions of the corrupted tensor via the Hadamard product, which guides feature reconstruction while keeping the known valid data unchanged. The reconstruction problem is formulated by minimizing the discrepancy between the reconstructed tensor $\mathcal{Y}$ and its low-rank decomposition, while enforcing consistency with reliable observations:

$$\min_{\mathcal{Y},\mathcal{U}} \quad \frac{1}{2}\|\mathcal{Y} - \mathrm{TD}(\mathcal{U}_1,\mathcal{U}_2,\mathcal{U}_3)\|_F^2 + \mathcal{L}_s(\mathcal{Y}), \tag{B1}$$

where $\mathrm{TD}(\mathcal{U}_1,\mathcal{U}_2,\mathcal{U}_3)$ denotes the tensor decomposition operator parameterized by the latent factors $\mathcal{U}_1$,$\mathcal{U}_2$, and $\mathcal{U}_3$, and $\mathcal{L}_s(\mathcal{Y})$ enforces consistency between the reconstructed tensor and the reliable observations, as detailed in Eq. (B3). Building on this setting, decomposition ranks R_{n_1,n_2} are introduced to describe the shared latent dimensions between the n_1-th and n_2-th factors which determine the sizes of the contracted dimensions in tensor decomposition and affect the balance between reconstruction fidelity and computational cost. The decomposition ranks are selected according to the intrinsic dimension of the feature manifold, typically by preserving more than 95% of the energy in the singular value spectrum of the latent features, which helps maintain a trade-off between reconstruction fidelity and computational overhead. In addition, we denote the set of pairwise decomposition ranks by $R = \{R_{1,2}, R_{1,3}, R_{2,3}\}$ for simplicity. The element-wise form of the decomposition can be formulated as:

$$\begin{aligned}\mathcal{X}(i_1,i_2,i_3) = \sum_{r_{1,2}=1}^{R_{1,2}} \sum_{r_{1,3}=1}^{R_{1,3}} \sum_{r_{2,3}=1}^{R_{2,3}} & \mathcal{U}_1(i_1,r_{1,2},r_{1,3}) \\ & \cdot \mathcal{U}_2(r_{1,2},i_2,r_{2,3}) \cdot \mathcal{U}_3(r_{1,3},r_{2,3},i_3),\end{aligned} \tag{B2}$$

which defines a latent feature manifold, where each factor $\mathcal{U}_k$ represents the elements located at position (i_1, i_2, i_3) within the decomposed tensor, capturing structural dependencies along one mode and interacting with the other modes through shared ranks. Such factorization also provides a memory advantage: by representing the dense feature tensor with low-rank components, the total number of parameters is reduced from exponential growth to linear growth with respect to the dimensions, enabling efficient deployment on memory-constrained autonomous driving hardware while maintaining semantic consistency. The decomposition can be denoted by $\mathcal{X} = \mathrm{TD}(\mathcal{U}_1,\mathcal{U}_2,\mathcal{U}_3)$.

To ensure the reconstruction process goes beyond pixel-wise completion while also preserving global consistency and contextual dependencies, we introduce a regularization term $\mathcal{L}_s(\mathcal{Y})$, which consists of a mask tensor $\mathcal{M}$ with the same dimensions as the corrupted feature tensor and an indicator function $\iota_S(\cdot)$. Specifically, the mask tensor $\mathcal{M}$ is used to distinguish reliable observations from corrupted parts, while the indicator function $\iota_S(\cdot)$ constrains the recovery process to preserve not only local observations but also the global consistency, contextual dependencies, and semantic coherence of the feature streams. Although $\mathcal{M}$ is used to isolate corrupted regions in simulation, in real-world deployment it can be generated by an onboard self-diagnosis mechanism. For instance, the system can estimate $\mathcal{M}$ by monitoring sensor hardware error codes or analyzing anomalous feature variance caused by lens occlusion or strong glare, thereby identifying invalid regions during inference. The regularization term $\mathcal{L}_s(\mathcal{Y})$ is formulated as

$$\mathcal{L}_s(\mathcal{Y}) = \lambda_1 \left(\frac{1}{2}\|(\mathcal{Y} - \mathcal{X}) \odot \mathcal{M}\|_F^2\right) + \lambda_2 \iota_S(\mathcal{Y}), \tag{B3}$$

where

$$\mathcal{M} = \begin{cases} 0, & \mathcal{X}_t \in \Omega, \\ 1, & \mathcal{X}_t \notin \Omega, \end{cases} \qquad \iota_S(\mathcal{Y}) = \begin{cases} \infty, & \text{if } \mathcal{Y} \notin S, \\ 0, & \text{if } \mathcal{Y} \in S, \end{cases}$$

with

$$S := \{\mathcal{Y} : P_{\Omega^c}(\mathcal{Y} - \mathcal{X}) = 0\}.$$

Here, Ω denotes the index set of corrupted elements, Ω^c denotes its complement corresponding to reliable observations, $\odot$ denotes the Hadamard product, and $P_{\Omega^c}(\cdot)$ denotes the projection operator associated with the reliable entries, while S represents the constraint set defined by the consistency condition. The mask tensor $\mathcal{M}$ ensures that the restoration process focuses on corrupted entries while keeping the reliable observations unchanged. Accordingly, the masked difference term $(\mathcal{Y}-\mathcal{X})\odot\mathcal{M}$ enforces data consistency only on the available observations, while allowing the corrupted entries to be inferred from the latent high-dimensional feature space. During optimization, $\mathcal{M}$ guides the learning process to update only the corrupted entries, while the observed entries act as hard constraints that anchor the reconstruction. In this way, the recovery problem is reformulated as the reconstruction of the multi-modal feature manifold encoded in $\mathcal{X}$. Meanwhile, the indicator function $\iota_S(\cdot)$ further enforces consistency between the reconstructed tensor $\mathcal{Y}$ and the available observations at known locations, while allowing the missing entries to be inferred from the global structure of the feature manifold.

