SCIENCE CHINA
Information Sciences

• LETTER •

October 2026, Vol. 69, Iss. 10, 209203:1–209203:2
https://doi.org/10.1007/s11432-025-4946-2

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, Brighton BN1 9RH, UK*
[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 071000, China*

Received 15 October 2025/Revised 12 February 2026/Accepted 30 March 2026/Published online 26 August 2026

Citation Zhang Y C, Pang H B, Zhou J S, et al. Ro3DOP: robust 3D object perception via multi-modal feature reconstruction and adaptive fusion for autonomous vehicles. Sci China Inf Sci, 2026, 69(10): 209203, https://doi.org/10.1007/s11432-025-4946-2

As a critical task in autonomous driving perception, 3D object detection is essential for environmental understanding and operational safety. Recent studies have advanced from simple bird's-eye-view (BEV) transformations to sophisticated multi-modal fusion paradigms. Representative frameworks such as TransFusion [1] and ProFusion3D [2] demonstrate the effectiveness of transformer-based queries for bridging LiDAR and camera streams, while BEVFusion [3], DeepInteraction [4], and Mv2DFusion [5] further improve performance through tighter cross-modal integration. However, these approaches largely rely on a "perfect sensor" assumption, namely that both modalities provide reliable observations with accurate spatial calibration. In real-world scenarios, this assumption is often violated by environmental interference and hardware malfunctions, which may corrupt one modality and propagate errors through the fusion pipeline. Consequently, robust 3D perception under severe sensor degradation remains challenging in two aspects: (i) recovering semantically consistent features from corrupted observations, and (ii) suppressing cross-modal noise propagation during fusion. To address these challenges, we propose Ro3DOP, which combines high-dimensional feature reconstruction with geometry-constrained dual-attention fusion for robust 3D object perception. Methodologically, Ro3DOP is characterized by three aspects: active reconstruction of corrupted features, LiDAR-driven initialization of input-aware object queries, and LiDAR-guided attention for noise-resilient cross-modal interaction.

Method. Ro3DOP addresses the above challenges through two tightly coupled components, as illustrated in Figure 1(a). HD-FR reconstructs semantically consistent visual features from corrupted observations, while the dual-attention fusion module performs geometry-constrained cross-modal interaction to suppress noise propagation and refine 3D object detection.

HD-FR. We model the corrupted feature $\mathcal{X} \in \mathbb{R}^{C\times W\times H}$ as an incomplete observation of a high-dimensional tensor which encodes multi-dimensional structural and semantic dependencies. To address incomplete tensors caused by sensor failure, HD-FR focuses on structured recovery and exploits the structural and semantic correlations within multi-modal inputs, ensuring that the restored features maintain global coherence. The reconstruction task is formulated as an optimization problem, seeking to recover a semantically consistent tensor $\mathcal{Y}$:

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

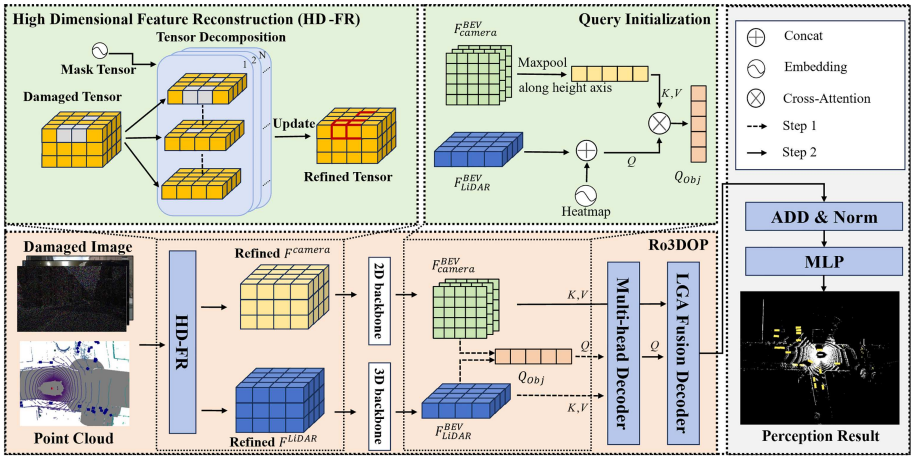
where $TD(\cdot)$ represents the tensor decomposition into latent factor tensors $\mathcal{U}_k$ and the decomposition rank R is determined by the feature manifold's intrinsic dimension, effectively balancing representation fidelity and memory consumption. To anchor the structural recovery, the regularization term $\mathcal{L}_s(\mathcal{Y})$ incorporates a mask tensor $\mathcal{M}$ (identified via sensor self-diagnosis) and an indicator function to enforce global consistency with available observations. This unsupervised proximal optimization framework ensures stable convergence and preserves local details under extreme noise.

