UDC 004.891

SENSOR FUSION FOR AUTONOMOUS VEHICLE ENVIRONMENTAL PERCEPTION: A REPRODUCIBLE COMPARISON

XINYU ZHANG

Belarusian State University of Informatics and Radioelectronics, Republic of Belarus

Received April 10, 2026

Abstract. Autonomous vehicles (AVs) rely heavily on advanced sensor technologies to ensure safe navigation and obstacle avoidance in dynamic and diverse environments, yet single sensors have inherent limitations in perceptual information completeness, environmental adaptability and functional safety, failing to support stable autonomous driving in complex open scenarios. Sensor fusion thus becomes a core approach for autonomous driving perception systems, which synthesizes multi-source sensory data to boost perception accuracy, enrich environmental cognition and enhance system robustness beyond individual sensor capabilities. This paper reproduces and compares two different sensor fusion methods, intending to offer practical references for the real-world application of such systems in autonomous driving.

Keywords: autonomous vehicles, sensor fusion, object detection.

Introduction

As a core component of intelligent transport systems, autonomous driving technology hinges on precise, comprehensive and robust perception of the driving environment, with the reliability of environmental perception directly determining the decision-making safety and operational stability of autonomous driving systems [1]. Single-sensor perception solutions have inherent limitations rooted in their physical characteristics: RGB cameras capture rich visual semantic information for accurate identification of object categories, textures and scenes, yet they are vulnerable to environmental interferences such as lighting changes, object occlusion and adverse weather, and cannot directly obtain the 3D geometric and distance information of objects [2]. In contrast, LiDAR generates high-precision 3D point cloud data via laser ranging to realize illumination-independent detection of target distance and spatial position with strong anti-interference ability [3], but it lacks target semantic information, making it hard to distinguish the category attributes of detected objects [4], and its point cloud data is plagued by high noise and sparse features for small targets.

These inherent drawbacks render single-sensor solutions incapable of meeting the perception demands of autonomous driving in complex scenarios, thus making multimodal fusion perception a key technological approach to address this issue [3]. Fusing the visual semantic information of RGB cameras with the 3D geometric information of LiDAR leverages the complementary strengths of the two sensors, compensates for the information gaps of monomodal perception [2], enhances the completeness, accuracy and robustness of environmental perception, and provides more reliable environmental data support for the decision-making and planning modules of autonomous driving systems, which has become a research focus and development trend in the field of autonomous driving perception [5].

In recent years, extensive research has been conducted on vision-LiDAR multimodal fusion perception, with fusion levels evolving from raw data-level to feature-level and decision-level, and fusion algorithms advancing from traditional manual feature fusion to deep learning-based end-to-end fusion [6]. Although existing studies have verified the effectiveness of multimodal fusion perception, most focus on the algorithm optimization of a single fusion scheme, with few comparative analyses of different-level fusion schemes and a lack of intuitive validation of their detection performance in real-world scenarios. Additionally, experimental research on multimodal fusion for small and medium-sized

datasets is relatively scarce, which fails to meet the research needs of university academic research and introductory engineering practice.

To fill these gaps, this paper takes the KITTI [7] autonomous driving dataset as the experimental carrier and constructs a vision-LiDAR multimodal fusion perception experimental framework to conduct end-to-end experimental verification from monomodal perception to multimodal fusion perception. The research mainly includes four aspects: first, building RGB-only and LiDAR-only monomodal perception modules to verify their inherent limitations in acquiring semantic and geometric information; second, designing and implementing a data-level (low-level) fusion scheme based on point cloud projection, which projects LiDAR point clouds onto the image plane to achieve pixel-level fusion of visual pixel and LiDAR depth information for raw data-level information complementarity; third, designing and implementing a feature-level (middle-level) fusion scheme based on feature extraction and weighted fusion [4], which adopts lightweight encoders to extract image semantic features and point cloud geometric features respectively, and realizes high-level feature alignment and fusion for feature-level information interaction; fourth, systematically comparing and analyzing the experimental results of monomodal perception and different-level multimodal fusion perception from multiple dimensions including the number of detected targets, category recognition ability, distance information acquisition accuracy and perception robustness. Through this research, the technical advantages and application limitations of each perception scheme are explored in depth, providing experimental evidence and technical references for the engineering design and introductory research of vision-LiDAR multimodal fusion perception algorithms.

Related Works

Monomodal perception methods are divided into visual perception and LiDAR perception. The former relies on camera-based object detection (such as YOLO and Faster R-CNN), which excels at semantic recognition but lacks depth information and is susceptible to lighting conditions or occlusion; the latter (DBSCAN) excels at geometric localisation and ranging but lacks semantic information [8].