Since the optimization objective jointly depends on the reconstructed tensor $\mathcal{Y}$ and the factor tensors $\mathcal{U}=\{\mathcal{U}_1,\mathcal{U}_2,\mathcal{U}_3\}$, we adopt an alternating optimizing strategy. By updating one group of variables while fixing the others, the optimization problem is decomposed into a sequence of sub-problems, and the solution is iteratively refined toward a stationary point of the objective function $f(\mathcal{Y},\mathcal{U})$ defined in Equation (B1). While fixing the remaining latent factors, the update process can be formulated as follows:

$$\begin{cases}\mathcal{U}_k^{(t+1)}=\arg\min\limits_{\mathcal{U}_k}\left\{f\left(\mathcal{U}_k,\mathcal{Y}^{(t)}\right)+\dfrac{\lambda}{2}\left\|\mathcal{U}_k-\mathcal{U}_k^{(t)}\right\|_F^2\right\},\\ \mathcal{Y}^{(t+1)}=\arg\min\limits_{\mathcal{Y}}\left\{f\left(\mathcal{U}^{(t+1)},\mathcal{Y}\right)+\dfrac{\lambda}{2}\left\|\mathcal{Y}-\mathcal{Y}^{(t)}\right\|_F^2\right\},\end{cases}\tag{B4}$$

where $\lambda>0$ is a proximal parameter that controls the update magnitude, thereby improving numerical stability and promoting convergence to a semantically consistent high-dimensional representation.

Factor update: To update the latent factors, we first introduce the mode-k matricization of the reconstructed tensor:

$$\mathcal{Y}_{\text{mode-}k}=\text{reshape}\left(\mathcal{Y},\prod_{i\neq k}I_i,I_k\right),\tag{B5}$$

where I_i denotes the size of $\mathcal{Y}$ along its i-th mode. Equivalently, this operation can be written as $\mathcal{Y}_{\text{mode-}k}=\text{unfold}(\mathcal{Y},k)$, and its inverse operation is given by $\mathcal{Y}=\text{fold}(\mathcal{Y}_{\text{mode-}k},k)$. Based on this matricized representation, the update of the latent factor $\mathcal{U}_k$ can be derived as

$$\begin{aligned}\mathcal{U}_k^{(t+1)}&=\text{fold}\left((\mathcal{U}_k^{(t+1)})_{\text{mode-}k}\right)\\&=\text{fold}\left(\left(\mathcal{Y}^{(t)}_{\text{mode-}k}(W_k^{(t)})_{\text{mode-}(m_1,m_2)}+\lambda(\mathcal{U}_k^{(t)})_{\text{mode-}k}\right)\left[\left((W_k^{(t)})_{\text{mode-}(m_1,m_2)}\right)^2+\lambda\mathbf{E}\right]^{-1}\right).\end{aligned}\tag{B6}$$

Here, $W_k^{(t)}$ denotes the contracted tensor associated with the update of $\mathcal{U}_k$, (m_1,m_2) denotes the corresponding contracted modes, and $\mathbf{E}$ is the identity matrix of compatible size. The matrix inversion in Eq. (B6) is performed on matrices whose sizes are determined by the decomposition ranks. Since these ranks are much smaller than the original feature dimensions, the computational cost of each iteration remains manageable and is compatible with real-time deployment on autonomous platforms.

Tensor update: With $\mathcal{U}_k$ fixed, the reconstructed tensor $\mathcal{Y}$ is updated by reassembling the latent factors while enforcing projection consistency:

$$\mathcal{Y}^{(t+1)}=\frac{1}{\lambda}\mathcal{I}_\Omega\left(\text{TD}(\mathcal{U}_1,\mathcal{U}_2,\mathcal{U}_3)\right)+\overline{\mathcal{I}}_\Omega(\mathcal{X}),\tag{B7}$$

where $\mathcal{I}_\Omega(\cdot)$ denotes the projection operator on the corrupted set Ω, while $\overline{\mathcal{I}}_\Omega(\cdot)$ denotes its complementary projection on the reliable entries. This update fills the missing entries with values inferred from the latent decomposition, while the observed entries remain fixed to preserve fidelity to the available data. In Eq. (B7), the first term reconstructs the latent tensor from the updated factors and projects it onto the corrupted region, whereas the second term preserves the reliable observations from $\mathcal{X}$. The overall computational complexity of Algorithm B1 is $O(N)$ per iteration, where N denotes the number of feature entries. In practice, the objective variation $\Delta f=\left|f^{(t+1)}-f^{(t)}\right|$ typically decreases below $\epsilon=10^{-6}$ within only a few iterations, indicating that the recovered tensor achieves both structural stability and computational efficiency for downstream tasks.

After each iteration, we evaluate the change in the objective function:

$$\Delta f=\left|f^{(t+1)}(\mathcal{Y},\mathcal{U})-f^{(t)}(\mathcal{Y},\mathcal{U})\right|,\tag{B8}$$

and terminate the optimization when $\Delta f<\epsilon$, where ϵ is a sufficiently small threshold. Here, $f(\cdot)$ denotes the objective function defined in Eq. (B1). This stopping criterion ensures that the recovered tensor reaches a stable and semantically consistent solution. After convergence, the reconstructed tensor $\mathcal{Y}$, which represents the recovered high-dimensional feature manifold, is mapped back to the original domain for downstream tasks or visualization.

Appendix C Refinement of dual-attention fusion

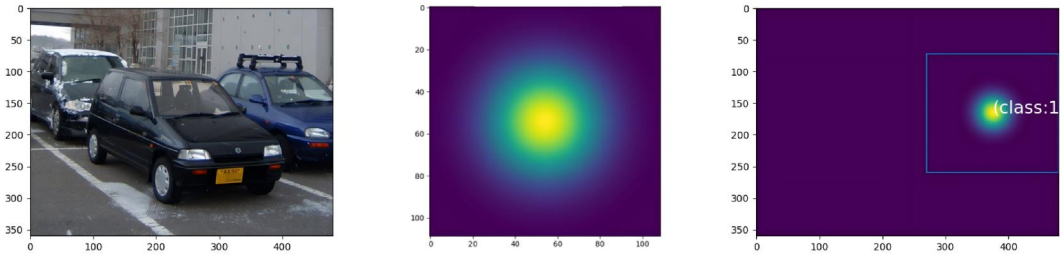
In the previous section, we addressed corrupted sensor observations through feature reconstruction, providing more reliable feature representations for our multi-modal fusion pipeline. On this basis, we introduce Ro3DOP as a robust dual-attention-based multi-modal

Algorithm B1 High-Dimensional Feature Reconstruction

Input: Corrupted sensor tensor $\mathcal{X}$, mask tensor $\mathcal{M}$, tensor decomposition rank collection R, proximal parameter λ, maximum iteration number $t_{\max}$, convergence threshold ϵ

Output: Reconstructed sensor tensor $\mathcal{Y}$

Initialize $\mathcal{Y}^{(0)} = \mathcal{X}$, initialize the latent factors $\{\mathcal{U}_k^{(0)}\}$, and set $t = 0$.
repeat
  Update the factor tensors $\mathcal{U}_k^{(t+1)}$ according to Eq. (B6).
  Update the reconstructed tensor $\mathcal{Y}^{(t+1)}$ according to Eq. (B7).
  Compute $\Delta f = \left| f^{(t+1)}(\mathcal{Y}, \mathcal{U}) - f^{(t)}(\mathcal{Y}, \mathcal{U}) \right|$.
  Set $t = t + 1$.
until $\Delta f < \epsilon$ or $t \geqslant t_{\max}$
return $\mathcal{Y}$.



