



Multi-Sensor Fusion Technology for Intelligent Driving

Jiajun Sun^{1*}

¹School of Mechanical and Power Engineering, East China University of Science and Technology, Shanghai, 200333, China

*23011323@mail.ecust.edu.cn

Abstract. Amidst the innovation of science and technology and the demand for new forms of transportation, intelligent driving has become a critical topic in modern automotive industry. Stable and efficient environmental recognition and monitoring serve as essential foundations for realizing intelligent driving. However, traditional single-sensor monitoring struggles to meet the basic operational requirements of intelligent driving systems when confronted with complex real-world driving environments and road conditions. Due to the detection limitations of a single sensor, full-scenario coverage for intelligent driving remains unattainable. Consequently, multi-sensor data fusion technology has been progressively developed and applied in intelligent driving. This review provides a comprehensive analysis of sensor technologies applied in intelligent driving. It details mainstream sensor technologies (including visual sensors, LiDAR, and millimeter-wave radar), analyses the multi-sensor fusion algorithm frameworks, and demonstrates specific 3D object detection examples through multi-sensor fusion algorithms categorized by fusion level. Additionally, the paper analyzes current technical challenges and future research directions in intelligent driving, concluding with a concise summary of this field.

Keywords: Intelligent Driving, Multi-Sensor Fusion Technology, Sensor

1 Introduction

1.1 Research Background and Significance

Development Status and Core Requirements of Intelligent Driving. Autonomous driving technology, as a profound integration of artificial intelligence and automotive engineering, is undergoing a transformative evolution from L2 partial automation to L4 high-level autonomous driving. Concurrently, intelligent driving systems are experiencing accelerated developmental phases. This paradigm shift has prompted an increasing number of automotive manufacturers to reorient their strategic priorities from conventional automotive technological innovations toward the research and development of intelligent driving systems.

Within operational contexts, intelligent driving systems necessitate three critical capabilities: real-time precise distance estimation, comprehensive road scenario recognition across diverse operational conditions, and instantaneous response

execution. However, current technological frameworks confront persistent challenges, including perceptual blind spots in complex traffic environments, misjudgment of dynamic objects, and insufficient adaptability to extreme environmental conditions. Specifically, the core demands concentrate on:

- Full-time perception: Continuous detection under extreme conditions (e.g., low-light, rain, fog) and correct operation under low light conditions in unstructured scenes.
- High-confidence decision-making: Millimeter-level localization accuracy and millisecond-level temporal consistency.
- Functional safety redundancy: Compliance with international safety standards for sensor redundancy.

Necessity of Multi-Sensor Fusion. Multi-Sensor Fusion (MSF) effectively improves the robustness and fault-tolerance of sensing systems through information complementarity and redundant verification mechanisms. Its technical necessity is reflected in three levels:

- Spatial: Cross-modal registration (e.g., LiDAR-camera fusion) improves 3D scene reconstruction, ensuring the spatial perception capabilities of intelligent driving systems.
- Temporal: Applying high refresh rate and high sampling rate sensor devices such as millimeter wave radar ensures that the intelligent driving system has real-time information about the road surface.
- Informational: In extreme environments or emergency scenarios, the provision of multi-dimensional information sources for intelligent driving systems ensures information redundancy, thereby preventing the systems from being constrained by inherent physical limitations of sensors that could impede timely and effective feedback.

Sensors constitute the fundamental mechanism through which intelligent driving systems perceive operational environments, with each sensor category possessing distinct advantages and inherent limitations. Single-sensor configurations exhibit physical constraints that introduce intrinsic deficiencies: visible-light cameras demonstrate vulnerability to illumination variations, LiDAR systems suffer significant performance degradation in rainy or foggy conditions, and millimeter-wave radars exhibit insufficient spatial resolution.

The objective of multi-sensor data fusion technology resides in harnessing redundant and complementary information provided by heterogeneous sensor arrays. By exploiting the spatiotemporal complementary properties of multi-sensor configurations, this methodology effectively overcomes the perceptual limitations of unimodal sensing paradigms. Such integration substantially reduces uncertainties and ambiguities in observational data while concurrently enhancing the reliability and operational robustness of perception systems.

The Critical Role of Multi-Sensor Fusion. In intelligent driving systems, sensor fusion technology delivers core value through a three-tier optimization mechanism:

- Perception Accuracy Enhancement: Sub-pixel-level object contour reconstruction is achieved through spatiotemporal registration and feature-level fusion, thereby enhancing information depth for precise characterization of object features.
- Decision Reliability Augmentation: Cross-validation of multi-source data effectively suppresses false detection rates while neutralizing uncertainties in observational data, thereby improving the reliability of system judgments.
- Safety Assurance: Redundant perception architectures establish fail-safe barriers, ensuring stable system operation in complex and extreme driving environments.