Multi-sensor fusion can be categorised into three types based on the level of fusion. Firstly, there is data-level fusion, which occurs at the raw data level (such as projecting point clouds onto images); whilst this preserves the most complete information, it involves significant computational effort and demands high registration accuracy. Secondly, there is feature-level fusion, which operates at the feature level (matching 2D detection boxes with 3D clusters). This approach balances accuracy and efficiency and is the current mainstream solution. Finally, there is decision-level fusion (high-level fusion), which operates at the decision-making level (voting on results from different modalities). This approach offers high flexibility but results in significant information loss.

Methodology

This research implements four perception modes: object detection using RGB cameras only, clustering of LiDAR point clouds only, data-level fusion and feature-level fusion. The overall processing workflow of the proposed multi-sensor fusion perception method is as follows: it sequentially performs the loading of multi-source sensor data; camera and LiDAR external parameter calibration and coordinate registration; single-modal perception comprising RGB image object detection and LiDAR point cloud clustering; and multi-modal information fusion at the raw data and intermediate feature levels, ultimately resulting in the visualisation of the fusion results.

1. Data and Sensor Configuration

The experiment uses the KITTI dataset, which includes: colour camera images, LiDAR point clouds, calibration files. Data formats: Images, PNG format, 2D object detection input. Point clouds, BIN format, with each point containing (x, y, z, r), where r is the reflectance. Calibration parameters, external parameters matrix R, translation vector T.

2. Sensor Calibration and Coordinate Transformation

2.1. Conversion from lidar to camera coordinate system

Converting LiDAR points from the LiDAR coordinate system to the camera coordinate system, the calculation formula is shown as equation

$$p_{cam} = R \cdot P_{velo} + T \,. \tag{1}$$

2.2. Projection from the camera onto the image plane

Use the projection matrix P_{rect_03} from the calibration file to project 3D points onto 2D pixel coordinates.

3. Unimodal Perception Module

3.1. RGB camera object detection

2D object detection using the YOLOv8 lightweight network [9], input RGB image and output object class, confidence score and 2D bounding. The purpose of this step is to provide semantic information about objects (vehicles, pedestrians, road signs, etc.).

3.2. LiDAR point cloud clustering

Extracting 3D targets using DBSCAN unsupervised clustering. The input is a ground-corrected LiDAR point cloud, and the output is a number of point cloud clusters, each representing a distinct target. The purpose of this step is to provide the 3D positions and distances of the targets. In experiments using LiDAR alone, the system can output precise distances but is unable to identify target categories.

4. Low-Level Fusion

The low-level fusion performs alignment directly at the raw data level. First, the image and point cloud are read; the point cloud is then projected onto the image via coordinate transformation and projection. This is subsequently fed into the YOLO detector to obtain 2D bounding boxes; the mean depth of all projected points within the box is calculated; and finally, the object class and average depth are output. The depth is calculated using the formula as equation

$$d_{mean} = \frac{1}{N} \sum_{i=1}^{N} d_i \,, \tag{2}$$

where N represents the number of valid points in the detection box. This fusion method preserves the original depth distribution without point cloud denoising or clustering; it is simple to implement and retains all information, but is susceptible to interference from background points, and the depth accuracy is generally moderate.

5. Middle-Level Fusion

The intermediate fusion performs matching at the feature level: the RGB component is responsible for YOLO's output of 2D bounding boxes and classes, whilst the LiDAR component is responsible for DBSCAN's output of 3D cluster centres. Subsequently, the 3D cluster centres are projected onto the image for matching; if a bounding box contains a cluster centre, it is identified as the same object. Finally, the object class and the precise depth of the centre are output. This approach combines cluster-based noise filtering with ground control points, utilising a depth map centred on the target, thereby achieving greater accuracy whilst offering both semantic and geometric advantages.

Ablation Study

To demonstrate the limitations of a single sensor and the benefits of multi-sensor fusion, this paper conducted ablation experiments on the KITTI dataset, designing four comparison scenarios: RGB vision only, LiDAR only, fusion at the raw data level, and fusion at the feature level.

1. RGB-only

When using only RGB images for YOLOv8 object detection, the model is able to identify objects such as cars and parking meters within the scene. However, due to the lack of depth information, the model is unable to provide the three-dimensional position or distance parameters of the objects. At the same time, visual inspection is susceptible to changes in lighting, occlusion and shadows, resulting in insufficient stability in object recognition. This experiment demonstrated the fundamental limitations of RGB-only: whilst it possesses semantic capabilities, it lacks the ability to determine distance and geometric positioning.

