

MAPLE: Modality-Aware Projection-free LiDAR-Camera Fusion for 3D Vehicular Object Detection

Abhishek Balasubramaniam and Sudeep Pasricha
 Department of Electrical and Computer Engineering
 Colorado State University, Fort Collins, Colorado, USA
 {abhishek.balasubramaniam, sudeep}@colostate.edu

Abstract— Accurate 3D object detection (3D-OD) is critical for autonomous vehicles, yet embedded platforms impose strict latency, power, and memory constraints. While LiDAR-camera fusion improves robustness, existing approaches depend on precise calibration and computationally expensive view projections. We present MAPLE, a projection-free and calibration-resilient fusion framework that adaptively balances LiDAR geometry and camera semantics using Gated Confidence Fusion (GCF) and low-rank adapter (LoRA) enhanced global attention refinement. MAPLE preserves fine-grained cross-modal interactions without view lifting and injects long-range context at low cost. On the nuScenes benchmark, MAPLE improves mean Average Precision (mAP) by up to 1.6% over the strongest prior fusion baseline, while reducing inference latency by 42.6% and energy consumption by 47% on the NVIDIA Jetson Orin Nano, demonstrating suitability for real-time embedded autonomous perception.

Keywords— Sensor fusion, 3D object detection, BEV perception, LiDAR, camera, LoRA, Autonomous driving.

I. INTRODUCTION

Autonomous vehicle (AV) perception requires accurate 3D object detection under strict power and resource constraints imposed by embedded automotive hardware [1] operating alongside multiple safety-critical workloads [2], [3]. This challenge is particularly acute for 3D object detection, which extends 2D detection by integrating depth, scale, and localization essential for high accuracy in complex driving scenarios, but at high memory and computational cost [5], [7]. Bird’s-eye-view (BEV) representations enable structured 3D reasoning [6] but typically rely on precise calibration and expensive view transformations [11]. Multi-sensor fusion, e.g., combining LiDAR’s geometric accuracy with cameras’ semantic richness, enables more robust perception than single modality approaches [10]. But existing methods suffer from calibration sensitivity [8], projection overhead [11], loss of fine-grained interaction [9], or high computational cost on embedded platforms [12], [13].

These challenges motivate MAPLE, a projection-free and modality-aware LiDAR-camera fusion framework that adaptively balances modalities and injects global context using lightweight LoRA-enhanced attention for efficient real-time embedded 3D perception. Our novel contributions are:

- A projection-free and calibration-resilient vehicular fusion framework for LiDAR BEV and image features.
- A confidence-aware gated fusion with LoRA-based global attention for efficient long-range reasoning.
- An embedded-optimized design with superior accuracy, efficiency, and robustness under sensor failures.

II. RELATED WORK

Modern 3D object detectors (ODs) are central to autonomous vehicle (AV) perception [14]. LiDAR-based methods (e.g., PointNet, PointPillars, SECOND, VoxelNext [15]–[18]) provide accurate localization at high computational

cost, while camera-only approaches (e.g., Monoflex, SMOKE [19], [20]) are lightweight but limited by unreliable depth, motivating LiDAR-camera fusion. Fusion strategies vary by integration stage: transformer-based methods such as FUTR3D [6] are computationally expensive, hybrid approaches like DeepInteraction++ [11] lack confidence awareness, attention-based designs such as TransFusion [13] incur high latency, and late fusion methods such as BEVFusion-e [9] sacrifice fine-grained cross-modal interaction. As a result, existing methods remain constrained by calibration dependence, projection overhead, limited robustness to sensor degradation, and poor accuracy-efficiency trade-offs on embedded platforms [4], [12]. In contrast, our MAPLE framework adopts a projection-free, confidence-aware fusion strategy with LoRA-enhanced global attention for robust and efficient embedded deployment.