(a) Original Image (b) Gaussian Kernel (c) Heatmap

Figure C1 Visualization of the Gaussian heatmap generation process for target localization. (a) Original image with the ground-truth annotation for a vehicle category. (b) Gaussian kernel, depicting a two-dimensional Gaussian function used to determine the spatial influence and radius around the object center. (c) Generated heatmap, illustrating the result of mapping the Gaussian kernel onto the target's spatial coordinates.

fusion perception framework. Motivated by early studies on dual-stage detection architectures, Ro3DOP employs dual decoders: the first generates initial predictions guided by LiDAR features, while the second refines them through attention-based fusion with the reconstructed sensor features. This dual-stage design improves both perception accuracy and robustness, even when either modality is partially degraded.

Appendix C.1 LiDAR-based detection

In this stage, Ro3DOP adopts an initial detection architecture dominated by LiDAR point clouds. The points are first projected into Bird's-Eye-View (BEV) space. To ensure computational efficiency and minimize inference latency, we omit the Transformer encoder following the 3D backbone. Instead, the subsequent cross-modal fusion stage is used to further refine local features. Inspired by query-based detection, we introduce input-aware object queries that interact with LiDAR BEV features via a cross-attention mechanism to aggregate task-relevant geometry. Unlike traditional input-agnostic queries, our queries are initialized directly from the BEV feature map, which significantly accelerates convergence and improves localization precision by providing high-fidelity spatial priors.

The initialization of object queries is achieved through heat maps generated from the LiDAR BEV feature map. A convolution operation is applied to reduce the channel dimension of the BEV feature map so that it matches the number of object categories. Local maxima in the resulting heat map are identified through pooling, and the top-N hotspots, representing potential object centers, are selected as query anchors. Each hotspot encodes spatial, semantic, and categorical cues, ensuring that the initialized queries are placed near likely object locations. For illustration, an example of the generated heat map, sourced from the VOC dataset [14], is shown in Fig. C1, which visualizes the object category and location information.

Let $\mathbf{p} = (x, y)$ represent a point on the image, and $\mathbf{c} = (x_c, y_c)$ denote the coordinates of the target center. The value of the two-dimensional Gaussian function $G_{\mathbf{c},\sigma}$ at point $\mathbf{p}$ is defined by the following equation [15, 16]:

$$G_{\mathbf{c},\sigma}(\mathbf{p}) = \exp\left(-\frac{(x - x_c)^2 + (y - y_c)^2}{2\sigma^2}\right), \tag{C1}$$

where σ is the standard deviation of the Gaussian kernel, controlling the degree of smoothing.

To generate the heat map, we place a Gaussian kernel at the centroid coordinates of each target in the image. We then compute the maximum value of these Gaussian kernels across all targets to obtain the final heat map H:

$$H(\mathbf{p}) = \max_{\mathbf{c} \in \mathbb{C}} G_{\mathbf{c},\sigma}(\mathbf{p}), \tag{C2}$$

where $\mathbb{C}$ is the set of all target center points in the image. Since the heatmap $H(\mathbf{p})$ represents the likelihood of each position being the centroid of a target, with larger values indicating a higher probability of the position being the center of an object, we select the positions corresponding to the top N maximum values from the heatmap as the initial query anchor points Q_{obj}:

$$Q_{\text{obj}} = \text{Top-N}(H), \tag{C3}$$

which represent potential object centers initialized close to the true object positions. They not only reflect the spatial locations of the objects but also incorporate contextual information from the image.

Appendix C.2 Cross-modal feature fusion

The objective of this stage is to achieve an adaptive alignment between the semantically recovered multi-view image features and the geometrically precise LiDAR BEV anchors. To maintain a streamlined architecture and minimize memory overhead, we refrain from employing a Transformer encoder in this stage and directly perform cross-modal interaction on the raw feature streams. First, the HD-FR module reconstructs semantically consistent features from corrupted inputs, which are then processed by a convolutional backbone to generate multi-view keys (K_{img}) and values (V_{img}). Meanwhile, the input-aware object queries (Q_{obj}), which encode the spatial priors initialized in the first stage (Eq. C3), are utilized to probe the image features. By leveraging these sparse, high-confidence anchors, the multi-head cross-attention (MHCA) mechanism effectively constrains the interaction within task-relevant regions, drastically reducing the computational complexity compared to dense global attention paradigms. The fusion process is formulated as:

$$\begin{aligned} \text{MHCA}(Q,K,V) &= \text{Concat}(\text{head}_1,\ldots,\text{head}_H)W_0, \\ \text{head}_k &= \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \\ F_{\text{BEV}} &= \text{MHCA}(Q_{\text{obj}}, K_{\text{img}}, V_{\text{img}}), \end{aligned} \tag{C4}$$

where F_{BEV} denotes the fused representation. This design ensures that the system maintains the geometric precision of the LiDAR stream while benefiting from the semantic enrichment of the camera modality, resulting in a robust perception manifold capable of handling severe modality degradation.

Appendix C.3 LiDAR-guided attention mechanism (LGA)

Although MHCA supports cross-modal fusion, the queries may still attend to the irrelevant regions instead of the informative regions when objects are sparse or the images are noisy, which reduces training efficiency and slows convergence [17]. To mitigate the sensitivity to sensor calibration misalignment and the propagation of visual noise in "hard-association" strategies, we introduce the LiDAR-Guided Attention (LGA) mechanism, which acts as a soft spatial constraint that uses the 3D bounding box predictions from the first stage as priors to guide the attention distribution and emphasize task-relevant regions. Taking a scene containing n objects as an example, we define a spatial weighting function W using Gaussian kernels centered at the predicted centroids $\mathbf{c}_i$:

$$\mathbf{W}_i = \exp\left(-\frac{\|\mathbf{p}-\mathbf{c}_i\|_2^2}{2\sigma^2}\right), \quad i \in \{1,2,\cdots,n\}, \tag{C5}$$

where $\mathbf{p}$ denotes the pixel coordinates on the feature map, and σ controls the spatial tolerance range. What's more, the resulting weight modifies the standard attention mapping (AM) as:

$$\begin{aligned} \text{AM} &= \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right), \\ \text{AM}' &= \mathbf{W}\cdot\text{AM}, \quad \mathbf{W} = (\mathbf{W}_1,\mathbf{W}_2,\cdots,\mathbf{W}_n), \end{aligned} \tag{C6}$$

where the modulation is applied element-wise over the attention map. As a result, the attention distribution is guided not only by features but also by explicit 3D geometric priors. By relaxing rigid point-to-pixel projection to a spatial probability distribution, the framework becomes more tolerant to minor extrinsic calibration errors. The object queries Q_{obj} are then updated via the modulated attention map AM' to capture finer semantic details:

$$Q_{\text{obj}}^{(s+1)} = \text{softmax}\left(\text{AM}'\cdot Q_{\text{obj}}^{(s)}\right), \tag{C7}$$

where $Q_{\text{obj}}^{(s+1)}$ and $Q_{\text{obj}}^{(s)}$ denote the updated and current queries, respectively. Finally, the fused BEV features refined by the LGA module are passed to the MLP to produce refined 3D detection results, as shown in Eq. C8:

$$\hat{b} = \text{MLP}(F'_{\text{BEV}}), \tag{C8}$$

where $\hat{b}$ denotes the predicted 3D bounding box, and F'_{BEV} represents the fused BEV feature map obtained after refining the query anchors Q_{obj} through Eqs. C4 to C7. The refined 3D bounding boxes contain the object's location, size, and orientation in 3D space, which can benefit downstream tasks and help the final predictions maintain high geometric accuracy and prediction robustness by integrating spatial and semantic information in the feature maps, even under severe sensor degradation. For the joint optimization of

Algorithm C1 LiDAR-camera fusion for perception

Input: Point cloud $\mathcal{P}$, restored image features $\mathcal{Y}$, corresponding labels and ground-truth bounding boxes, learning rate lr, batch size B, number of epochs I.
Output: Refined 3D bounding box predictions $\hat{b}$.

Initialize object queries Q_{obj} from the heatmap generated on the LiDAR BEV feature map F_L;
Construct the forward process F, including the voxel encoder, convolution layers, the MHCA-based fusion module, and the LGA module;
Define the loss function L according to Eq. (C9) for classification and bounding box matching;
Set the optimizer O with the adopted training configuration (e.g., AdamW with learning rate lr);
repeat
  Feed point cloud $\mathcal{P}$ and restored image features $\mathcal{Y}$ into the forward process F to obtain feature maps from both modalities;
  Fuse the image features and LiDAR BEV features through MHCA to produce the fused BEV feature map F_{BEV};
  Update the object queries using the modulated attention map together with LiDAR-derived spatial priors;
  Refine the fused features with the LiDAR-Guided Attention (LGA) module and pass them through the MLP to obtain refined bounding box predictions;
  Compute the loss L and backpropagate the gradients;
  Update the network parameters with optimizer O.
until the number of epochs I is reached
Return the final refined 3D bounding box predictions $\hat{b}$.

semantic discrimination and geometric consistency, we employ a composite loss function during training. The total objective function L minimizes the discrepancy between the predicted 3D bounding box $\hat{b}$ and the ground-truth b, and is formulated as:

$$\begin{aligned} L &= \sigma_1 L_{cls}(y_i, p(y_i)) + \sigma_2 L_{match}(b, \hat{b}) \\ &= -\frac{\sigma_1}{N}\sum_{i=1}^{N}\left[y_i \cdot \log(p(y_i)) + (1-y_i)\cdot\log(1-p(y_i))\right] \\ &\quad + \sigma_2\left[\alpha \cdot L_{\text{IoU}}(b,\hat{b}) + (1-\alpha)\cdot L_{\text{L2}}(b,\hat{b})\right], \end{aligned} \tag{C9}$$

where L_{cls} denotes the binary cross-entropy loss for objects and background classification, and L_{match} represents the hybrid matching loss for bounding box regression. The localization term L_{match} combines the Intersection over Union (IoU) loss L_{IoU} and the L2 distance loss L_{L2}. Here, the hyperparameter α controls the balance between structural overlap and coordinate-wise precision. Using only L2 loss may cause the model to overfit to the noisy local centroids under severe sensor degradation (e.g., 90% noise), so we introduce L_{IoU} as a global geometric constraint which helps the model obtain more stable localization and more robust perception results under the degradation. Futhermore, the scaling factors σ_1 and σ_2 are used to balance the classification and regression gradients, accelerating model convergence.

Appendix D Supplementary experimental analysis

Appendix D.1 Dataset

NuScenes-ArtNoise. All final evaluations of Ro3DOP are performed on the nuScenes dataset [18], while the nuScenes-ArtNoise protocol is used to simulate sensor degradation observed in real-world autonomous driving scenarios. In our setting, random white and light-gray pixel distributions are injected to imitate unstructured feature-level corruption, which can be caused by CMOS hardware failures, packet loss during data transmission, and severe lens occlusions. The framework is tested under two extreme noise levels, 60% and 90%, to establish a rigorous "stress test" setup, which allows us to evaluate whether the model can preserve the intrinsic manifold structure of the feature space when local semantic information is heavily degraded.

Appendix D.2 Implementation details

All experiments are conducted on an NVIDIA RTX A6000 GPU (48G), which provides sufficient computational capacity for training high-dimensional tensor operations and the multi-modal perception fusion pipeline. The framework is implemented based on the modular design of MMDetection3D [19] and uses CenterPoint-Voxel as the primary 3D backbone, enabling our HD-FR and LGA modules to be intergrated into standard 3D perception pipelines. For a fair comparison with existing state-of-the-art (SOTA) methods, all backbone hyperparameters, including voxel size, receptive field, and anchor configurations, are kept consistent with the official implementation of the nuScenes benchmark. Object queries Q_{obj} are initialized through the Gaussian heatmap generation mechanism described in Appendix C.1, together with category-specific convolutions and a 3×3 max-pooling operation. Ro3DOP is optimized with the AdamW optimizer and a one-cycle learning rate policy for efficient and stable convergence. A two-stage training strategy is adopted to ensure geometric stability, and a constant batch size of 16 is used throughout the optimization process: training process begins with 20 epochs of pre-training for the 3D backbone to enhance structural feature extraction, followed by 6 epochs of joint LiDAR-image fusion training