1.2 Review Objectives and Structure

Today, fundamental autonomous driving technologies have achieved widespread adoption across the automotive industry, while continuous advancements are being made toward realizing fully intelligent driving systems. Automotive manufacturers have begun implementing autonomous driving functionalities—including automated steering, acceleration, and braking—under active driver supervision through Advanced Driving Assistance Systems (ADAS). These capabilities are being continuously refined, with new functionalities progressively emerging. Multi-sensor fusion technology stands as a pivotal technology in the development of future intelligent driving systems. This paper presents a systematic review of multi-sensor fusion technologies in intelligent driving, elucidating current technical solutions, primary challenges, and emerging trends.

This work methodically organizes the theoretical frameworks and practical advancements in multi-sensor fusion technologies. Below is a concise overview of the core contents:

- Section II comprehensively analyzes core sensor technologies in intelligent driving, detailing their operational principles and technical characteristics.
- Section III introduces the architectural framework of multi-sensor fusion algorithms.
- Section IV demonstrates practical implementations of multi-sensor fusion algorithms through 3D object detection case studies.
- Section V discusses current challenges and proposes future research directions in this domain.

2 Classification of sensors

Enter methodology section here. Autonomous driving systems and advanced driver assistance technologies have now achieved widespread implementation, with sensor modalities such as vision sensors, LiDAR and radar serving as their core operational components [1]. For next-generation intelligent driving systems, these sensors remain the indispensable foundation for environmental perception, enabling critical functionalities in situational awareness and decision-making processes.

2.1 Vision Sensors

Cameras are the most widely used visual sensors and hold the predominant position in intelligent driving applications. Cameras acquire image information through optical imaging principles to perceive the surrounding environment, serving as low-cost passive sensors with excellent imaging performance [2]. For image information captured by cameras, digital image processing methods are utilized for object feature recognition. The advantage of image information lies in its ability to determine the spatial position and category of target objects more efficiently. By employing trained image recognition models, reliable identification of conventional road scenarios such as vehicles, pedestrians, and non-motorized vehicles can be achieved [3].

The most critical information source for drivers during operation relies on visual perception, and cameras largely simulate the real-world image data perceived by human vision. The operational principle of cameras involves image sensors receiving optical projections of physical entities generated by lenses, which are subsequently converted into corresponding digital signals. Simultaneously, machine vision algorithm techniques process these signals to extract essential target parameters, including object type, positional coordinates, and velocity information [4].

However, cameras exhibit significant limitations in practical applications: First, their imaging performance deteriorates in low-light environments. Although certain algorithms can approximate normal illumination imaging quality under such conditions, this approach unavoidably increases exposure time, thereby failing to meet the fundamental sampling rate requirements for sensor data acquisition in intelligent driving systems. Second, cameras demonstrate constrained depth perception capabilities. Depth prediction through 2D image processing via trained neural networks typically introduces substantial uncertainties and low accuracy [2]. Additionally, cameras remain highly sensitive to weather conditions, with imaging performance under low-visibility scenarios such as rain and fog significantly inferior to that in clear weather [4].

2.2 LiDAR

LiDAR measures distance by calculating the round-trip delay of laser signals transmitted to targets [5], as illustrated in Figure 1. A LiDAR sensor comprises two primary components: a transmitter and a receiver. The transmitter incorporates a laser emitter generating light with wavelengths between 250-1600 nm, while the receiver executes signal collection, analysis, and computation. The receiver typically consists of three subsystems: a telescope for photon collection, an optical analyzer converting optical signals into electrical signals, and a data acquisition module responsible for pulse timing calculations and information storage [6].

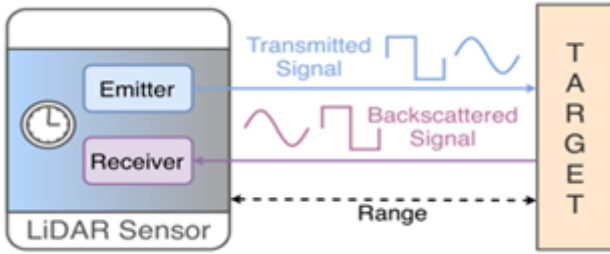


Fig. 1. Operation principle of LiDAR technology [6].

In LiDAR ranging, the target distance can be derived using the time of flight (ToF) of the laser pulse, which is calculated by modulating parameters such as the frequency, phase, or intensity of the transmitted light and measuring the temporal interval at which it is detected by the receiver [6].