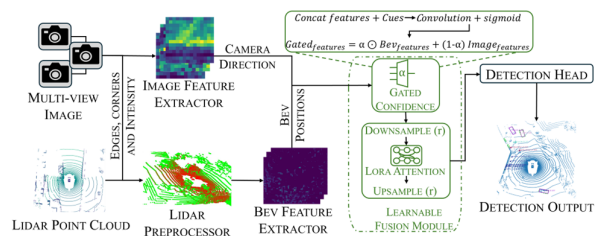


Fig 1 Overview of MAPLE framework with a learnable fusion module that uses confidence-based gating for individual sensor modalities along with low-rank adapters (LoRA) for global attention refinement.

III. MAPLE FRAMEWORK OVERVIEW

In this section, we describe our novel LiDAR-camera fusion framework MAPLE (Fig. 1). The proposed framework consists of four components: (1) a LiDAR preprocessor, (2) modality-specific feature extractors, (3) a learnable fusion module, and (4) a lightweight detection head.

A. LiDAR Preprocessor

Raw LiDAR point clouds are sparse and uneven, particularly at long ranges. MAPLE applies a structure-guided preprocessing step that enriches LiDAR points using image-derived edge, corner, and intensity priors extracted from image projections [22], generating pseudo-points along object contours without heavy voxelization. The enriched point cloud is processed by a sparse 3D backbone (e.g., CenterNet3D or VoxelNext [18]) and collapsed into BEV features encoding road layout and object geometry, improving robustness under occlusion and sensor degradation.

B. Feature Extractor

After LiDAR preprocessing, MAPLE extracts modality-specific features from LiDAR and camera inputs to form complementary representations for fusion. Densified LiDAR point clouds are voxelized and processed by a sparse 3D backbone to produce Bird’s-Eye-View (BEV) features that encode road layout, lane structure, and object geometry. In parallel, each camera view is processed by a lightweight

TABLE 1 COMPARISON OF DETECTION PERFORMANCE OF MAPLE WITH PRIOR FUSION BASELINES ON THE nuSCENES DATASET.

Models	mAP	Car	Motorcycle	Pedestrian	RTX 4060 Ti Latency (ms)	Jetson Latency (ms)	RTX 4060 Ti Energy (J)	Jetson Energy (J)	mAP with Camera failure	
									1 Camera	2 Camera
TransFusion [16]	0.689	0.871	0.736	0.884	204.41	697.39	9.811	9.34	-	-
FUTR3D [8]	0.694	0.780	0.614	0.757	237.63	619.83	7.32	8.05	-	-
BEVFusion-e [12]	0.750	0.905	0.844	0.918	119.05	316.67	5.714	6.72	0.721	0.584
DeepInteraction++ [14]	0.756	0.883	0.854	0.925	181.8	498.04	8.726	6.97	0.732	0.594
MAPLE (CenterNet3D)*	0.757	0.914	0.852	0.927	98.53	310.56	3.436	4.22	0.743	0.615
MAPLE (VoxelNext)*	0.768	0.927	0.864	0.942	76.85	285.93	2.439	3.69	0.758	0.697

convolutional backbone to extract semantic feature maps, which are tokenized and aligned with the BEV space using positional encodings without explicit projection or calibration. The resulting geometry-aware BEV features and semantic image features are passed to the learnable Gated Confidence Fusion module (Section III.C) where modality importance is adaptively determined per region based on spatial reliability and sensor quality, enabling robust fusion under occlusion and sensor degradation.

C. Learnable Fusion Module

The learnable fusion stage in MAPLE unifies LiDAR BEV and multi-view image features in a projection-free, calibration-resilient manner using a two-stage design optimized for robustness and efficiency. We use Gated Confidence Fusion (GCF) to adaptively balance LiDAR and image tokens based on geometry-aware cues (camera direction, BEV position) and learned modality confidences, allowing the most reliable modality to dominate under sensor sparsity, occlusion, or noise. This confidence-aware gating suppresses unreliable or hallucinated features while preserving fine-grained cross-modal interactions, enabling voxel-level fusion without explicit 2D–3D projection.

Global context is then injected through a lightweight Global Refinement Block tailored for embedded deployment. Sparse BEV tokens are pooled into compact spatial bins with bounded sequence length and processed using LoRA-enhanced multi-head self-attention to capture long-range dependencies at low parameter and memory cost [23]. A lightweight channel-wise MLP refines the contextualized tokens, which are upsampled and residually fused back into the BEV map. Together, confidence-aware local fusion and efficient global context modeling preserve geometric detail and scene-level consistency while maintaining low latency and predictable memory usage, enabling robust real-time 3D perception under sensor degradation.