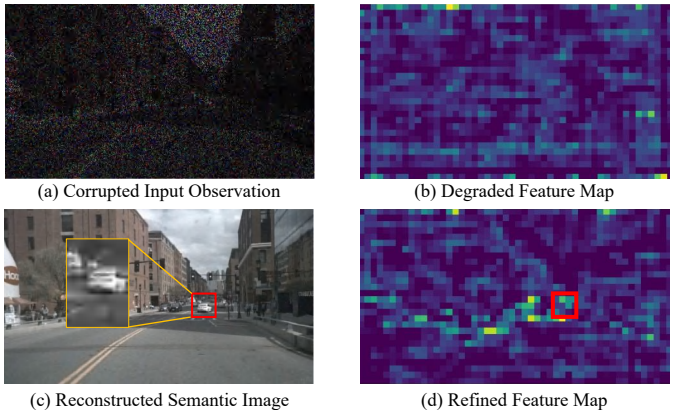


Figure D1 Qualitative analysis of the feature reconstruction process. (a) Corrupted observation exhibiting severe noise and irregular smearing. (b) Degraded feature map with disordered activation patterns. (c) Reconstructed image restored via the HD-FR module, showing semantic recovery. (d) Refined feature map, where red bounding boxes highlight the restoration of fine-grained local details and stabilized feature manifolds.

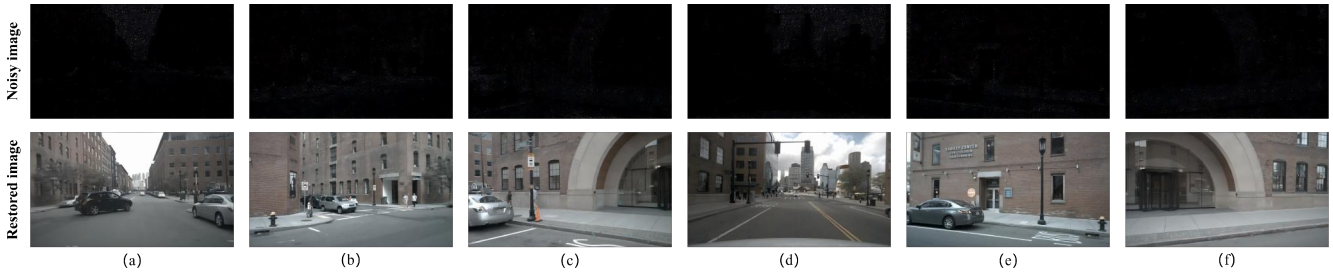


Figure D2 Results of feature reconstruction under 90% noise. Brackets one to six denote six different camera views in the same scene.

to refine the cross-modal attention mechanisms. To improve robustness and reduce overfitting, several data augmentation operations are applied, including random flipping, global rotation, and spatial scaling. Meanwhile, gradient clipping is also applied to mitigate numerical instabilities caused by the iterative reconstruction process within the HD-FR module, ensuring the optimization stable during training.

Appendix D.3 Feature reconstruction results

Appendix D.3.1 *Qualitative analysis.*

To qualitative analyze the restoration effect, Fig. D1 presents a detailed case study of image feature reconstruction. While the raw input (a) and damaged feature map (b) exhibit unstructured distortion due to extreme noise, our module successfully restores the overall semantic content (c). As highlighted by the red bounding boxes in the reconstructed feature map (d), HD-FR effectively captures small and distant local details, providing a stabilized geometric foundation for subsequent fusion stages. Extending this analysis to a broader scale, Fig. D2 illustrates the reconstruction results across multiple scenes.

Appendix D.3.2 *Quantitative and efficiency analysis.*

To quantitatively evaluate the effectiveness of our feature reconstruction module, we employ the Peak Signal-to-Noise Ratio (PSNR), a standard metric for assessing the quality of reconstructed images. PSNR is defined as a normalized logarithmic function of the Mean Squared Error (MSE) between the original sensor observation I and the processed K, while higher PSNR values indicate that the reconstructed feature is closer to the original observation.

$$\mathrm{MSE} = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2, \tag{D1}$$

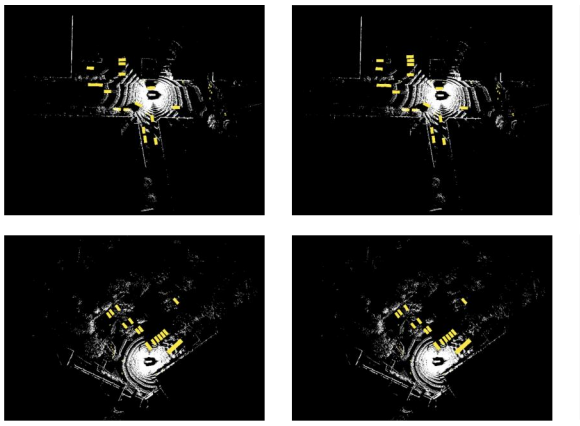
$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}_I^2}{\mathrm{MSE}}\right), \tag{D2}$$

where m and n are the input tensor dimensions, and MAX_I is the maximum possible pixel value. A higher PSNR value indicates lower distortion and better reconstruction quality. The corresponding PSNR values are summarized in Table D1. Our HD-FR module achieves robust reconstruction performance across all camera perspectives even under an extreme noise intensity of 90%, while reducing the peak memory usage by approximately 45% compared with dense feature processing by operating in a low-rank latent space.

Table D1 Results on the feature reconstruction module under 90% noise

	(1)[a]	(2)	(3)	(4)	(5)	(6)
PSNR	34.281	32.753	35.006	34.245	33.733	36.931

[a]The different views at the same time stamp in Fig. D2.



(a) Ours w/o (b) Ours

Figure D3 Qualitative comparison of perception results under noisy conditions. (a) Perception results without the image completion module. (b) Perception results with the proposed robust perception framework.

Table D2 Ablation study of the HD-FR module under different noise intensities. AP (%) results are reported, with performance gains (↑) attributed to the reconstruction module shown in brackets.

Noise Level	HD-FR Module	Cars	Pedestrians	Bicycle
60% Noise	w/o HD-FR	90.9 (-0.2)	94.1 (-0.2)	54.6 (-10.4)
	Full Ro3DOP	**91.1**	**94.3**	**65.0**
90% Noise	w/o HD-FR	90.7 (-0.4)	93.9 (-0.4)	53.2 (-11.8)
	Full Ro3DOP	**91.1**	**94.3**	**65.0**

Values in brackets denote the performance drop relative to the full Ro3DOP framework.

Appendix D.4 Ablation study of feature reconstruction

To quantitatively assess the contribution of the HD-FR module, we conduct comprehensive ablation experiments on the full nuScenes validation set. The primary evaluation metric is Average Precision (AP), which provides a holistic measure of detection performance by integrating Precision and Recall:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{D3}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively. The AP is derived by computing the area under the Precision-Recall curve, representing the model's mean detection capability across varying confidence thresholds.