However, LiDAR systems face notable limitations: Their high deployment costs and the short emission wavelengths of laser beams render them susceptible to interference from airborne particulates. Consequently, LiDAR alone cannot independently fulfill the operational requirements for vehicle autonomy [2].

2.3 Millimeter-Wave Radar

Radar stands as one of the most technologically mature and widely implemented sensors, characterized by operational stability across most environmental conditions. However, conventional radar systems exhibit limited resolution and accuracy, incapable of directly capturing object contours or detecting small-scale targets [2].

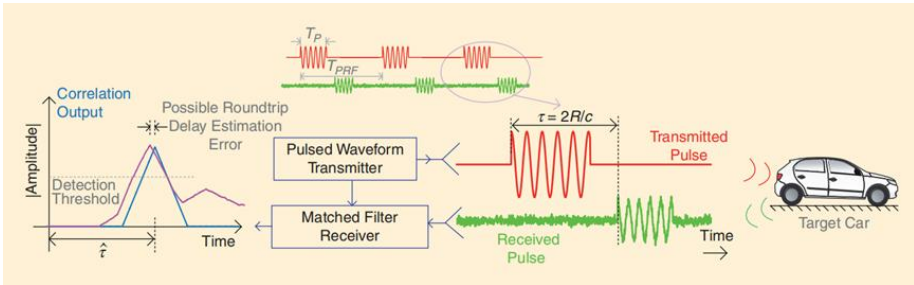


Fig. 2. Schematic of pulsed continuous wave radar ranging of target car [1].

Millimeter-wave radar, an active sensor, shares operational principles with LiDAR by employing time-of-flight (ToF) calculations for target detection. Its primary distinction lies in the utilization of longer-wavelength radio waves, enabling extended detection ranges. The waveform of electromagnetic signals significantly impacts ToF delay estimation in distance measurement. As shown in Figure 2, pulse-modulated continuous waves (CW) consist of short-duration power pulses separated by silent intervals, which allow for reflected signals reception and serve as temporal markers for ranging calculations, unlike unmodulated continuous waves. Furthermore, reflected

signals from targets must arrive before the initiation of subsequent pulses. Consequently, the maximum detectable range is constrained by the pulse repetition interval. The transmitted signals experience attenuation from path loss and imperfect target reflections before returning to the receiver. Additionally, received target signals are contaminated by internal electronic noise and external interference, which may originate from non-target object reflections or artificial sources [1].

3 Multi-Sensor Fusion Algorithm Framework

3.1 Technical Background

Multi-sensor data fusion is a technology that combines data from multiple sensors and relevant databases to achieve higher accuracy and specific inferences. It is used by humans and animals to assess their surroundings, identify threats, and enhance their survival prospects by combining multiple sensory modalities, which has evolved to be utilized concurrently [7].

While the concept of data fusion is not new, the emergence of novel sensors, advanced processing techniques, and improved computational hardware has driven the development of real-time data fusion. Artificial intelligence technologies have also injected momentum into this field. Data fusion techniques originate from multiple traditional disciplines, including digital signal processing, statistical estimation, control theory, artificial intelligence, and classical numerical methods [7].

Data fusion can be categorized into feature-level fusion and decision-level fusion. In feature-level fusion, representative features are extracted from sensor data, creating a single feature vector. This vector is used in pattern recognition techniques like neural networks or clustering algorithms. Decision-level fusion involves integrating sensor data after each sensor has determined entity position, characteristics, and identity [7].

3.2 ARCHITECTURE OF MFI

The MFI architecture is classified based on the types of processed input data and the resultant information provided to the system, encompassing low-level incoming signal estimation, intermediate-level extracted feature classification, and high-level decision-making processes involving symbolic and sub-decision outputs [8]. Its general structure is shown in Figure 3. The fusion functionalities across these levels and their corresponding fusion methodologies are summarized as follows:

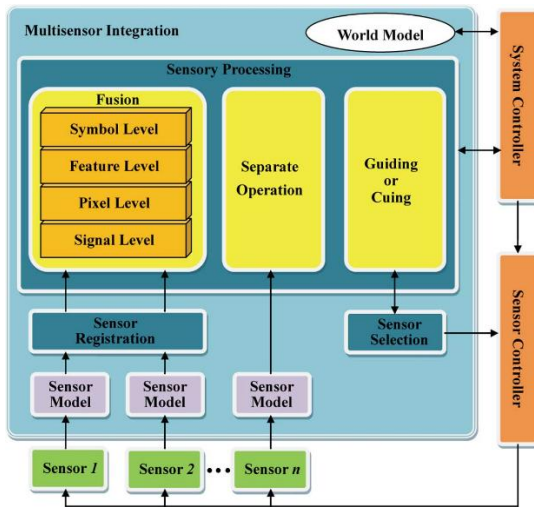


Fig. 3. Schematic diagram of MFI structure [8].

