Theory and Practice of Science and Technology
2025, VOL. 6, NO. 8, 81-85
DOI: 10.47297/taposatWSP2633-456915.20250608

Multi-Modal Fusion for 3D Object Detection in Autonomous Driving: A Review

Haozhou Fei

Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu, 210023, China

ABSTRACT
Autonomous driving technologies have seen significant advancements due to improvements in artificial intelligence and sensor systems. A crucial element for safe and effective autonomous driving is the environmental perception system, with 3D object detection serving as a core task for vehicle awareness and decision-making. However, single-modality sensors, such as cameras, LiDAR, and radar, each have inherent limitations. Cameras provide rich semantic data but struggle under adverse conditions, LiDAR offers precise geometric data but is costly, and radar, while robust in all weather, lacks spatial resolution. Multi-modal sensor fusion addresses these challenges by combining complementary data from different sensor types, leading to a more comprehensive environmental understanding. This paper surveys the latest advancements in 3D object detection using multi-modal fusion for autonomous driving, focusing on feature-level fusion techniques, particularly the employment of Bird's-Eye-View (BEV) representation and Transformer architectures. The paper also explores the development of datasets, the evolution of fusion strategies, and the critical challenges faced, such as data alignment, performance efficiency, and robustness in harsh conditions. Finally, it outlines the future research directions aimed at optimizing the fusion process, including lightweight architectures and the integration of foundation models for improved performance and generalization.

KEYWORDS
Multi-modal fusion; Bird's-Eye-View (BEV); 3D object detection; Transformer; Feature fusion; Autonomous driving

1 Introduction

In recent years, substantial advancements in artificial intelligence and sensor technology have established autonomous driving as a critical technology for improving road safety, primarily by mitigating the human error responsible for up to 90% of traffic accidents. The cornerstone of this endeavor is a robust environmental perception system, for which 3D object detection is a core function that directly informs the vehicle's situational awareness and subsequent decision-making. However, perception systems that rely on a single sensor modality are constrained by inherent limitations. Cameras, for instance, provide rich semantic information but are highly susceptible to variations in lighting and weather, and they lack the ability to sense depth directly. LiDAR offers precise 3D geometric data but is impeded by high costs and an inability to capture texture or color information. Similarly, while millimeter-wave radar is robust in all weather conditions and can measure velocity, it suffers from low spatial resolution.

To overcome these shortcomings, multi-modal sensor fusion has become an indispensable strategy. A fusion system integrates complementary data: cameras provide rich semantic information, while LiDAR provides precise geometric information. This integration yields a more comprehensive, accurate, and robust environmental representation, leading to significantly improved object detection reliability and robustness.

This paper provides a comprehensive survey of 3D object detection techniques that leverage multi-modal fusion for autonomous driving. We focus on the dominant feature-level fusion paradigm and its various technical approaches. We analyze the pivotal roles that the unified Bird's-Eye-View (BEV) representation and the Transformer architecture play in modern fusion frameworks. By examining representative algorithms and their associated challenges, this review offers a clear overview of the state of the art and provides insights into future research directions.

2 Background

2.1 Sensor characteristics analysis

Camera: As the most mature sensor in autonomous driving perception, cameras provide high-resolution color/texture semantics at low cost. However, their passive sensing nature precludes direct depth acquisition, limiting 3D localization accuracy. Performance degrades significantly under low illumination, glare, and adverse weather, failing to meet all-scenario safety requirements [1].

LiDAR: By actively emitting laser pulses, LiDAR delivers precise 3D geometric structures with consistent day/night performance. Key limitations include high cost, lack of semantic cues, and point cloud sparsity causing high miss rates for distant objects. Laser attenuation in dense fog/heavy rain reduces effective [1].

Radar: Leveraging the Doppler effect for direct radial velocity/distance measurement, its electromagnetic wave penetration ensures performance degradation in rain/snow/fog, providing exceptional all-weather robustness. However,

low angular resolution blurs object contours, and susceptibility to metallic clutter increases false alarms, rendering it inadequate for high-fidelity standalone perception [1].

Table 1 ComparISON OF THREE SENSORS IN 3D object detection.