The ablation results are summarized in Table D2, which reveal the important role of the HD-FR module in maintaining perception robustness under extreme modality degradation. A key observation is the variation in modality dependency across different object classes. For larger objects such as Cars and Pedestrians, the perception accuracy remains relatively stable even without feature reconstruction (within 0.4% AP), which is attributed to the abundant geometric information provided by dense LiDAR point cloud data, partially compensating for the loss of visual information. In contrast, small-scale categories such as Bicycles exhibit high sensitivity to sensor noise due to their sparse LiDAR reflections: removing the HD-FR module leads to an 11.8-point AP drop for bicycles under 90% noise intensity. This performance gap suggests that while LiDAR provides sufficient structural cues for large objects, small objects rely more heavily on the camera modality for semantic perception. By reconstructing the latent feature manifold, our module recovers semantic cues that would otherwise be corrupted by sensor failure. Meanwhile, this structural reconstruction is further supported by the qualitative results in Fig. D3. While the baseline model in (a) suffers from significant localization failures and false negatives under noisy conditions, the full Ro3DOP framework in (b) accurately identifies and localizes the objects. The widening performance gap as noise increases from 60% to 90% indicates that the HD-FR module becomes increasingly important in extreme "stress test" scenarios, stabilizing semantic perception for the multi-modal fusion pipeline.

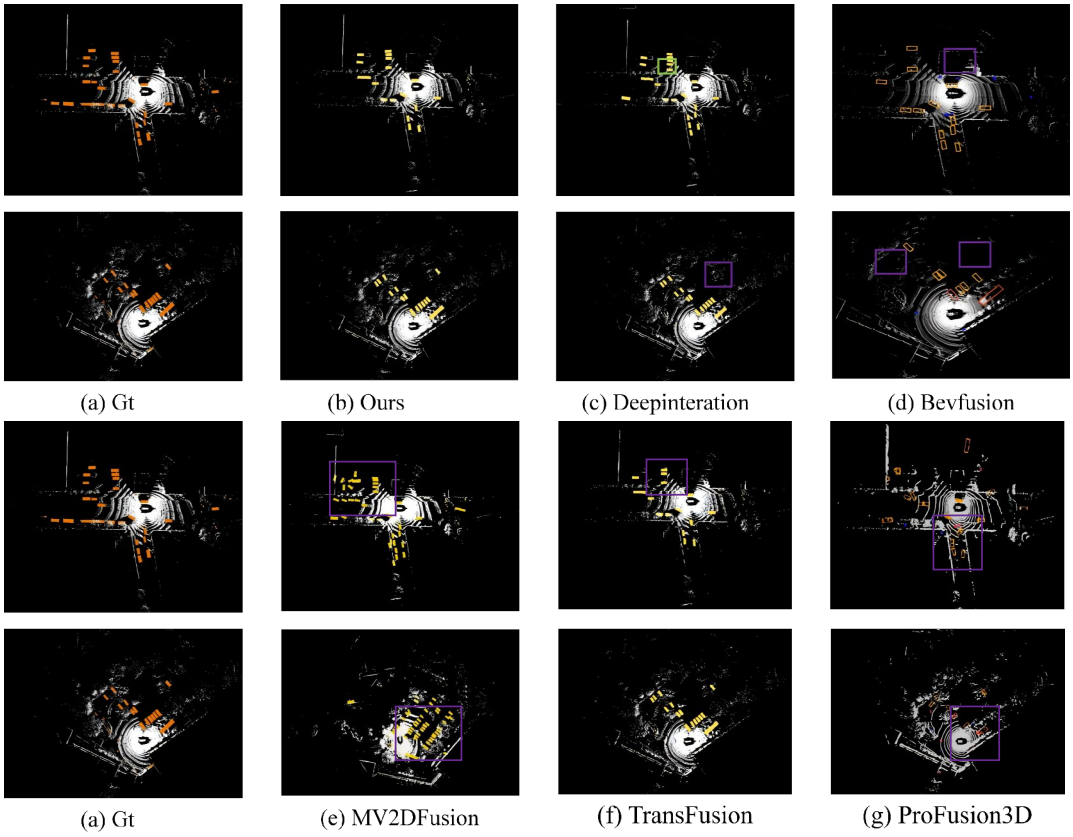


Figure E1 Qualitative comparison of 3D object detection under extreme sensor noise. (a) Ground-truth (GT) distribution across two distinct traffic scenarios. (b) Perception results of the proposed Ro3DOP framework, showing improved localization and recall. (c)–(g) Comparative results of DeepInteraction, BEVFusion, MV2DFusion, TransFusion, and ProFusion3D. Purple boxes highlight prominent false positives and hallucinated detections in the baseline models.

Appendix E Quantitative results

Appendix E.1 Comparison with LiDAR-camera fusion models

A comprehensive comparison between Ro3DOP and leading LiDAR-camera fusion frameworks under varying noise intensities is summarized in Table E1 and Table E2, with qualitative visualizations in Bird's-Eye-View (BEV) space provided in Fig. D4. The proposed framework demonstrates consistently strong robustness across all evaluated object categories and noise intensities, maintaining stable performance even under severe sensor degradation. For major categories such as cars and pedestrians, Ro3DOP maintains detection accuracies of 91.1% and 94.3% AP, respectively, as noise intensity increases from 60% to 90%. This performance stability indicates that the architecture effectively reduces the sensitivity of the final perception task to degraded sensor. Compared with competing methods, while other fusion models such as MV2DFusion, BEVFusion(MIT), and TransFusion exhibit noticeable performance degradation as noise intensifies, Ro3DOP better preserves the structural integrity of the fused feature manifold. The most noticeable performance is observed in bicycle detection, where Ro3DOP achieves 65.0% AP under 90% noise, surpassing DeepInteraction by 21.6%. Compared with ProFusion3D, which benefits from a higher inference throughput of 4.7 FPS and remains competitive on large-scale objects such as cars, our method provides stronger small-object robustness under severe sensor corruption. Although it is efficient, the progressive inference strategy of ProFusion3D degrades substantially on bicycles, dropping to 33.9% AP under 90% noise, and is more likely to hallucinated detections . Qualitative results from Fig. D4 further show that ProFusion3D produces more false positives under global sensor noise, which reflects the limitation of skip-based logic. In contrast, our LiDAR-Guided Attention (LGA) mechanism serves as a spatial constraint to reduce the interference of corrupted visual features with geometric anchors. Beyond precision metrics, the engineering feasibility of Ro3DOP is also validated by its consistent inference speed of 3.34 FPS on an NVIDIA RTX A6000 GPU: even with the integration of an iterative feature reconstruction module, the $O(N)$ linear complexity of our optimization keeps the system computationally efficient and competitive with dense attention-based paradigms such as BEVFormer and DeepInteraction. Overall, these results indicate that Ro3DOP achieves a balance between semantic reconstruction and real-time efficiency, while improving robustness across different object scales.

