International Journal of Multidisciplinary Research and Publications
ISSN (Online): 2581-6187

A YOLO-Based Multimodal Object Detection Method for Intelligent Driving

Yuqi Huang [1], Chenlin Ma[2], Jing Zhao[3], Yuhuan Fang[4], Wenxin Zhang[5]
[1]Anhui University of Finance and Economics, Bengbu 233030, Anhui, China
[2]School of Management Science and Engineering, Anhui University of Finance and Economics, Bengbu 233030, Anhui, China
[3]School of Liberal Arts, Anhui University of Finance and Economics, Bengbu 233030, Anhui, China
[4, 5]School of Public Finance and Administration, Anhui University of Finance and Economics, Bengbu 233030, Anhui, China
Email: 2978997730@qq.com

Abstract*—This paper proposes a YOLO-based multi-modal object detection method for intelligent driving, which improves the system fault tolerance by fusing visual, infrared, LiDAR and Radar data. The YOLOv5 network is optimized by building a dual-stream backbone network for parallel extraction of multi-scale depth and RGB information, and inserting the CBAM module after the SPPF layer to enhance detection accuracy and scene adaptability. In addition, LiDAR point cloud is integrated, and 3D box regression is added to the YOLO head for 3D target detection. The findings might be used to advance the development of intelligent transportation systems, as well as autonomous vehicles and intelligent traffic monitoring systems.*

***Keywords*—** *YOLOv5; Multi-modal fusion; Intelligent driving; Object detection; CBAM module.*

I. Introduction

Object detection is the core of the intelligent driving perception system, and accurate real - time detection is the premise of autonomous vehicle safety[1]. Traditional single - modal detection has obvious limitations: visual detection is affected by light and weather, LiDAR has high data sparsity, and infrared detection has low resolution[4], which leads to poor performance in complex scenarios such as extreme weather and target occlusion.

China's Intelligent Vehicle Innovation and Development Strategy clearly advocates multi-sensor fusion perception technology, which provides a direction for the upgrade of intelligent driving perception[2].

Based on the high-real-time YOLOv5 algorithm, this paper fuses visual, infrared, LiDAR, and radar data. It improves the YOLOv5 network structure and designs a multi-modal detection framework for intelligent driving. This research enriches the multi-modal perception theory of intelligent driving and provides a feasible technical scheme for high-precision detection in complex environments, with important practical value for the development of intelligent transportation systems.

The functional modules of an intelligent driving system are shown in Fig. 1.

II. Related Theoretical Basis

YOLOv5 Algorithm

YOLOv5 is a classic single-stage target detection algorithm with end-to-end detection capability. It is composed of Backbone, Neck, and Head[5]. The Backbone adopts CSPDarknet53 for multi-scale feature extraction. The Neck integrates FPN and PAN to fuse shallow position and deep semantic information, solving the multi-scale detection information imbalance problem. The Head outputs detection results for large, medium, and small targets through three different scale feature maps, with the advantages of high speed and good accuracy.

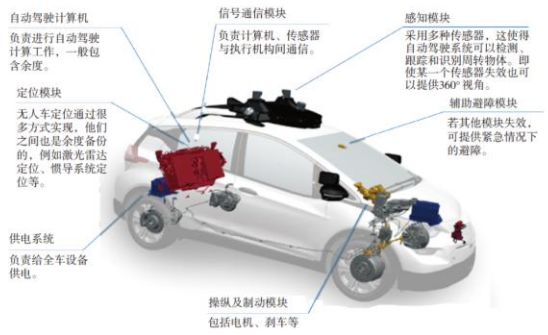

Fig. 1. Functional module diagram of intelligent driving vehicle.

The architecture of the YOLOv5s model used in this paper is shown in Fig. 2.

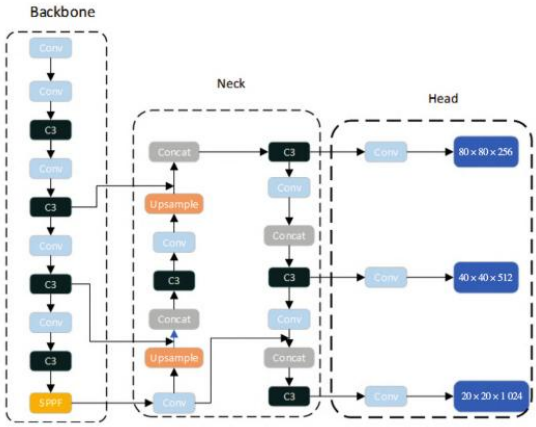

Fig. 2. Architecture of YOLOv5s model.