Signal and Pixel Level Fusion. Sensory data at the signal level is a result of device measurements, and due to heterogeneous sampling characteristics, multi-sensory data requires synchronization and adaptation. Statistical estimation methods, categorized into non-recursive and recursive methods, have been successfully applied for data fusion. Non-recursive estimators like weighted averaging and least squares estimation are used for redundant data merging, while recursive methods like Kalman filters and extended Kalman filters can be applied for broader fusion objectives [8]. Pixel-level fusion involves processing raw pixel information from multiple sensors through image synthesis [9].

Feature Level Fusion. Fusion at this level operates on features extracted from signals and images. Enhanced feature distinctiveness is achieved by concatenating feature points acquired from disparate sources [10]. A typical feature-level fusion process comprises three sequential steps: feature set unification and normalization, feature dimensionality reduction and concatenation, followed by feature matching. Data classification techniques prove particularly suitable for feature fusion [11,12].

Symbol Level Fusion. Human-interpretable descriptors and other symbolic representations of process parameters make up the data processed at this level. The integration of symbols with corresponding uncertainty metrics into a comprehensive decision is known as high-level fusion. Another name for symbolic fusion is decision-level fusion. Algorithms that support approximation, uncertainty, partial truth, and imprecision are ideally suited for symbolic-level fusion. Fuzzy logic systems, neural networks, genetic algorithms, and evolutionary algorithms are examples of common

implementations. Furthermore, inferential techniques like Dempster-Shafer (D-S) theory and Bayesian inference have been effectively used for symbolic-level fusion [8].

4 Case Studies: 3D Object Detection via Multi-Sensor Fusion

4.1 Feature-Level Fusion

Feature-level fusion involves extracting distinctive features from the raw observation data furnished by each sensor and subsequently integrating these features into a unified feature vector. [2]. The subsequent section briefly introduces selected data processing methods employed in feature-level fusion.

Fusion Input Representation. (1) LiDAR Points & image feature map: PointNet makes it possible to extract features directly from point clouds, and solves the problem that point clouds cannot be directly convolved due to their irregularity [13]. However, the laser radar still has the problem that the detection effect for remote and small objects is not ideal due to the uneven distribution of radar points. A common solution is to integrate LiDAR point clouds with camera images to achieve enhanced 3D object detection [2]. Here are some examples of different solutions:

PointFusion is a multi-sensor fusion framework that processes point clouds by combining spatial information from raw point clouds with texture data from images. It uses ResNet to extract image features, fully exploiting RGB information [14, 15].

F-PointNet uses mature 2D object detection networks to identify 2D bounding boxes in images, and through camera projection transformations, it establishes 3D frustums that constrain relevant point cloud data [16].

A dynamic cross-attention module (DCAN) compensates for limitations in conventional fusion methods and improves LiDAR-camera data alignment accuracy [17].

EPNet enhances point cloud features by directly integrating image features without relying on image annotations, addressing two critical challenges in 3D object detection: multi-sensor data fusion and resolving inconsistencies in classification confidence [18].

FusionRCNN generates preliminary 3D proposals using the SECOND network and creates regions of interest (RoIs). It enhances features within RoIs through point cloud processing, extracts multi-view image features via ResNet and FPN, and uses a self-attention mechanism to strengthen domain-specific features. FusionRCNN demonstrates operational capability without cross-attention modules, indicating its dual-branch architecture can independently perform single-sensor 3D detection [19]-[20].

(2) LiDAR Points & image mask: Compared to 2D object detection, which usually only attains pixel-level localization accuracy, image segmentation allows for more accurate object identification. Foreground and background elements are frequently present in conventional 2D detection frames, which can result in imprecise image feature extraction and challenges with 3D feature fusion. Semantic segmentation, on the other hand, accomplishes per-pixel object delineation, allowing for precise

correspondence with LiDAR points and more accurate semantic information for fusion [2].

PI-RCNN is a camera image feature-based framework consisting of a segmentation network and a detection network. It uses a PACF module to interconnect these subnetworks, combining multi-modal features for 3D prediction generation [21].

Point Painting enhances LiDAR point features by combining 2D semantic segmentation information with point clouds through projection matrices. This method allows for direct correspondence between point clouds and pixel data for identical objects, capturing fine-grained structural details and achieving pixel-level partitioning [22].

