Proceedings of CONF-CDS 2026 Symposium: Data-Centric AI Security: Securing Models, Learning Agents, and Autonomous Systems
DOI: 10.54254/2755-2721/2026.CH35450

Application of Multimodal Fusion Based on Sensors and Machine Vision in Autonomous Driving

Yuhang Lu

Faculty of Innovation Engineering, Macau University of Science and Technology, Macau, China
lyh060914@gmail.com

Abstract. Autonomous driving has become a transformative technology poised to reshape modern transportation systems. This paper explores multimodal fusion techniques that integrate various sensors with machine vision for autonomous driving. We examine the integration of various sensor modalities, including cameras, LiDAR, and millimeter-wave radar, alongside advanced machine vision algorithms such as YOLO, Faster R-CNN, Point Pillars, and MVX-Net for environment perception. This work addresses major challenges in sensor fusion, including data synchronization, coordinate transformation, real-time computation and conflict resolution of heterogeneous sensor data. We systematically analyze three typical fusion architectures: data-level, feature-level and decision-level fusion, and compare their performance in information retention, computational efficiency and system robustness. Through representative application cases in object detection and classification, high-precision localization and mapping, and decision-making and path planning, we demonstrate how multimodal fusion significantly enhances the robustness, accuracy, and reliability of autonomous vehicle perception systems. The paper further discusses current limitations including computational overhead, adverse-weather robustness, and lack of standardized evaluation, and outlines future directions such as end-to-end learning, 4D radar integration, and V2X-enabled cooperative perception. The results prove that reliable multimodal fusion is a core prerequisite for realizing safe and stable autonomous driving.

Keywords: Autonomous driving, Multimodal fusion, Machine vision, Sensor integration, Perception systems

1. Introduction

Autonomous driving is one of the most influential technological breakthroughs in the 21st century. Benefiting from the rapid development of artificial intelligence, sensor technology and machine learning, autonomous vehicles are gradually moving from laboratory prototypes to real-world applications. The global autonomous vehicle market is projected to experience substantial growth, driven by demands for improved traffic safety, reduced congestion, and enhanced mobility for all demographics [1]. Nevertheless, fully autonomous driving relies on high-precision perception systems to accurately perceive complex and dynamic road environments.

Single-sensor systems face inherent limitations in addressing the diverse challenges of autonomous driving environments. Each sensor modality possesses unique strengths and

DOI: 10.54254/2755-2721/2026.CH35450

weaknesses: cameras provide rich visual information but struggle in low-light conditions; LiDAR offers precise 3D measurements but is affected by adverse weather; millimeter-wave radar excels in detecting object velocity but has limited resolution [2]. Therefore, multimodal perception systems that fuse complementary sensor data have become indispensable for autonomous driving. Such integration ensures stable perception under diverse environments and further improves driving safety and overall performance.

This paper mainly discusses the principles, methodologies and typical applications of multimodal fusion combining sensors and machine vision for autonomous driving. Specifically, we aim to address three core questions: (1) how complementary sensor modalities can be effectively combined at different architectural levels to overcome the limitations of single-sensor perception; (2) which deep-learning-based fusion strategies best balance accuracy, robustness, and real-time performance; and (3) how multimodal fusion translates into measurable performance gains in downstream tasks such as object detection, localization, and path planning. This research has two-fold significance. In terms of safety, robust multimodal perception is essential to reduce traffic accidents caused by perception errors in harsh environments. Technically, the systematic analysis of fusion architectures can summarize design guidelines for next-generation perception systems.

2. Overview of autonomous driving perception systems

2.1. Basic framework of perception systems

An autonomous driving perception system consists of multiple collaborative functional layers. The foundation layer involves obtaining raw sensor data from various modalities. The preprocessing layer performs calibration, noise reduction, and data normalization. The feature extraction layer identifies relevant patterns and objects within the sensed environment. The feature extraction layer extracts target features from environmental data. The fusion layer integrates multi-sensor information to build a complete environmental model. The final interpretation layer converts fused data into effective control commands for vehicles [3]. This modular architecture enables flexible integration of diverse sensing technologies while maintaining system robustness and scalability.