Appendix E.2 Comparative analysis with vision-only models

The comparative evaluation provides a clearer view of the structural weakness of vision-only perception under progressive sensor degradation. As shown in Tables E1 and E2, representative vision-based architectures such as BevFormer and Fast-BEV suffer significant performance degradation as the noise level increases, and this trend is especially pronounced at 90% noise, where pedestrian detection

Table E1 Quantitative Comparison of 3D Detection Performance under 60% Sensor Noise. AP (%) values are reported, with absolute improvements over baselines shown in brackets.

Method	Modality	FPS ↑	Cars	Pedestrians	Bicycle
Vision-only					
BevFormer	Camera	2.3	42.2 (-26.4)†	20.7 (-36.8)†	-
Fast-BEV	Camera	**8.5**	31.6 (-22.8)†	16.7 (-27.3)†	-
Multi-modal Fusion					
MV2DFusion	L + C	2.1	89.1 (-2.0)	90.3 (-4.0)	40.0 (-25.0)
BevFusion(MIT)	L + C	3.1	90.0 (-1.1)	91.8 (-2.5)	42.6 (-22.4)
TransFusion	L + C	2.8	89.7 (-1.4)	91.4 (-2.9)	52.0 (-13.0)
DeepInteraction	L + C	2.5	90.6 (-0.5)	91.0 (-3.3)	43.4 (-21.6)
ProFusion3D	L + C	**4.7**	90.5 (-0.6)	91.2 (-3.1)	44.1 (-21.1)
Ours (Ro3DOP)	**L + C**	**3.34**	**91.1**	**94.3**	**65.0**

† Performance drop compared to the noise-free scenario for vision-only models.

Negative values in brackets denote the AP gap between the corresponding method and Ro3DOP.

Table E2 Quantitative Comparison of 3D Detection Performance under 90% Sensor Noise. AP (%) values are reported, with absolute improvements over baselines shown in brackets.

Method	Modality	FPS ↑	Cars	Pedestrians	Bicycle
Vision-only					
BevFormer	Camera	2.3	20.4 (-48.2)†	3.2 (-54.3)†	-
Fast-BEV	Camera	**8.5**	14.0 (-40.4)†	4.0 (-40.0)†	-
Multi-modal Fusion					
MV2DFusion	L + C	2.1	85.2 (-5.9)	73.6 (-20.7)	36.1 (-28.9)
BevFusion(MIT)	L + C	3.1	87.6 (-3.5)	91.1 (-3.2)	42.3 (-22.7)
TransFusion	L + C	2.8	89.7 (-1.4)	90.6 (-3.7)	51.2 (-13.8)
DeepInteraction	L + C	2.5	90.6 (-0.5)	91.0 (-3.3)	43.4 (-21.6)
ProFusion3D	L + C	**4.7**	89.9 (-1.2)	89.0 (-5.3)	33.9 (-31.1)
Ours (Ro3DOP)	**L + C**	**3.34**	**91.1**	**94.3**	**65.0**

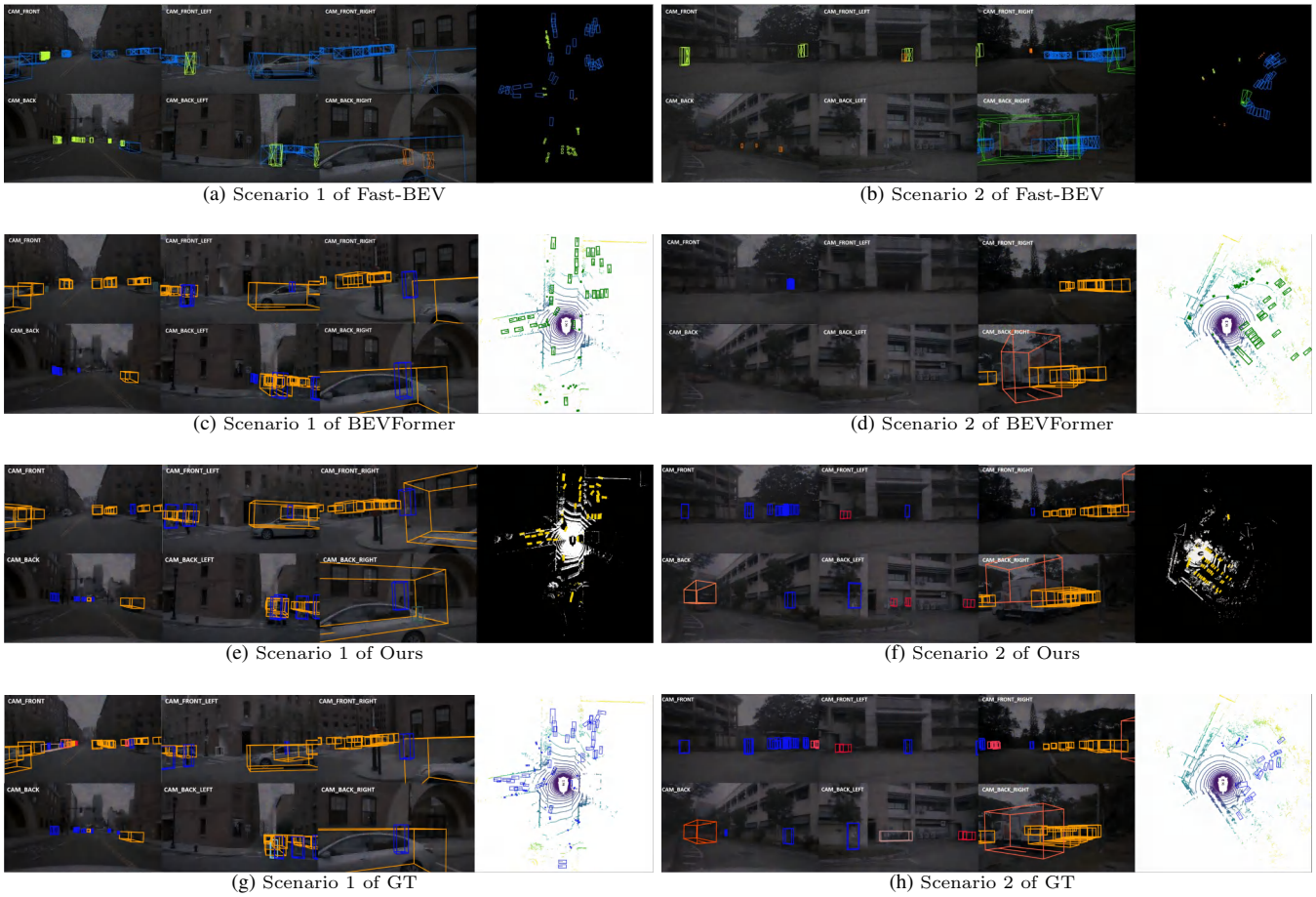
† Performance drop compared to the noise-free scenario for vision-only models.