D. Detection Head

The detection head operates on the refined BEV features produced by the global refinement block and converts them into final 3D bounding box predictions. It consists of lightweight convolutional classification and regression branches that predict object presence and 3D box parameters, including position, dimensions, orientation, and velocity, at each BEV location. Final detections are obtained using non-maximum suppression to remove redundant predictions. This design enables accurate and temporally consistent 3D object detection while remaining computationally efficient for real-time embedded deployment.

IV. EXPERIMENTAL RESULTS

A. Experimental Setup

We evaluate MAPLE on the multimodal nuScenes dataset [21] using the standard 80/20 train–validation split. Experiments are conducted on an NVIDIA RTX 4060 Ti [25]

based workstation (370W) for training and evaluation and an NVIDIA Jetson Orin Nano (15W) [24] for embedded inference profiling. MAPLE is implemented in PyTorch [26] on top of OpenPCDet [27] and processes all six camera views jointly with each LiDAR frame for end-to-end latency measurement. CenterNet3D [28] and VoxelNext [18] are used as LiDAR backbones with MobileNetV3-Large [29] for image features. Robustness is evaluated via camera-view dropout and embedded power is measured using NVPower [30]. Lightweight fusion in MAPLE is implemented with four attention heads with 64-dimensional projections, LoRA rank 8, and a compact global refinement block to enable efficient real-time deployment.

B. Evaluation results

MAPLE is evaluated against state-of-the-art fusion methods including BEVFusion-e [9], TransFusion [13], DeepInteraction++ [11], and FUTR3D [6] on the nuScenes validation set (Table 1). With a CenterNet3D backbone, MAPLE achieves 0.757 mAP, marginally surpassing the strongest prior baseline (DeepInteraction++ at 0.756), while adopting a stronger VoxelNext backbone further improves performance to 0.768 mAP, yielding gains of up to 1.6% over DeepInteraction++ and more than 10% over TransFusion and FUTR3D. MAPLE delivers consistent improvements across object categories, with strong performance on safety-critical classes such as pedestrians and motorcycles, while remaining competitive on challenging classes like construction vehicles and trailers. Efficiency results highlight MAPLE’s favorable accuracy–efficiency trade-off. On an RTX 4060 Ti, MAPLE (VoxelNext) reduces inference latency and energy consumption by 58% and 72%, respectively, compared to DeepInteraction++, while on the Jetson Orin Nano it achieves 42.6% lower latency and 47% lower energy, consistently outperforming BEVFusion-e. Robustness experiments with induced camera failures further show that MAPLE maintains higher mAP under single and multi-camera dropouts by adaptively prioritizing reliable LiDAR cues. Together, these results demonstrate that MAPLE achieves superior accuracy, efficiency, and robustness for real-time embedded autonomous perception.

V. CONCLUSIONS

In this paper, we presented MAPLE, a projection-free and modality-aware fusion framework for efficient 3D perception in autonomous driving. MAPLE fuses LiDAR and camera features directly in BEV space using confidence-aware gating and LoRA-enhanced attention, avoiding explicit projection and calibration. MAPLE outperforms the strongest prior baseline by 1.6% mAP while reducing inference latency and energy consumption on Jetson Orin Nano by 42.6% and 47%, respectively, and maintains robust performance under partial and multi-camera sensor failures. Our ongoing work is exploring integration of sensor selection/placement [31]–[33], security primitives [34]–[38], and GPS-free localization [39]–[43] with perception in complex vehicular environments.