2.2. Major sensors and systems

Contemporary autonomous vehicles employ a diverse sensor suite. LiDAR sensors generate high-resolution 3D point clouds of the surrounding environment, providing accurate distance measurements up to 200 meters with centimeter-level precision. Cameras collect rich visual information such as color, texture and semantic features to recognize traffic signs, lane lines and traffic lights. Millimeter-wave radar operates effectively in adverse weather conditions, measuring object velocity through Doppler shift while penetrating fog and precipitation. Ultrasonic sensors provide short-range obstacle detection for low-speed maneuvering and parking scenarios. GPS and IMU systems supply vehicle positioning and motion information. Each sensor provides unique data features for full environmental perception. Table 1 summarizes the properties of mainstream sensors and systems [4].

Table 1. Main sensors and systems in autonomous driving

Sensor/Systems	Strengths	Weaknesses
Camera	High resolution, semantic detail	Sensitive to lighting/weather; no direct depth

DOI: 10.54254/2755-2721/2026.CH35450

Table 1. (continued)

LiDAR	High-accuracy 3D geometry (dense point clouds)	Expensive; range reduced by weather; limited update rate
Millimeter-wave Radar	Long-range distance & velocity measurements; all-weather	Low spatial resolution; noisy sparse data
Ultrasonic	Low-cost short-range detection	Very short range, low resolution
GPS	Global positioning	Signal blockage; limited accuracy

2.3. Machine vision fundamental algorithms

Machine vision algorithms are used to extract effective information from image data. Object detection algorithms, which can identify and locate multiple targets simultaneously, are the core of machine vision systems. Modern deep learning architectures, particularly convolutional neural networks, have revolutionized this domain by enabling real-time detection with remarkable accuracy. The You Only Look Once (YOLO) family of algorithms exemplifies this progress, treating object detection as a single regression problem that directly predicts bounding boxes and class probabilities from full images in one evaluation. This approach achieves processing speeds exceeding 30 frames per second while maintaining high detection accuracy, making it particularly suitable for the real-time requirements of autonomous driving. Similarly, region-based methods like Faster R-CNN employ region proposal networks to generate candidate object locations, which are then refined through additional neural network stages, offering superior accuracy for complex scenarios despite slightly higher computational costs.

Beyond simple object detection, semantic segmentation algorithms provide dense pixel-wise classification of the entire visual field, distinguishing between road surfaces, vehicles, pedestrians, vegetation, buildings, and other environmental elements. This comprehensive scene understanding proves essential for path planning and navigation decisions. Instance segmentation further refines this capability by differentiating individual objects within the same category, enabling the system to track multiple vehicles or pedestrians independently. These segmentation networks typically employ encoder-decoder architectures, where the encoder progressively extracts hierarchical features while reducing spatial resolution, and the decoder reconstructs the full-resolution segmentation map through learned upsampling operations.

Depth estimation algorithms complement geometric sensors by inferring three-dimensional structure from two-dimensional images. Monocular depth estimation networks learn to predict distance information from single camera images by recognizing visual cues such as object size, texture gradients, and perspective effects. Stereo vision systems exploit the parallax between synchronized camera pairs to triangulate depth through correspondence matching, achieving more reliable geometric measurements. Optical flow analysis tracks pixel-level motion between consecutive frames, providing velocity estimates for dynamic objects and enabling motion prediction crucial for anticipating the behavior of surrounding vehicles and pedestrians. The integration of these complementary vision algorithms, increasingly powered by transformer architectures that capture long-range dependencies in visual data, has achieved performance levels approaching human perception capabilities while operating at the millisecond timescales required for safe autonomous navigation.

DOI: 10.54254/2755-2721/2026.CH35450

3. Fundamental methods of multimodal fusion

3.1. Basic concepts of multimodal fusion

Multimodal fusion combines data from multiple sensors to construct a unified environmental model, which is more accurate than the output of any single sensor. The fusion process needs to solve several core problems: temporal synchronization, spatial alignment, data association and data conflict resolution. Temporal synchronization ensures that data from different sensors captured at different rates are properly aligned in time. Spatial calibration transforms measurements from diverse sensor coordinate frames into a common reference system. Data association determines which measurements from different sensors correspond to the same physical objects. Conflict resolution mechanisms handle inconsistencies between sensor readings, often employing uncertainty quantification and reliability assessment [5].