The superiority of semantic segmentation over conventional 2D object detection lies in its capacity to precisely separate foreground targets from background interference, thereby minimizing background impacts on fused pixel and point cloud data [2].

(3) Point clouds view & image feature map: TransFusion's fusion framework incorporates attention mechanisms. The first decoder layer uses the bird's-eye view (BEV) map to predict initial bounding boxes after converting LiDAR point clouds into a BEV representation. The second decoder then links object queries to image features to produce final predictions [23].

Diverging from mainstream approaches, BEVFusion primarily focuses on camera sensors with LiDAR serving as supplementary inputs. This framework transforms multi-view camera images into 3D representations subsequently converted to BEV maps, while concurrently projecting LiDAR points into BEV space. A fusion module combines camera-derived and LiDAR-derived BEV representations, achieving implementation simplicity through their dimensional consistency. Notably, BEVFusion represents the first framework capable of addressing LiDAR sensor failures or insufficient LiDAR data due to its camera-centric architecture [24].

Wang et al. implement adaptive fusion by integrating three data representations: LiDAR BEV maps, LiDAR range-view point clouds, and 2D camera images. Their designed Point-Attention Fusion (PAF) module employs attention mechanisms to dynamically weigh the importance of each data source during feature combination [25].

(4) Point clouds voxels & image feature map: VoxelNet is a method that transforms point cloud data into ordered high-dimensional feature representations by sampling them into sparse voxels. This approach has inspired numerous studies on point cloud voxelization [26].

Li et al. proposed a new approach that partitions point cloud space into voxels, using a sampler to select key image regions and mapping them onto light rays [27].

AutoAlignV2 is a 3D object detector framework that efficiently aggregates image features using a deformable cross-attention network. It also has an image-level dropout training strategy that improves detection accuracy [28].

MVX-Net establishes correlations between point clouds and pixels by projecting voxelized features onto image planes [29].

The MSMDFusion framework uses multi-granularity progression of multi-scale LiDAR and phase features, sampling point clouds and multi-view images into voxel representations, and projecting the resulting features into BEV space [30].

Fusion Granularity. The most straightforward multi-sensor fusion approach involves fusing the minimal data units from each sensor, yet this method demands substantial computational resources. To balance detection accuracy with inference speed, fusion operations can be implemented across entire regions or localized areas depending on network architectures and parameter configurations [2]. Common fusion granularities include Region of Interest (RoI), Voxel, and Point, which are elaborated below:

(1) RoI-Wise: RoI partitioning represents a conventional operation in image processing, functionally analogous to attention mechanisms. Through algorithmic methods that divide distinct regions into RoIs, computational resources are concentrated on these critical areas, thereby enhancing learning efficiency while reducing computational overhead. In multi-modal fusion-based 3D object detection, a common practice involves leveraging 2D image-based object detection results to define RoIs, which are subsequently projected into 3D space to obtain corresponding 3D regions. These spatially constrained 3D regions are then processed by specialized 3D detectors for refined object localization and characterization [31].

(2) Voxel-Wise: In some situations, ROI-based methods are not the best for detecting small objects due to their excessive perceptual scope. Since voxels can roughly depict object geometries in 3D space, voxel-based techniques outperform ROI-based techniques in terms of accurately aligning 3D objects with 2D imagery and successfully distinguishing between foreground and background elements. Sparse LiDAR points can be regarded as empty voxels since voxels are essentially downsampled representations of point clouds. This feature allows for the complete use of available data while eliminating irrelevant information, giving voxel-based fusion techniques more accuracy than their ROI-based counterparts [2].

(3) Point-Wise: Point-wise fusion typically enhances the characteristics of LiDAR points through feature augmentation. A conventional approach involves calculating distances from individual LiDAR points to predefined geometric references within fixed-size bounding boxes, such as box centers and corners, as illustrated in Figure 4a. Alternatively, Xie et al. employed distances between LiDAR points and their k -nearest neighboring points for feature enrichment, as depicted in Figure 4b [21].

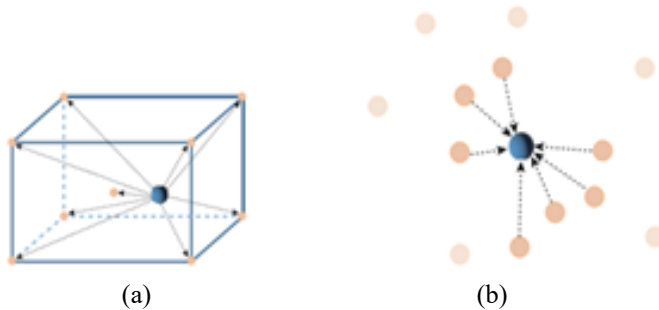


Fig. 4. Schematic diagram of Point-wise fusion enhancement [2]. (a) Enhance features using point-to-box distance. (b) Enhance features using the distance from a point to k -neighbor points.