2. LiDAR-only

Using only LiDAR point clouds for DBSCAN clustering, the experiment partitioned the raw point cloud into 354 target clusters. A bird's-eye view (BEV) visualization (Figure 1) shows that the point

cloud exhibits a typical distribution of scanning arcs, with each cluster distinguished by a different colour.

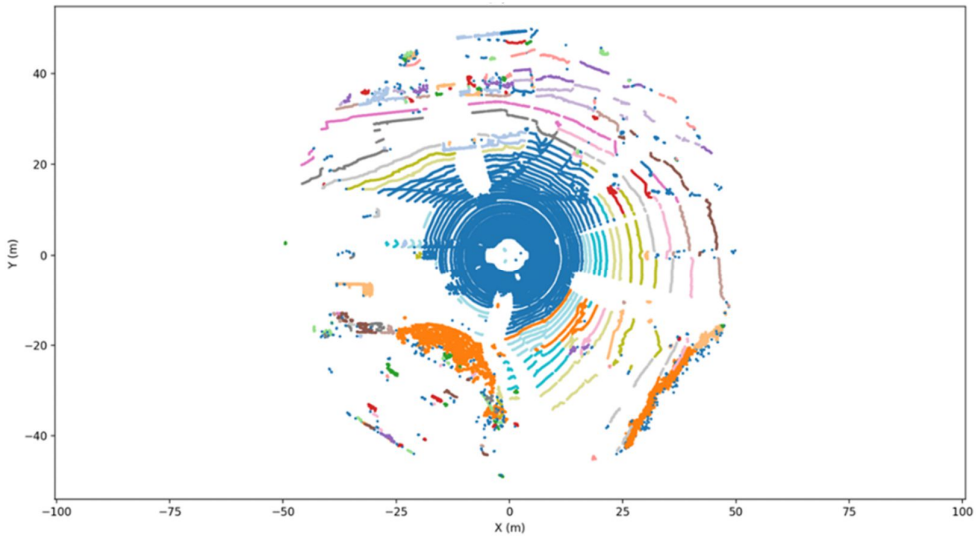

Figure 1. LiDAR obstacle clustering results

This approach provides precise 3D position and distance information for targets, but lacks semantic classification and cannot distinguish whether a target is a vehicle, a pedestrian or a road sign. Consequently, a pure LiDAR approach can only provide a geometric foundation and cannot achieve target-level semantic understanding. Experimental results indicate that a single LiDAR modality suffers from a significant lack of semantic information in target recognition.

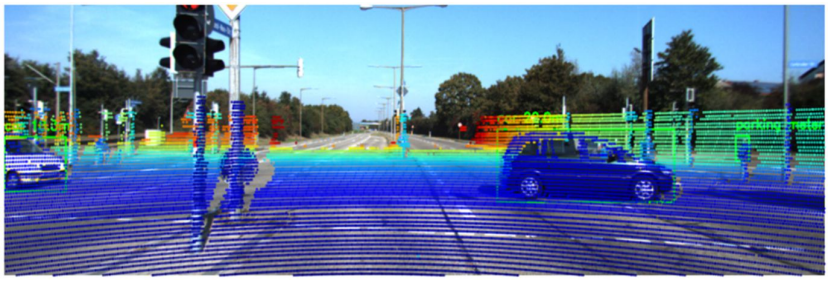

Figure 2. In-depth visualisation of data-level Fusion

3. Data-Level Fusion (low-level fusion)

The low-level fusion projects the LiDAR point cloud directly onto the image plane, thereby achieving spatial alignment of the raw sensor data. This model is capable of outputting target categories and approximate depths: the depth of the vehicle and that of the parking timer (Figure 2). Although this method preserves the full scene depth, it does not perform clustering on the point cloud; consequently, a large number of background points are included within the detection boxes, leading to noise interference in the depth calculations.

4. Feature-Level Fusion (middle-level fusion)

The middle-level fusion stage first extracts visual features (YOLOv8 2D detection) and point cloud features (DBSCAN 3D clustering) separately, and then achieves multimodal information fusion through feature matching (Figure 3).

Compared with other approaches, feature-level fusion offers three key advantages: Firstly, more precise target localisation: by filtering out noise through clustering, only the target's central point cloud is retained, resulting in distances that are closer to the true values; Sencondly, greater robustness: only targets successfully matched across modalities are retained, filtering out isolated false positives; Thirdly, complementary semantic and geometric information: it combines visual class recognition with the precise ranging capabilities of LiDAR.