REFERENCES

- [1] A. Balasubramaniam, S. Pasricha, "Object detection in autonomous vehicles: Status and open challenges", *arXiv preprint arXiv:2201.07706*, 2022.
- [2] G. B. Thieu, S. Gesper, G. Payá-Vayá, C. Riggers, O. Renke, T. Fiedler, C. Sauer, "ZuSE Ki-Avf: Application-Specific AI Processor for Intelligent Sensor Signal Processing in Autonomous Driving," *2023 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Belgium, 2023.
- [3] V. Kukkala, J. Tunnell, S. Pasricha, "Advanced Driver Assistance Systems: A Path Toward Autonomous Vehicles", *IEEE Consumer Electronics*, vol. 7, no. 5, Sept 2018.
- [4] T. Kotrba, M. Lechner, O. Sarwar, A. Jantsch, "Multispectral Feature Fusion for Deep Object Detection on Embedded NVIDIA Platforms," *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Antwerp, Belgium, 2023
- [5] A. Balasubramaniam, F. Sunny, S. Pasricha, "R-TOSS: A framework for real-time object detection using semi-structured pruning", *2023 60th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 2023.
- [6] X. Chen, T. Zhang, Y. Wang, Y. Wang, H. Zhao, "Futr3d: A unified sensor fusion framework for 3d detection", In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 2023.
- [7] A. Balasubramaniam, F. Sunny, S. Pasricha, "UPAQ: A Framework for Real-Time and Energy-Efficient 3D Object Detection in Autonomous Vehicles", *2025 Design, Automation & Test in Europe Conference (DATE)*, Lyon, France, 2025.
- [8] S. Vora, A.H. Lang, B. Helou, O. Beijbom, "PointPainting: Sequential Fusion for 3D Object Detection." In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.
- [9] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. Rus, S. Han, "Befusion: Multi-task multi-sensor fusion with unified bird's-eye view representation", *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.
- [10] L. Wang, X. Zhang, Z. Song, J. Bi, G. Zhang, H. Wei, L. Zhao, "Multi-modal 3d object detection in autonomous driving: A survey and taxonomy" *IEEE Transactions on Intelligent Vehicles*, 2023.
- [11] Z. Yang, N. Song, N. W. Li, X. Zhu, L. Zhang, P.H. Torr, "Deepinteraction++: Multi-modality interaction for autonomous driving", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [12] J. Ku, M. Mozifian, J. Lee, A. Harakeh, S.L. Waslander, "Joint 3d proposal generation and object detection from view aggregation", *IEEE/RSJ international conference on intelligent robots and systems (IROS)*, 2018.
- [13] C. Zhou, L. Yu, A. Babu, K. Tirumala, M. Yasunaga, L. Shamis, O. Levy, "Transfusion: Predict the next token and diffuse images with one multi-modal model", *arXiv preprint arXiv:2408.11039*, 2024.
- [14] V. Kukkala, S. Pasricha, "Machine Learning and Optimization Techniques for Automotive Cyber-Physical Systems", *Springer Nature Publishers*, 2023.
- [15] Q. Charles, H. Su, K. Mo, J. Leonidas, "Pointnet: Deep learning on point sets for 3d classification and segmentation", *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.
- [16] H. Alex, S. Vora, H. Caesar, L. Zhou, J. Yang, O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds", *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019.
- [17] Y. Yan, Y. Mao, B. Li, "Second: Sparsely embedded convolutional detection", *Sensors* 18.10, 2018.
- [18] Y. Chen, J. Liu, X. Zhang, X. Qi, J. Jia, "Voxelnext: Fully sparse voxelnet for 3d object detection and tracking", In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2023.
- [19] Z. Yunpeng, J. Lu, J. Zhou, "Objects are different: Flexible monocular 3d object detection", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.
- [20] L. Zechen, Z. Wu, R. Tóth, "Smoke: Single-stage monocular 3d object detection via keypoint estimation", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020.
- [21] H. Caesar, B. Varun, H. A. Lang, S. Vora, V.E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, O. Beijbom, "nusenes: A multimodal dataset for autonomous driving", In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.