Beyond the aforementioned point-wise feature enhancement methods, laser radar point characteristics can be further augmented by associating image features with

corresponding LiDAR points. This approach fully exploits the rich textual information inherent in visual data. Compared with RoI-level and voxel-based methods, point-wise techniques demonstrate superior performance improvements [22]. However, such enhancements come at the expense of increased memory consumption. Particularly when implemented without sampling, point-level fusion necessitates feature augmentation for most of the LiDAR points, thereby imposing significant computational resource demands [2].

4.2 Feature-Level Fusion

Decision-level fusion uses single-sensor detection networks to process distinct data streams for object detection. A designed fusion module refines individual sensor predictions, generating enhanced accuracy. Compared to feature-level fusion, it benefits from modular design, allowing straightforward evaluation through interchangeable detection heads. This methodology eliminates direct pixel-to-point correspondence handling while maintaining lower computational complexity [2].

CLOCs is a multi-sensor fusion framework of low-complexity that uses geometric consistency to generate refined detection results [32]. MV3D is a multi-sensor fusion framework that uses LiDAR point clouds and camera images as inputs. It generates 3D proposals from BEV LiDAR data, refines them using features from alternative views, and fuses them across three perspectives for final prediction [33]. AVOD is a hybrid approach that combines feature-level and decision-level fusion, extracting BEV LiDAR and camera image features separately [34].

The incapacity of decision-level fusion to utilize rich intermediate features is a significant drawback, as the performance of specific single-sensor detection networks limits the detection accuracy [2].

Most object detection algorithms are developed on open-source frameworks, making it easy for researchers to use pre-existing detection networks for experimental validation. This is particularly beneficial for research on decision-level fusion, which can be used as a standalone solution or an adjunct to feature-level fusion. This method reduces operational failures caused by single-sensor malfunctions and improves system robustness by combining outputs from multiple models and sensors [2].

5 Open Challenges and Future

5.1 Challenges

Data Alignment. Each sensor possesses unique acquisition perspectives or information dimensions. Due to vehicle engineering constraints or practical requirements, onboard sensors are typically installed at distinct positions on the vehicle body, resulting in varying viewpoint discrepancies. Simultaneously, inherent modality differences exist across sensors—for instance, cameras capture environmental data through optical imaging while LiDAR perceives surroundings via 3D point clouds. These heterogeneous data representations create significant challenges in post-acquisition sensor data alignment [2].

A prevalent solution involves rigorous calibration based on each sensor's physical placement, device parameters, and detection ranges to establish point-to-point correspondence across modalities. This constitutes a rigorous and time-intensive task with elevated error rates, while simultaneously escalating computational demands during real-world system operation. Furthermore, minor sensor displacements caused by vehicular vibrations during daily operation or natural device aging inevitably degrade calibration accuracy, introducing systematic errors into the perception framework [2].

Lost Information. Data loss typically arises from sensor discrepancies, processing limitations, and algorithmic flaws. Precision variations across sensors and computational load reduction strategies during data processing often result in missing pixels, where the absence of critical data points significantly impacts feature determination [2].

Sensor misjudgments in extreme scenarios can also induce catastrophic data loss:

- Case 1 (May 2016): A Tesla Model S operating in autonomous mode collided with a turning semi-trailer truck on a Florida highway, resulting in a fatal crash. The vehicle was equipped with a Mobileye EyeQ3 vision system (central windshield-mounted camera), a front-bumper-mounted millimeter-wave radar, and 12 ultrasonic sensors. During the incident, the camera's view was obstructed by intense ambient light, failing to detect the white truck. Concurrently, the radar's low mounting position and the truck's elevated chassis exceeded its detection range. Ultrasonic sensors proved ineffective due to their limited measurement range at highway speeds [35].
- Case 2 (June 2020): Another Tesla vehicle with Autopilot engaged crashed into a white truck. The vehicle was equipped with eight cameras and twelve millimeter-wave radars distributed around its chassis. While the cameras were designated for object recognition, the radars primarily functioned for velocity measurement and tracking of preceding vehicles, demonstrating limited effectiveness in identifying complex static objects. Within the sensor fusion framework, radar-derived velocity measurements were considered valid only when the cameras confirmed the presence of a vehicle ahead. This operational logic stemmed from the camera's dependence on ambient lighting conditions and physical color characteristics for obstacle detection. In this incident, the truck's coloration exhibited high similarity to the surrounding environment. Additionally, intense ambient light interference contributed to erroneous camera judgments, resulting in a false perception of unobstructed forward space. Another contributing factor potentially involved limitations in the training dataset for the camera's visual algorithm. The deployed deep learning models likely failed to accurately classify the truck's container roof due to insufficient representation in training data, ultimately causing perception failure [35].