Dual-attention fusion mechanism. Ro3DOP adopts a two-stage decoding process. Initially, input-aware object queries Q_{obj} are initialized from LiDAR BEV feature maps via heatmap-based hotspots, providing accurate spatial anchors.

$$Q_{obj} = \text{Top-}N(\text{Pool}(\phi(F_L))), \tag{2}$$

where F_L denotes the LiDAR BEV feature map, ϕ is the heatmap perception head, and Pool is a 3×3 max-pooling operation for local-maximum extraction. By selecting the Top-N high-confidence hotspots as initial spatial anchors, this input-driven

* Corresponding author (email: jianshanzhou@foxmail.com)
† These authors contributed equally to this work.



(a)

Method	Modality	FPS ↑	Cars	Pedestrians	Bicycle
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**

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

(b)

Figure 1 (a) The proposed framework consists of a high-dimensional feature reconstruction module and a dual-attention fusion decoder. (b) Quantitative comparison of average precision (AP) and inference speed (FPS) with representative multi-modal fusion baselines under 90% camera noise. Bold values indicate the best results for each metric. Detailed explanations are provided in the Supplementary file.

strategy ensures that queries are localized in potential target regions from the outset, significantly accelerating convergence compared to input-agnostic methods. Subsequently, we introduce the LiDAR-guided attention (LGA) mechanism to achieve resilient cross-modal integration. For a scene with n targets, a multi-peak spatial weighting function $\mathbf{W}_i$ is constructed by aggregating individual Gaussian kernels centered at $\mathbf{c}_i$:

$$\mathbf{W}_i = \exp\left(-\frac{\|\mathbf{p}-\mathbf{c}_i\|_2^2}{2\sigma^2}\right), \quad i \in \{1, 2, \cdots, n\}, \tag{3}$$

where $\mathbf{p}$ denotes a location on the image feature map, $\mathbf{c}_i$ is the center of the i-th object hypothesis, and σ controls the spatial tolerance. The resulting weight reweights the cross-attention map to emphasize object-related regions and suppress noise-contaminated visual responses. By using LiDAR-derived geometric priors as soft constraints, LGA improves robustness under severe sensor degradation and mitigates the effect of minor calibration errors.

Experiments. We evaluate Ro3DOP on the nuScenes-ArtNoise benchmark, which is specifically designed to simulate severe sensor degradation, and compare it against recent state-of-the-art fusion methods under the same challenging conditions. As summarized in Figure 1(b) and further detailed in the Supplementary Material, existing multi-modal baselines generally maintain relatively high accuracy for large objects, but their performance deteriorates markedly on small and geometrically sparse targets when exposed to severe noise. In contrast, Ro3DOP achieves 91.1% AP for cars and 94.3% AP for pedestrians, while showing a particularly notable advantage for bicycles, where it reaches 65.0% AP and surpasses DeepInteraction by 21.6% under 90% camera noise. These results demonstrate that HD-FR effectively restores semantically critical visual information from corrupted observations, while LGA suppresses noise propagation during cross-modal fusion and improves the reliability of feature interaction. Despite the additional iterative reconstruction step, Ro3DOP still runs at 3.34 FPS on an NVIDIA RTX A6000 GPU, remaining competitive with recent fusion baselines and demonstrating a favorable balance between robustness and efficiency.

Acknowledgements This work was partly supported by National Natural Science Foundation of China (Grant Nos. 62571014, 62432002, T2588101), Beijing Nova Program (Grant No. 20250484936), Beijing Municipal Natural Science Foundation (Grant No. 4252005), Beijing-Tianjin-Hebei Basic Research Cooperation Project (Grant No. F2024201070), Fundamental Research Funds for the Central Universities (Grant No. JKF-2025056347186), and Xplorer Prize.

Supporting information Appendixes A–E. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 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
2 Mohan R, Cattaneo D, Drews F, et al. Progressive multi-modal fusion for robust 3d object detection. In: Proceedings of Conference on Robot Learning (CoRL), 2024
3 Liang T, Xie H, Yu K, et al. Bevfusion: a simple and robust lidar-camera fusion framework. In: Proceedings of Advances in Neural Information Processing Systems, 2022. 10421–10434
4 Yang Z, Chen J, Miao Z, et al. Deepinteraction: 3d object detection via modality interaction. In: Proceedings of Advances in Neural Information Processing Systems, 2022. 35: 1992–2005
5 Wang Z, Huang Z, Gao Y, et al. MV2DFusion: leveraging modality-specific object semantics for multi-modal 3D detection. IEEE Trans Pattern Anal Mach Intell, 2026, 48: 609–623