3.2. Key technologies in vision-radar fusion

A common fusion scenario is vision-radar fusion, which integrates camera images with radar returns. In typical autonomous vehicles, high-resolution cameras are installed at the front, while millimeter-wave radars are mounted on bumpers or windshields [6]. To fuse them, one key step is geometric alignment: radar detections must be projected into the camera image plane using known calibration. Conversely, image-based proposals can be lifted into 3D space to associate radar points.

Two mainstream algorithm frameworks are adopted for vision-radar fusion: ROI-based fusion and end-to-end fusion. In ROI-based fusion, one sensor first proposes regions of interest (ROIs) and the other sensor's data is used to validate or refine those ROIs. For example, a radar might generate coarse 3D proposals that are projected into the image to extract visual features, or a vision detector's bounding box is mapped onto radar to gather range/velocity points. This approach has the advantage that if one sensor fails, the other can still propose detections. However, ROI-based methods can be limited by the initial proposal quality and may underutilize cross-modal information.

In contrast, end-to-end fusion processes both modalities jointly in a unified model. Deep neural networks simultaneously ingest image pixels and radar signals, learning correspondences without a separate ROI step. End-to-end fusion can "more comprehensively integrate features from both sensor data", leading to potentially more robust perception. Recent research trends apply such networks to fuse camera and radar in bird's-eye-view (BEV) or 3D detection frameworks.

A notable recent advancement is 4D radar: next-generation radars that provide dense point clouds with elevation data. 4D radars offer richer information that is "similar to LiDAR but with additional velocity details". Early studies show that 4D radar can help overcome limitations of traditional radar, although effectively aligning 4D radar data with camera imagery remains challenging. Overall, vision–radar fusion technology relies on accurate sensor calibration, data alignment, and clever network architectures to merge the distinct sensor inputs [7].

3.3. Deep learning-based fusion methods

Deep learning has enabled more sophisticated fusion strategies. In deep fusion networks, multimodal data can be merged at various layers. In data-level fusion, raw sensor inputs are stacked and fed into a network. In feature-level fusion, each modality is first processed through separate subnetworks and the resulting feature maps are merged midway through the network. Finally, decision-level fusion ensembles the outputs of separate pipelines. Many state-of-the-art approaches

DOI: 10.54254/2755-2721/2026.CH35450

perform a hybrid of these, learning optimal fusion weights or using transformers to let the model decide how to integrate modalities [8].

Architectures often employ CNNs for image processing and tailor-made layers for radar. Recurrent networks (RNNs) or temporal models may fuse sequences of data for dynamic scenes. Some methods use gating or attention mechanisms to weigh each sensor's contribution depending on context. In practice, popular deep fusion models include variant fusion branches in detectors like Faster R-CNN, YOLO, and SSD adapted for multi-sensory input. CNNs and RNNs remain the predominant building blocks in these fusion networks. Research shows deep fusion often outperforms classical Kalman-filter-based approaches when ample data are available. In summary, deep learning enables fusion by jointly learning cross-modal feature representations, making the resulting perception model more powerful than modality-specific approaches.

4. Application cases of multimodal fusion in autonomous driving

4.1. Object detection and classification

Multimodal fusion greatly improves the object detection and classification performance of autonomous driving systems. The integration of camera and LiDAR data exemplifies this synergy, where cameras provide rich semantic information, including color, texture, and appearance features, while LiDAR contributes precise three-dimensional spatial measurements and geometric structure. Modern deep learning architectures have been specifically designed to exploit these complementary strengths. For instance, PointPillars converts sparse LiDAR point clouds into structured pillar representations that can be efficiently processed alongside camera image features, while MVX-Net employs sophisticated attention mechanisms to dynamically weight and fuse multi-sensor features based on their reliability and relevance in different scenarios [9, 10].

This fusion approach proves particularly valuable under challenging conditions where single sensors fail. In low-light environments or nighttime driving, cameras struggle to capture sufficient visual detail, but LiDAR maintains consistent performance by actively emitting laser pulses. Conversely, in adverse weather conditions, such as heavy rain or fog, where LiDAR returns become noisy and unreliable, camera data combined with millimeter-wave radar provides essential continuity. The radar component adds another critical dimension by measuring object velocity through Doppler shift analysis, enabling the system to distinguish between stationary obstacles and moving vehicles, predict collision trajectories, and track fast-moving objects with high temporal resolution. Experimental validation across diverse datasets demonstrates that these multimodal fusion systems consistently outperform single sensor approaches by 15%-25% in detection accuracy, with the most significant improvements observed for distant objects beyond 50 meters, partially occluded targets where only portions are visible, and small objects such as pedestrians and cyclists that present minimal LiDAR returns [11].