These incidents demonstrate that partial sensor data loss can induce critical perception failures in autonomous systems, leading to severe operational consequences.

5.2 Research Trends and Future

Deep learning has shown promise in autonomous driving applications by achieving success in object detection. Neural network operations are facilitated by transforming sparse point clouds into ordered voxel representations. Techniques like sparse convolution and voxel indexing help efficiently identify and process non-critical regions, reducing computational complexity and speeding up network inference. Neighboring points in point clouds usually reside within the same or nearby voxels, maintaining spatial relationships and creating contextual information. Voxels facilitate advanced processing and application of point cloud data by providing standardized representations, streamlined data processing, and spatial relationship management [2].

Additionally, image data is important for multi-sensor fusion, with depth prediction enhancing safety and reliability in intelligent driving technologies by mapping features to 3D space and pixel-level depth estimation [2].

Moreover, the tri-modal fusion of millimeter-wave radar, LiDAR, and cameras remains underexplored. Radars exhibit advantages including rapid response, cost-effectiveness, and weather resilience, positioning them as reliable emergency sensors under extreme conditions [2].

6 Conclusion

Over the past three decades, autonomous driving technology has achieved substantial progress, with its rapid advancement being fundamentally attributed to innovations in sensor technology and applications of multi-sensor fusion techniques. To consolidate existing knowledge on multi-sensor fusion technologies in the autonomous driving domain, this paper provides a comprehensive review of various technical details regarding multi-sensor fusion algorithms.

Firstly, the paper systematically analyzes core sensor technologies employed in autonomous driving systems, elucidating their operational principles and technical characteristics. Subsequently, it introduces the technical background and framework paradigms of multi-sensor fusion algorithms, accompanied by concrete implementation methods of specific fusion algorithms in 3D object detection. Multiple advanced algorithms in feature-level fusion and decision-level fusion are examined and compared.

Building upon this foundation, the paper further summarizes current challenges and proposes future research directions in this field. Persistent technical obstacles, including data alignment inaccuracies and data loss issues, pose significant risks to the safe operation of autonomous driving systems by compromising environmental perception integrity and potentially triggering severe traffic accidents. Ongoing research endeavors focus on optimizing existing multi-sensor fusion algorithms through deep learning techniques, which demonstrate potential in reducing computational complexity and accelerating network inference speed via predictive reasoning mechanisms, to enhance the reliability of intelligent driving. It is anticipated that emerging methodologies developed in the coming years will effectively mitigate these challenges, ultimately advancing the realization of fully intelligent driving solutions.