	Camera	LiDAR	Radar
Operating Principle	Passive optical (ambient light)	Active laser ToF pulses	Active EM waves (Doppler effect)
Strengths	High-res color/texture; low cost; semantic detail	Precise 3D geometry; consistent day/night; metric accuracy	All-weather reliability; direct velocity/distance; rain/fog penetration
Limitations	No direct depth; fails in low light/glare/weather	High cost; no color/material semantics; sparse distant points; rain/fog attenuation	Low angular resolution (blurs shapes); metallic clutter → false alarms
Standalone Capability	Semantic understanding only (no depth)	Geometric modeling only (no semantics)	Insufficient for full perception (low detail)

2.2 Dataset

Key datasets for 3D object detection relevant to autonomous driving, prominent current datasets include KITTI [2], nuScenes [3], Waymo Open [4], Argoverse [5], ApolloScape [6], PandaSet [7], H3D [8], among others, along with the most recent datasets for 3D object detection.

Table 2 Frequently utilized datasets for object detection

Dataset Name	Year	Loc	camera	lidar	radar	Classes	3D Annos	Primary Application
KITTI [2]	2012	Germany	4	1	-	8	~200k	Classic benchmark for 3D object detection and tracking
nuScenes [3]	2020	Singapore, USA	6	1	5	23	~1.4M	360° full-surround, multi-modal perception
Waymo Open [4]	2020	USA	5	5	-	4	~12M	Large-scale learning and model generalization
Argoverse [5]	2019	USA	7	2	-	15	~993k	3D tracking and trajectory forecasting
ApolloScape [6]	2019	China	6	1	-	6	~475k	Complex urban scene understanding, multi-task learning
PandaSet [7]	2021	USA	6	2	-	28	~1.3M	Perception with diverse sensor configurations
H3D [8]	2019	USA	3	1	-	8	~1.1M	Full-surround 3D multi-object detection and tracking
A2D2 [9]	2020	Germany	6	5	-	14	>12k frames with 3D labels	Audi autonomous driving dataset, multi-modal perception
RADIATE [10]	2021	UK	1	-	1	8	~37k	Robust perception in adverse weather conditions

3 Hierarchical framework for multimodal fusion methods

Early multimodal fusion frameworks classified approaches by processing stage (early, middle, late fusion). Contemporary frameworks instead characterize fusion hierarchically by information abstraction level: data-level, feature-level, and decision-level. This shift arose due to limitations in traditional classification frameworks. Within complex multi-branch deep networks, the category "middle fusion" became increasingly broad, subsuming most contemporary methods and thereby losing its discriminative utility. Consequently, the new hierarchical perspective emerged: data-level fusion integrates information at the raw data layer; feature-level fusion performs deep interaction on intermediate feature representations extracted by the network; and decision-level fusion integrates the independent predictions from each modality. The majority of current research is conducted within the framework of feature-level fusion. This paper will summarize the concepts of data-level and decision-level fusion in brief, before undertaking a detailed analysis of feature-level fusion, which constitutes the primary focus of this study.

3.1 Data-level fusion

In data-level fusion, raw or minimally preprocessed multimodal data is integrated prior to deep feature extraction in the backbone network. This construction of a unified input representation eliminates the need for modality-specific adaptation in subsequent networks. Its primary advantage lies in enabling fused data to be processed by any single-modality architecture, though it imposes stringent requirements on sensor synchronization.

Region-guided fusion constrains 3D spatial processing through 2D image detection results, with F-PointNet pioneering

the projection of 2D bounding boxes into 3D frustums using camera-LiDAR calibration parameters; subsequent point cloud segmentation and 3D bounding box regression are confined within these frustums, reducing computational complexity [11]. Semantic-enhancement fusion improves sparse data quality via cross-modal semantic transfer, where PointPainting maps image segmentation scores onto LiDAR points through "pixel painting" to boost small-object classification accuracy [12]. PointAugmenting directly superimposes image data with point cloud information, thereby enhancing the representational capability of point clouds. However, the alignment accuracy between point clouds and images has a significant impact on the overall performance of the detection model [13].

3.2 Feature-level fusion

Feature-level fusion, also known as middle fusion, is currently the most widely used and actively researched technical route in the field of multi-modal 3D object detection. This strategy involves the deep interaction and integration of feature maps at the intermediate layers of the network, after deep features have been extracted from different modality data by their respective encoders. The advantage of feature-level fusion is that it effectively balances information fidelity with implementation complexity: compared to data-level fusion, it is less sensitive to spatio-temporal calibration errors between sensors; compared to decision-level fusion, it can fully exploit the deep semantic correlations and complementarities between modality features before a final decision is made [14].