4.2. High-precision localization and mapping

Accurate vehicle localization and environmental mapping are fundamental requirements for autonomous driving, demanding centimeter-level positioning accuracy across diverse operational scenarios. The solution lies in fusing complementary sensor data through Simultaneous Localization and Mapping (SLAM) algorithms that simultaneously build environmental maps and determine vehicle position within them.

DOI: 10.54254/2755-2721/2026.CH35450

LiDAR serves as the primary mapping sensor, generating detailed three-dimensional point clouds that are matched against pre-built high-definition maps to achieve precise localization. This matching process, combined with scan-to-scan alignment, provides robust position estimates with centimeter-level accuracy. However, LiDAR-only approaches accumulate drift errors over long distances and struggle in environments lacking distinctive geometric features, such as highways or tunnels. Camera-based visual odometry addresses these limitations by tracking distinctive visual features across frames, providing independent motion estimates that complement LiDAR measurements. The integration of IMU data further enhances robustness by measuring vehicle acceleration and rotation at high frequencies, maintaining accurate state estimates during rapid maneuvers or when other sensors temporarily fail.

GPS integration provides absolute position references that prevent unbounded error accumulation, though its reliability varies significantly with environment. In urban canyons and tunnels where GPS signals degrade or disappear, the system relies on LiDAR-camera-IMU fusion. Modern implementations employ Kalman filtering or graph optimization to fuse these heterogeneous measurements optimally, weighting each sensor according to its reliability in the current context. This multi-sensor redundancy ensures continuous localization capability across all driving scenarios, automatically adapting to sensor availability while maintaining the precision required for safe autonomous navigation [12].

4.3. Decision making and path planning

The culmination of multimodal perception lies in its support for intelligent decision-making and path-planning systems that must operate safely in complex, dynamic environments populated by unpredictable human drivers and pedestrians. Fused sensor data provides a comprehensive environmental representation that captures not only the current state but also the semantic context essential for understanding scene dynamics. The integration of geometric information from LiDAR with semantic labels from camera-based vision algorithms enables the system to distinguish between fundamentally different scenarios that might appear geometrically similar: a parked vehicle versus one with brake lights illuminated preparing to merge, a pedestrian standing at a crosswalk versus one actively stepping into the roadway, or construction barriers marking a temporary lane closure versus permanent road infrastructure.

This rich multimodal representation feeds into prediction modules that forecast future trajectories of surrounding dynamic agents. By analyzing historical motion patterns captured through optical flow and object tracking, combined with contextual cues such as traffic signal states, turn indicators, and pedestrian gaze direction detected through vision algorithms, the system generates probabilistic predictions of likely future behaviors. These predictions account for multiple possible futures, recognizing the inherent uncertainty in human behavior while identifying high-risk scenarios requiring defensive responses. The velocity measurements from radar prove particularly valuable here, providing direct observation of object speeds without relying on frame-to-frame differentiation that introduces noise and latency.

Path planning algorithms utilize this comprehensive understanding to generate trajectories that satisfy multiple competing objectives: safety constraints that maintain sufficient clearance from obstacles and respect kinematic limits of the vehicle, comfort considerations that minimize harsh accelerations and ensure smooth steering, efficiency goals that optimize travel time and energy consumption, and legal compliance that adheres to traffic rules and road geometry. The real-time nature of multimodal fusion enables reactive planning that continuously adapts to evolving situations, such as vehicles suddenly changing lanes, pedestrians entering crosswalks, or emergency

DOI: 10.54254/2755-2721/2026.CH35450

vehicles approaching with active sirens. This tight integration between perception and planning, made possible through reliable multimodal sensor fusion, represents the essential foundation for navigating the complexity of real-world autonomous driving in both structured highway environments and chaotic urban settings [13].