Figure 3. Feature-level Fusion obstacle detection

Result and discussion

1. Results of monomodal perception

When relying solely on an RGB camera for object detection, the model identified a total of six objects, comprising two cars, three traffic lights and one parking meter. Visual detection is capable of clearly distinguishing between object categories, but it cannot provide information on the distance to the objects and is susceptible to environmental factors such as lighting and occlusion; consequently, the confidence level for detecting small or distant objects is relatively low. However, when relying solely on LiDAR for point cloud clustering, a total of 354 3D target clusters were obtained. Whilst LiDAR provides precise distance information and offers more stable target detection at long ranges and in low-light conditions, it is unable to distinguish between target categories. Consequently, the clustering results contain a large number of noise points and non-target clusters, resulting in a significant lack of semantic information regarding the targets.

2. Results of multimodal fusion perception

In the initial fusion stage, the LiDAR point cloud is projected onto the image plane to achieve pixel-level fusion of depth information with visual information. This method yields a total of 184999 valid projected points, enabling the assignment of precise distance information to visually detected targets: The average distances to the two vehicles are 22,01 m and 18,29 m respectively, the average distance to the parking timer was 28,89 m.

It simultaneously retains the class information from visual detection and the distance information from LiDAR, achieving a preliminary alignment of semantic and geometric information for the targets.

After performing weighted fusion of image features and point cloud features at the feature level, the frature-level fusion results are fed into YOLOv8 for object detection. Ultimately, a total of two car objects were identified; the detection results were consistent with those of single-modal visual detection, whilst also achieving more reliable object distance estimates. Target 1 (green box): car, 0,92 confidence, 13,13m, labelled as Vision and LiDAR fusion type. Target 2 (orange box): car, 0,63 confidence, unknown distance, labelled as Vision-only type.

The fused model significantly improves the accuracy of distance estimation for close-range targets whilst retaining the high-precision class recognition capability of visual detection, thereby achieving complementary information compared to single-modal perception (as shown in Table1).

Table 1. **Performance Comparison of Different Perception Methods**

Methods	Target	Ctegory recognition	Distance information	Advantages	Limitation
RGB-only	6	Yes	No	clear category recognition	no distance information
LiDAR-only	354	No	Yes	accurate distance measurement	no category information
Low-level Fusion	3	Yes	Yes	pixel-level information alignment	relies on precise calibration
Mid-level Fusion	2	Yes	Yes	feature-level information complementarity and stronger robustness	complex implementation

In the current experiment, middle-level fusion is still unable to provide distance information for some targets (such as Target 2). The primary reason for this is the absence of a valid LiDAR point cloud projection for that target area, which prevents the acquisition of geometric information during feature fusion. In future, the completeness and accuracy of the fused perception can be further improved by optimising the point cloud feature generation strategy and refining the feature alignment algorithm.

Conclusion

The experimental results demonstrate that: Single-modal perception has clear limitations: visual detection lacks geometric information, whilst LiDAR lacks semantic information; neither can meet the requirements of autonomous driving for comprehensive target perception. Low-level fusion achieves basic information complementarity: assigning distance information to visual targets via point cloud projection is the most intuitive fusion method, but it relies on precise sensor calibration and incurs high computational costs. Mid-level fusion enhances perception robustness: by fusing the strengths of the two modalities at the feature level, this approach retains the class recognition capabilities of visual detection whilst incorporating the distance information from LiDAR, significantly improving the perception accuracy of close-range targets. It is a fusion scheme of greater engineering value.

References

1. Y. Zhang, S. Lu.// SEC 2025.
2. M. Nawaz, J. K. T. Tang, K. Bibi, S. Xiao, H. P. Ho, W. Yuan.// IEEE 2024.
3. D. J. Yeong, G. Velasco-hernandez, J. Barry, J. Walsh.// Sensors 2021.
4. F. García, D. Martín, A. de La Escalera, J. M. Armingol.// IEEE 2017.
5. W. Li, Z. Xie, P. He.// ICSI 2025.
6. B. S. Jahromi, T. Tulabandhula, S. Cetin.// Sensors (Switzerland). 2019.
7. IEEE Staff.// 2012 IEEE Conference on Computer Vision and Pattern Recognition 2012.
8. W. Hadikurniawati, K. D. Hartomo, I. Sembiring.// ICMERALDA 2023.
9. I. Logeswaran, H. Eissa, R. Almadhoun.// ICAC 2024.