- [22] W. Gao, X. Zhang, L. Yang, H. Liu, "An improved Sobel edge detection". In *3rd International conference on computer science and information technology*, IEEE, 2010.
- [23] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen "Lora: Low-rank adaptation of large language models", *The International Conference on Learning Representations (ICLR)*, 2022.
- [24] "Jetson Orin nano developer kit getting started - nvidia developer," <https://developer.nvidia.com/embedded/learn/get-started-jetson-orin-nano-devkit>, (last accessed on Aug. 24, 2025).
- [25] "Nvidia RTX 4060Ti workstation", <https://www.nvidia.com/en-us/geforce/graphics-cards/40-series/rtx-4060-4060ti>, (Accessed on August 24, 2025).
- [26] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, M. Raison, "Pytorch: An imperative style, high-performance deep learning library", *Advances in neural information processing systems*, 2019.
- [27] OpenPCDet Development Team. "Openpcdet: An open-source toolbox for 3d object detection from point clouds", <https://github.com/open-mmlab/OpenPCDet>, 2020.
- [28] Q. Wang, J. Chen, J. Deng, and X. Zhang, "3D-CenterNet: 3D object detection network for point clouds with center estimation priority", *Pattern Recognition*, 2021.
- [29] B. Koonce, "MobileNetV3", In *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization*, Berkeley, CA: Apress, 2021.
- [30] wildkid1024, "NVpower at master wildkid1024/NVpower," <https://github.com/wildkid1024/NVpower/tree/master/NVpower> [last accessed Aug. 24, 2025].
- [31] J. Dey, W. Taylor, S. Pasricha, "VESPA: Optimizing Heterogeneous Sensor Placement and Orientation for Autonomous Vehicles", *IEEE Consumer Electronics*, 10(2), Mar 2021.
- [32] J. Dey, S. Pasricha, "Co-Optimizing Sensing and Deep Machine Learning in Automotive Cyber-Physical Systems", *IEEE Euromicro Conference on Digital Systems Design*, 2022.
- [33] J. Dey, S. Pasricha, "Robust Perception Architecture Design for Automotive Cyber-Physical Systems", *IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 2022.
- [34] V. K. Kukkala, S. V. Thiruloga, S. Pasricha, "Roadmap for Cybersecurity in Autonomous Vehicles", *Vol. 11, Iss. 6, pp. 13-23, IEEE Consumer Electronics*, Nov 2022.
- [35] V. K. Kukkala, S. V. Thiruloga, and S. Pasricha, "LATTE: LSTM Self-Attention based Anomaly Detection in Embedded Automotive Platforms", *ACM TECS, Volume 20, Issue 5s*, Oct 2021.
- [36] V. K. Kukkala, S. V. Thiruloga, and S. Pasricha, "INDRA: Intrusion Detection using Recurrent Autoencoders in Automotive Embedded Systems", *IEEE TCAD*, 39(11), Nov 2020.
- [37] V. Kukkala, S. Pasricha, T. H. Bradley, "SEDAN: Security-Aware Design of Time-Critical Automotive Networks", *IEEE Transactions on Vehicular Technology (TVT)*, vol. 69, no. 8, Aug 2020.
- [38] V. Kukkala, S. Pasricha, "Machine Learning and Optimization Techniques for Automotive Cyber-Physical Systems", *Springer Nature Publishers*, 2023.
- [39] S. Tiku, S. Pasricha, "Machine Learning for Indoor Localization and Navigation", *Springer Nature Publishers*, 2023.
- [40] S. Tiku, P. Kale, S. Pasricha, "QuickLoc: Adaptive Deep-Learning for Fast Indoor Localization with Mobile Devices", *ACM Transactions on Cyber-Physical Systems (TCPS)*, Vol. 17, Iss. 4, Oct 2021.
- [41] S. Tiku, D. Gufran, S. Pasricha, "Multi-Head Attention Neural Network for Smartphone Invariant Indoor Localization", *IEEE Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 2022
- [42] A. Singampalli, S. Pasricha, "Unified Class and Domain Incremental Learning with Mixture of Experts for Indoor Localization", *IEEE/ACM DATE, Verona, Italy*, Mar 2026.
- [43] D. Gufran, S. Tiku, S. Pasricha, "STELLAR: Siamese Multi-Headed Attention Neural Networks for Overcoming Temporal Variations and Device Heterogeneity with Indoor Localization", *IEEE Journal of Indoor and Seamless Positioning and Navigation*, 2023.