Transforming multi-modal features into a unified Bird's-Eye-View (BEV) space for fusion has become the most advanced and mainstream paradigm. The BEV representation provides a regular 2D grid that is consistent with the real world's scale, effectively solving the feature misalignment caused by different sensor views, while preserving the physical size of objects and significantly reducing occlusions. The foundational work, BEVFusion, pioneered this paradigm by using a differentiable view transformation method to accurately project multi-view image features into the BEV space. These features are then simply concatenated with LiDAR point cloud-derived BEV features and fed into a unified BEV encoder for processing. This framework demonstrated the immense potential of feature fusion in the BEV space but also suffered from larger projection errors for dynamic objects and performance degradation in adverse weather [15]. Subsequent research has optimized this foundation in various ways: BEVFusion4D introduced temporal information, using a Time Warping Alignment (TDA) module to dynamically aggregate historical BEV features, effectively mitigating ambiguity in dynamic scenes; to balance accuracy and efficiency [16], SimpleBEV improved real-time performance by simplifying the depth estimation process and reducing algorithm inference overhead [17].

As research has progressed, simple feature concatenation has become insufficient, leading researchers to explore more complex feature interaction mechanisms, with the application of the Transformer being particularly prominent. The Transformer architecture enables deeper interaction and alignment between features of different modalities through its core Cross-Attention Mechanism. Such methods typically treat features from one modality (e.g., camera images) as the Query, and features from another modality (e.g., LiDAR point clouds) as the Key and Value. By analyzing the similarity (attention scores) between the Query and the Key, the model can selectively aggregate the most relevant information from the Value, thereby effectively injecting the contextual information of one modality into the features of another. For deeper feature enhancement, ECL3D proposes an Instance BEV Feature Enhancement (IBFE) module, which innovatively uses Deformable Cross-Attention to allow image and LiDAR BEV features to serve as mutual queries, dynamically aligning and mutually enhancing object-relevant features during interaction, effectively suppressing background noise [18]. FusionFormer improves the Transformer encoder itself, using cross-attention and temporal fusion modules to enhance multi-modal interaction and context understanding while avoiding the geometric information loss in traditional BEV projection [19].

Within the feature-level fusion framework, proposal-level fusion has formed a significant technical branch. Its core idea is to leverage proposals generated from one sensor modality to guide the feature extraction and fusion process of another modality. F-PointNet is the pioneering work of this route, first using a 2D image detector to generate 2D bounding boxes (proposals), and then extruding them into 3D frustums in space. Subsequent 3D detection is performed only within these narrowed frustums, drastically reducing the 3D search space. However, this method relies heavily on the accuracy of the 2D detector and its rigid geometric projection. The more modern TransFusion represents an evolution towards the Transformer architecture [20]. It employs a two-stage process where LiDAR-only proposals are generated in the first stage, and in the second stage, these proposals act as queries to interact with image features via a cross-attention mechanism. This "soft alignment" uses rich image semantics to refine the geometric features of the LiDAR proposals. MV2DFusion is dedicated to solving the "modal bias" problem. Its core innovation is a dual-branch query generator: the image branch queries model depth uncertainty, while the point cloud branch queries retain precise geometric priors. This design allows the model to dynamically balance the contributions of different modalities based on the scene [21].

Although BEV is the mainstream, compressing 3D features into a 2D plane inevitably results in the loss of height information. To address this, some studies have explored alternative fusion paradigms. UniVoxel proposes an innovative idea: instead of directly converting features to BEV, it first fuses image and LiDAR information in 3D space via a Fusion

Voxel Encoder (FVE) to generate unified Multi-modal Voxels, aiming to preserve more complete 3D spatial information [22]. CMT (Cross Modal Transformer) takes a different path, avoiding explicit view transformation by treating both image and point cloud features as tokens and using a unified positional encoding and a Transformer architecture to achieve implicit alignment and fusion, thereby improving computational efficiency while maintaining high accuracy [23].

Overall, the strength of feature-level fusion lies in its ability to fully integrate multi-source data, provide rich contextual information, and effectively capture complex relationships between different modalities. However, its challenges are also significant, including high computational complexity, strong dependency on sensor poses, and the implementation complexity of designing intricate interaction networks. Future research trends are moving towards more efficient, robust, and lightweight designs.