5. Challenges and future development

Despite remarkable progress, multimodal fusion for autonomous driving still faces multiple challenges. The real-time processing of massive sensor data requires high computational resources, particularly for deep learning-based fusion approaches. Ensuring robustness across diverse environmental conditions, including extreme weather and unusual scenarios rarely encountered during training, remains difficult. Sensor degradation and failure modes require fault detection and degradation strategies. The lack of standardized fusion architectures and evaluation metrics complicates cross-platform comparison and validation.

Future development directions include advancement of end-to-end learning approaches that jointly optimize sensing, fusion, and control. Edge computing and specialized hardware accelerators will enable more sophisticated fusion algorithms to execute in real time. Improved uncertainty quantification methods will better characterize fusion reliability for safety-critical decisions. Vehicle-to-everything (V2X) communication will introduce additional data sources requiring novel fusion paradigms. Simulation-based testing and synthetic data generation will accelerate algorithm development and validation. As these technologies mature, multimodal fusion will play an increasingly central role in realizing safe, reliable, and widely deployable autonomous driving systems [14].

6. Conclusion

This paper has examined the critical role of multimodal fusion combining sensors and machine vision in autonomous driving systems. We have demonstrated that effective integration of diverse sensor modalities including LiDAR, cameras, radar, and GPS provides comprehensive environmental perception capabilities exceeding those of any individual sensor. The analysis of fusion methodologies across different architectural levels reveals trade-offs between information preservation, computational efficiency, and system modularity. Application case studies in object detection, localization, and path planning illustrate the substantial performance improvements achieved through multimodal fusion, with accuracy gains of 15%-25% compared to single-sensor approaches.

This work mainly discusses traditional sensor combinations and fails to fully explore emerging technologies such as event cameras, solid-state LiDAR and V2X communication. Future research should investigate end-to-end learning frameworks that optimize entire perception-planning-control pipelines, advanced uncertainty quantification for safety assurance, and fusion algorithms robust to sensor failures and adversarial conditions. The integration of V2X data streams presents opportunities for cooperative perception extending beyond individual vehicle sensor ranges.

As autonomous driving technology progresses toward widespread deployment, multimodal fusion will remain a cornerstone capability enabling safe, reliable, and efficient autonomous vehicles. Continued advancement in sensor technology, computing hardware, and fusion algorithms will drive the evolution toward fully autonomous transportation systems, ultimately transforming mobility and improving road safety for all users.

DOI: 10.54254/2755-2721/2026.CH35450

References

[1] Badue, C., et al. (2021). Self-driving cars: A survey. *Expert Systems With Applications, 165*, 113816.

[2] Yeong, J., et al. (2021). Sensor and sensor fusion technology in autonomous vehicles: A review. *Sensors (Basel), 21*(6).

[3] Fayyad, J., et al. (2020). Deep learning sensor fusion for autonomous vehicle perception and localization: A review. *Sensors, 20*(15), 4220.

[4] Li, Y., & Ibanez-Guzman, J. (2020). Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems. *IEEE Signal Processing Magazine, 37*(4), 50–61.

[5] Feng, D., et al. (2020). Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems, 22*(3), 1341–1360.

[6] Yao, S., et al. (2023). Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review. *IEEE Transactions on Intelligent Vehicles, 9*(1), 2094–2128.

[7] Wang, L., et al. (2022). Multi-modal and multi-scale fusion 3D object detection of 4D radar and LiDAR for autonomous driving. *IEEE Transactions on Vehicular Technology, 72*(5), 5628–5641.

[8] Zhang, X., et al. (2023). Multi-modal fusion technology based on vehicle information: A survey. *IEEE Transactions on Intelligent Vehicles, 8*(6), 3605–3619.

[9] Lang, A. H., et al. (2019). Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

[10] Sindagi, V. A., Zhou, Y., & Tuzel, O. (2019). Mvx-net: Multimodal voxelnet for 3D object detection. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE.

[11] Caesar, H., et al. (2020). nuScenes: A multimodal dataset for autonomous driving. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)*, 11618–11628.

[12] Cadena, C., et al. (2017). Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics, 32*(6), 1309–1332.

[13] Schwarting, W., Alonso-Mora, J., & Rus, D. (2018). Planning and decision-making for autonomous vehicles. *Annual Review of Control, Robotics, and Autonomous Systems, 1*(1), 187–210.

[14] Rosique, F., et al. (2019). A systematic review of perception system and simulators for autonomous vehicles research. *Sensors (Basel), 19*(3).