References

1. S. M. Patole, M. Torlak, D. Wang and M. Ali, Automotive radars: A review of signal processing techniques, in *IEEE Signal Processing Magazine*, 34(2), 22-35(2017).
2. X. Wang, K. Li and A. Chehri, Multi-Sensor Fusion Technology for 3D Object Detection in Autonomous Driving: A Review, in *IEEE Transactions on Intelligent Transportation Systems*, 25(2), 1148-1165(2024).
3. J. Bai, S. Li, L. Huang and H. Chen, Robust Detection and Tracking Method for Moving Object Based on Radar and Camera Data Fusion, in *IEEE Sensors Journal*, 21(9), 10761-10774(2021).
4. Z. Liu et al., Robust Target Recognition and Tracking of Self-Driving Cars With Radar and Camera Information Fusion Under Severe Weather Conditions, in *IEEE Transactions on Intelligent Transportation Systems*, 23(7), 6640-6653(2022).
5. R. O. Dubayah and J. B. Drake, LiDAR remote sensing for forestry, *J. Forestry*, 98(6), 44-46(2000).
6. R. Roriz, J. Cabral and T. Gomes, Automotive LiDAR Technology: A Survey, in *IEEE Transactions on Intelligent Transportation Systems*, 23, 7, July 2022, 6282-6297.
7. D. L. Hall and J. Llinas, An introduction to multisensor data fusion, in *Proceedings of the IEEE*, 85, 1, Jan. 1997, 6-23.
8. R. C. Luo and C. -C. Chang, Multisensor Fusion and Integration: A Review on Approaches and Its Applications in Mechatronics, in *IEEE Transactions on Industrial Informatics*, 8(1), 49-60(2012).
9. Y. Yang, C. Han, X. Kang, and D. Han, An overview on pixel-level image fusion in remote sensing, in *Proc. IEEE Int. Conf. Autom. Logist.*, Aug. 2007, 2339-2344.
10. A. Rattani, D. R. Kisku, M. Bicego, and M. Tistarelli, Feature level fusion of face and fingerprint biometrics, in *Proc. IEEE Int. Conf. Biometr.: Theory, Appl., Syst.*, Sep. 1-6(2007).
11. G. A. Wilkin and X. Huang, K-means clustering algorithms: Implementation and comparison, in *Proc. Int. Multi-Symp. Comput. Comput. Sci.*, Aug. 133-136(2007).
12. K. Venkatalakshmi and S. M. Shalinie, Classification of multispectral images using support vector machines based on PSO and k-means clustering, in *Proc. Int. Conf. Intell. Sens. Inf. Process.*, Jan. 2005, 127-133.
13. R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, PointNet: Deep learning on point sets for 3D classification and segmentation, in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, 77-85.
14. D. Xu, D. Anguelov, and A. Jain, PointFusion: Deep sensor fusion for 3D bounding box estimation, in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, 244-253.
15. K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, 770-778.
16. C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, Frustum PointNets for 3D object detection from RGB-D data, in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, 918-927.
17. R. Wan, S. Xu, W. Wu, X. Zou, and T. Cao, From one to many: Dynamic cross attention networks for LiDAR and camera fusion, 2022, arXiv:2209.12254.
18. T. Huang, Z. Liu, X. Chen, and X. Bai, EPNet: Enhancing point features with image semantics for 3D object detection, in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Glasgow, U.K.: Springer, Aug. 2020, 35-52.
19. X. Xu, S. Dong, L. Ding, J. Wang, T. Xu, and J. Li, FusionRCNN: LiDAR-camera fusion for two-stage 3D object detection, 2022, arXiv:2209.10733.

20. Y. Yan, Y. Mao, and B. Li, SECOND: Sparsely embedded convolutional detection, *Sensors*, 18, 10, p. 3337, Oct. 2018.
21. L. Xie et al., PI-RCNN: An efficient multi-sensor 3D object detector with point-based attentive cont-conv fusion module, in *Proc. AAAI Conf. Artif. Intell.*, 34(7), 12460-12467(2020).
22. S. Vora, A. H. Lang, B. Helou, and O. Beijbom, PointPainting: Sequential fusion for 3D object detection, in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, 4603-4611.
23. X. Bai et al., TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers, in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, 1080-1089.
24. T. Liang et al., BEVFusion: A simple and robust LiDAR-camera fusion framework, 2022, arXiv:2205.13790.
25. G. Wang, B. Tian, Y. Zhang, L. Chen, D. Cao, and J. Wu, Multi-view adaptive fusion network for 3D object detection, 2020, arXiv:2011.00652.
26. Y. Zhou and O. Tuzel, VoxelNet: End-to-end learning for point cloud based 3D object detection, in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, 4490-4499.
27. Y. Li et al., Voxel field fusion for 3D object detection, in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, 1110-1119.
28. Z. Chen, Z. Li, S. Zhang, L. Fang, Q. Jiang, and F. Zhao, AutoAlignV2: Deformable feature aggregation for dynamic multimodal 3D object detection, 2022, arXiv:2207.10316.
29. V. A. Sindagi, Y. Zhou, and O. Tuzel, MVX-Net: Multimodal Voxelnet for 3D object detection, in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, 7276-7282.
30. Y. Jiao, Z. Jie, S. Chen, J. Chen, L. Ma, and Y.-G. Jiang, MSMDFusion: Fusing LiDAR and camera at multiple scales with multi-depth seeds for 3D object detection, 2022, arXiv:2209.03102.
31. Z. Wang and K. Jia, Frustum ConvNet: Sliding frustums to aggregate local point-wise features for amodal 3D object detection, in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, 1742-1749.
32. S. Pang, D. Morris, and H. Radha, CLOCs: Camera-LiDAR object candidates fusion for 3D object detection, in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, 10386-10393.
33. X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, Multi-view 3D object detection network for autonomous driving, in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, 6526-6534.
34. J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, Joint 3D proposal generation and object detection from view aggregation, in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, 1-8.
35. C. Gao, G. Wang, W. Shi, Z. Wang and Y. Chen, Autonomous Driving Security: State of the Art and Challenges, in *IEEE Internet of Things Journal*, 9(10), 7572-7595(2022).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