3.3 Decision-level fusion

Decision-level fusion, also known as late fusion, is a strategy that performs integration at the backend of the information processing pipeline. Its core principle involves allowing each sensor modality, such as a camera or LiDAR, to first process data through its own independent perception network to generate high-level semantic outputs, such as bounding boxes with class confidences. Subsequently, a dedicated fusion module arbitrates and integrates these independent decisions from different sources to form a final, more reliable detection result. The notable advantage of this method lies in its high degree of flexibility and modular design, as it allows for the direct use of mature, optimized networks for each sensor without needing to handle complex upfront data alignment issues, thus effectively enhancing system fault tolerance. However, its primary shortcoming is the inability to exploit the rich, intermediate features generated during each sensor's processing, leading to a loss of deep correlations between modalities.

CLOCs is a canonical example of a decision-level fusion approach. This algorithm moves beyond simple bounding box integration by learning and applying geometric and semantic consistency constraints between 2D proposals from images and 3D proposals from point clouds to improve detection accuracy, particularly at long distances [24]. Other representative works have focused on different aspects of decision-level optimization. For instance, some approaches use a Semantic Consistency Filter (SCF) with 2D segmentation masks to filter out false positives. Shahian Jahromi et al. implement late fusion by first processing LiDAR and radar data independently to the object level, and then fusing these object states using an Extended Kalman Filter (EKF)[25]. Ren et al. address the specific challenge of fusing outputs when only two sensors are available, a scenario where traditional "majority rules" methods fail. They propose a framework based on the Evidential Reasoning (ER) rule to fuse the final classification decisions from two independent sensor models [26].

4 Discussion and outlook

A systematic review of existing works reveals a clear developmental trajectory for multi-modal fusion detection technology in autonomous driving. Currently, feature-level fusion, which performs deep interaction at the intermediate feature layers of the network, has become the definitive mainstream paradigm due to its effective balance between information fidelity and implementation complexity. Within this paradigm, unifying multi-modal features in a Bird's-Eye-View (BEV) space has emerged as the core solution for addressing the spatial alignment challenge, popularized by seminal works like BEVFusion [15]. The Transformer architecture, with its powerful cross-attention mechanism, has in turn replaced traditional concatenation or convolution as the core engine for enabling efficient and deep cross-modal feature interaction. Furthermore, the latest research trend has evolved from "how to fuse" to "how to prepare data for fusion," pioneering a new paradigm of "pre-enhancement"[27].

Notwithstanding the substantial technological progress that has been made, multi-modal fusion continues to encounter substantial challenges on the path to large-scale, all-scenario commercial deployment. The primary concern is the trade-off between performance and efficiency. Advanced fusion models, particularly those based on Transformers, are frequently computationally onerous, thereby challenging the real-time capabilities of embedded vehicle platforms [28]. Information fidelity and alignment accuracy persist as significant impediments to progress. Within the BEV paradigm, the misalignment of features resulting from depth estimation errors and the loss of details and height information during the compression of 3D data into a 2D plane continue to be primary factors hindering the enhancement of model accuracy [29]. The necessity of fortifying the resilience of systems within unfavorable environmental conditions is paramount. Despite the enhancement in robustness imparted by the fusion process, there remains an unresolved challenge concerning the reliability of the system in extreme weather scenarios, such as heavy precipitation and dense fog.

Addressing these challenges will guide the field into a new phase of development. Research will continue to focus on lightweight and adaptive fusion architectures aimed at reducing model deployment costs and enabling systems to dynamically adjust modality weights based on environmental changes. A more significant trend is the end-to-end joint optimization with downstream tasks, as the ultimate value of fusion is realized in its empowerment of the entire autonomous driving task chain.

References

[1] H. Li *et al.*, "The Effect of Rainfall and Illumination on Automotive Sensors Detection Performance," *Sustainability*, vol. 15, no. 9.

[2] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 16–21 June 2012 2012, pp. 3354–3361.

[3] H. Caesar *et al.*, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11621–11631.

[4] P. Sun *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2446–2454.

[5] M.-F. Chang *et al.*, "Argoverse: 3d tracking and forecasting with rich maps," 2019, pp. 8748–8757.

[6] X. Huang, P. Wang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, "The ApolloScape Open Dataset for Autonomous Driving and Its Application," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 42, no. 10, pp. 2702–2719, 2020.

[7] P. Xiao *et al.*, "PandaSet: Advanced Sensor Suite Dataset for Autonomous Driving," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, 19–22 Sept. 2021 2021, pp. 3095–3101.

[8] A. Patil, S. Malla, H. Gang, and Y. T. Chen, "The H3D Dataset for Full-Surround 3D Multi-Object Detection and Tracking in Crowded Urban Scenes," in *2019 International Conference on Robotics and Automation (ICRA)*, 20–24 May 2019 2019, pp. 9552–9557.