Negative values in brackets denote the AP gap between the corresponding method and Ro3DOP.

precision drops to approximately 3.2% and 4.0%, respectively. Such a drop is not a minor variation but a severe collapse in detection reliability, suggesting that these algorithms become highly unrobust once the visual sensor is heavily corrupted. In other words, when the image modality can no longer provide stable texture and appearance information, the ability of vision-only methods to support reliable 3D inference is weakened significantly, while this tendency is also reflected in the qualitative results shown in Fig. E1. Under severe sensor corruption, the vision-only baselines produce frequent false negatives and noticeable localization bias, for instance, many targets are either missed entirely or placed at incorrect spatial positions. These failure cases are consistent with the behavior observed in the quantitative results. We infer that the semantic queries used by such methods are closely tied to texture-dependent depth information, once such information is disturbed by strong noise, the model may still retain partial semantic anchors, but it becomes much harder to anchor them accurately in 3D space. As a result, both target recognition and geometric localization decline at the same time. By comparison, Ro3DOP remains much more stable under the same degradation levels, which indicates that although the vision modality contributes important semantic information, robust perception in degraded settings still depends heavily on the geometric certainty provided by LiDAR. As a result, semantic richness alone is not enough when the visual stream becomes unreliable, and the model also needs a stable spatial reference to prevent the perception from drifting. The results indicate that LiDAR continues to play this anchoring role even when the vision modality is severely corrupted, which explains why the proposed framework can preserve stronger detection performance than vision-only baselines. Another point worth noting is that the advantage of Ro3DOP does not come simply from combining two sensor modalities while its effectiveness is more closely related to the way corrupted information is handled during fusion. The proposed framework adopts a modality-decoupled interaction mechanism, so the system can leverage the reconstructed latent manifold together with stable geometric priors from LiDAR instead of being forced to rely on the degraded visual stream in a rigid manner, which helps retain useful semantic information from corrupted image features while limiting the propagation of unreliable sensor responses into the final fused perception output. In this sense, the framework is not only intergarting multi-modal information but also actively controlling how degraded features influence cross-modal interaction. Overall, these observations explain why Ro3DOP maintains more robust and efficient perception results under severe sensor degradation. The comparison with vision-only models shows that performance robustness in corrupted environments depends not only on semantic representation quality, but also on whether the model can preserve stable geometric foundation and suppress the negative effect of corrupted feature streams during fusion.

References

1 Tian D, Zhou J, Han X, et al. Robust platoon control of mixed autonomous and human-driven vehicles for obstacle collision avoidance: A cooperative sensing-based adaptive model predictive control approach. Engineering, 2024, 42: 244–266



(a) Scenario 1 of Fast-BEV (b) Scenario 2 of Fast-BEV

(c) Scenario 1 of BEVFormer (d) Scenario 2 of BEVFormer

(e) Scenario 1 of Ours (f) Scenario 2 of Ours

(g) Scenario 1 of GT (h) Scenario 2 of GT

Figure E2 Qualitative evaluation of vision-centric 3D detection under simulated sensor failure. (a)–(d) Predictions from baseline models Fast-BEV and BEVFormer, showing substantial recall loss and orientation inaccuracies in noisy scenarios. (e)–(f) Results from the proposed Ro3DOP, showing predictions that are more consistent with (g)–(h) the ground truth across two distinct traffic scenes.

2 Li Y, Huang B, Chen Z, et al. Fast-bev: A fast and strong bird's-eye view perception baseline. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46: 8665–8679
3 Yang C, Chen Y, Tian H, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 17830–17839
4 Liu H, Teng Y, Lu T, et al. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In: Proceedings of Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 18580–18590
5 Zhang D, Zheng Z, Niu H, et al. Fully sparse transformer 3-d detector for lidar point cloud. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1–12
6 Liu Z, Hou J, Wang X, et al. Lion: Linear group rnn for 3d object detection in point clouds. Advances in Neural Information Processing Systems, 2024, 37: 13601–13626
7 Fan L, Wang F, Wang N, et al. Fsd v2: Improving fully sparse 3d object detection with virtual voxels. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024
8 Chen Y, Yu Z, Chen Y, et al. Focalformer3d: focusing on hard instance for 3d object detection. In: Proceedings of Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 8394–8405
9 Liang T, Xie H, Yu K, et al. Bevfusion: A simple and robust lidar-camera fusion framework. In: Proceedings of Koyejo S, Mohamed S, Agarwal A, et al., editors, Advances in Neural Information Processing Systems. Curran Associates, Inc., 2022. 10421–10434
10 Bai X, Hu Z, Zhu X, et al. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In: Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 1080-1089
11 Ge C, Chen J, Xie E, et al. Metabev: Solving sensor failures for bev detection and map segmentation. arXiv preprint arXiv:2304.09801, 2023
12 Park K, Kim Y, Kim D, et al. Resilient sensor fusion under adverse sensor failures via multi-modal expert fusion. In: Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025. 6720-6729
13 Mohan R, Cattaneo D, Drews F, et al. Progressive multi-modal fusion for robust 3d object detection. Conference on Robot Learning (CoRL), 2024
14 Everingham M, Gool L V, Williams C K I, et al. The pascal visual object classes (voc) challenge. International Journal of Computer Vision, 2010, 88: 303–338
15 Law H, Deng J. Cornernet: Detecting objects as paired keypoints, 2019
16 Zheng L, Tang M, Chen Y, et al. Improving multiple object tracking with single object tracking. In: Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 2453-2462
17 Gao P, Zheng M, Wang X, et al. Fast convergence of detr with spatially modulated co-attention, 2021
18 Caesar H, Bankiti V, Lang A H, et al. nuscenes: A multimodal dataset for autonomous driving. arXiv preprint arXiv:1903.11027, 2019
19 Chen K, Wang J, Pang J, et al. Mmdetection: Open mmlab detection toolbox and benchmark, 2019