[9] J. Geyer *et al.*, "A2d2: Audi autonomous driving dataset," *arXiv preprint arXiv:*2004.06320*,* 2020.

[10] M. Sheeny, E. D. Pellegrin, S. Mukherjee, A. Ahrabian, S. Wang, and A. Wallace, "RADIATE: A Radar Dataset for Automotive Perception in Bad Weather," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 30 May–5 June 2021 2021, pp. 1–7.

[11] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," 2018, pp. 918–927.

[12] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "Pointpainting: Sequential fusion for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4604–4612.

[13] C. Wang, C. Ma, M. Zhu, and X. Yang, "Pointaugmenting: Cross-modal augmentation for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11794–11803.

[14] M. Liang, B. Yang, S. Wang, and R. Urtasun, "Deep continuous fusion for multi-sensor 3d object detection," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 641–656.

[15] Z. Liu *et al.*, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," *arXiv preprint arXiv:*2205.13542*,* 2022.

[16] H. Cai, Z. Zhang, Z. Zhou, Z. Li, W. Ding, and J. Zhao, "BEVFusion4D: Learning LiDAR-Camera Fusion Under Bird's-Eye-View via Cross-Modality Guidance and Temporal Aggregation," *arXiv e-prints,* p. arXiv:2303.17099, 2023.

[17] A. W. Harley, Z. Fang, J. Li, R. Ambrus, and K. Fragkiadaki, "Simple-BEV: What Really Matters for Multi-Sensor BEV Perception?," *2023 IEEE International Conference on Robotics and Automation (ICRA),* pp. 2759–2765, 2022.

[18] M. Chen *et al.*, "Multi-Modal BEV Enhancement Fusion for 3D Object Detection in Autonomous Driving," *IEEE Transactions on Intelligent Transportation Systems,* pp. 1–13, 2025.

[19] C. Hu *et al.*, "FusionFormer: A Multi-sensory Fusion in Bird's-Eye-View and Temporal Consistent Transformer for 3D Object Detection," 2023.

[20] X. Bai *et al.*, "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1090–1099.

[21] Z. Wang, Z. Huang, Y. Gao, N. Wang, and S. Liu, "Mv2dfusion: Leveraging modality-specific object semantics for multi-modal 3d detection," *arXiv preprint arXiv:*2408.05945*,* 2024.

[22] K. Liu, Y. Deng, J. Tong, and W. Li, "UniVoxel: A Novel Framework for 3D Object Detection in Autonomous Vehicles with Multi-modal Voxel Representation," *IEEE Sensors Journal,* pp. 1–1, 2025.

[23] J. Yan *et al.*, "Cross modal transformer: Towards fast and robust 3d object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 18268–18278.

[24] S. Pang, D. Morris, and H. Radha, "CLOCs: Camera-LiDAR Object Candidates Fusion for 3D Object Detection," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 24 Oct.–24 Jan. 2021 2020, pp. 10386–10393.

[25] B. Shahian Jahromi, T. Tulabandhula, and S. Cetin, "Real-Time Hybrid Multi-Sensor Fusion Framework for Perception in Autonomous Vehicles," *Key Laboratory of Universal Wireless Communications, Ministry of Education, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China National Engineering Lab for Mobile Networ*, vol. 19, no. 20.

[26] M. Ren, P. He, and J. Zhou, "Decision fusion of two sensors object classification based on the evidential reasoning rule," *Expert Systems with Applications,* vol. 210, p. 118620, 2022/12/30/ 2022.

[27] X. Zhang, G. Liu, M. Li, Q. Ren, H. Tang, and D. P. Bavirisetti, "FusionNGFPE: An image fusion approach driven by non-global fuzzy pre-enhancement framework," *Digital Signal Processing,* vol. 156, p. 104801, 2025/01/01/ 2025.

[28] A. B. Amjoud and M. Amrouch, "Object Detection Using Deep Learning, CNNs and Vision Transformers: A Review," *IEEE Access,* vol. 11, pp. 35479–35516, 2023.

[29] K. Bajbaa, M. Usman, S. Anwar, I. Radwan, and A. Bais, "Bird's-Eye View to Street-View: A Survey," *City University of Hong Kong, Hong Kong;Delft University of Technology, Netherlands;Nanyang Technological University, Singapore,* vol. abs/2405.08961, 2024.