Multi-Modal Data Complementarity

The multi-modal data in this study includes RGB images, infrared images, LiDAR point cloud, and radar data, which

Yuqi Huang, Chenlin Ma, Jing Zhao, Yuhuan Fang, and Wenxin Zhang, "A YOLO-Based Multimodal Object Detection Method for Intelligent Driving," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 8, Issue 10, pp. 260-262, 2026.

have complementary characteristics. RGB images have rich texture information but are affected by the environment. Infrared images are not limited by light and are suitable for night or foggy days, but they have low resolution. LiDAR can obtain high-precision 3D spatial information, yet its data is sparse. Radar can accurately measure distance and speed, but it has low spatial resolution. The fusion of multi-modal data makes up for the defects of single-modal perception and improves the system's environmental adaptability.

CBAM Attention Mechanism

CBAM is a lightweight attention module that includes channel and spatial attention[7]. The channel attention assigns weights to feature channels based on importance, and the spatial attention focuses on key target areas while suppressing background noise. Inserting CBAM into YOLOv5 can enhance the network's ability to extract features of small and occluded targets and improve detection accuracy.

III. Design of Multi-Modal Detection Method Based on YOLOv5

Overall Framework

The method framework comprises four parts: multi-modal data preprocessing, spatio-temporal alignment, improved YOLOv5 feature extraction, and multi-modal fusion detection.

First, preprocess the original sensor data (such as image enhancement and point cloud denoising). Then, achieve spatio-temporal alignment of multi-modal data via LiDAR-camera chessboard calibration and infrared-RGB affine transformation. Next, extract features using the improved YOLOv5 network. Finally, fuse multi-modal features with a dynamic strategy and incorporate 3D box regression to realize 2D/3D joint detection.

Data Preprocessing and Spatio-Temporal Alignment

For RGB and infrared images, contrast adjustment and denoising are employed for enhancement. LiDAR point cloud is denoised through statistical filtering and converted into 3 - channel pseudo - images (height/intensity/density map) for YOLOv5 processing. Radar data is filtered to extract effective target information.

Spatio - temporal alignment ensures data consistency within the same coordinate system. LiDAR - camera alignment utilizes the chessboard calibration method to obtain the rotation matrix R and translation vector t for coordinate conversion[3]. Infrared - RGB alignment depends on affine transformation to map infrared pixel points to the corresponding positions of RGB images, achieving target coincidence in the two images.

YOLOv5 Network Optimization

Dual-Stream Backbone Network

A dual - stream backbone is constructed for parallel multi - modal feature extraction. The RGB branch uses the original CSPDarknet53 to extract texture and semantic features. The LiDAR branch adopts GhostNet to reduce parameters by about 40% and improve computational efficiency[3]. The input channel of the first convolution layer is adjusted from 3 to 6 to receive the spliced feature map of RGB (3 channels) and LiDAR pseudo - image (3 channels).

CBAM Module and Dynamic Fusion

The CBAM module is inserted after the YOLOv5 SPPF layer to adaptively enhance key feature information and suppress noise, thus improving the detection ability for small, occluded, and distant targets[3].

A dynamic feature interaction module is added in the Neck layer to fuse RGB and LiDAR features according to attention weights. The Head layer dynamically adjusts the NMS IoU threshold according to the confidence of each branch and weighted fuses the detection results to improve accuracy.

3D Detection and Loss Function Design

A 3D detection head is added to the YOLOv5 Head layer, fusing LiDAR point cloud to realize target 3D box regression. The output parameters include the target center depth z, size (l, w, h), and heading angle θ[6]. The model adopts multi - task joint loss, including 2D detection loss L_{2D}, 3D detection loss, L_{3D} and infrared feature loss $L_{thermal}$.

$$L_{3D} = \lambda_1 \cdot \text{SmoothL1}(z) + \lambda_2 \cdot \text{SmoothL1}(l, w, h) + \lambda_3 \cdot (1 - \cos(\theta))$$

$$L_{total} = L_{2D} + 0.5 \cdot L_{3D} + 0.1 \cdot L_{thermal}$$

$\lambda_1, \lambda_2, \lambda_3$ are weight coefficients to balance the contribution of each loss term to model training.

IV. Experimental Verification

Experimental Environment and Dataset

The experiment is based on the NVIDIA V100 GPU, using PyTorch, OpenCV, and ROS as the software framework[3]. The test datasets include the nuScenes multi - modal dataset (covering rain/snow/fog scenarios), the FLIR ADAS dataset (infrared - RGB aligned data for night detection), and a self - collected custom dataset (for verifying the model's generalization ability).

Evaluation Metrics and Results

The model is evaluated from three aspects: accuracy (mAP@0.5, 3D IoU, night missing detection rate), real - time performance (FPS, end - to - end delay), and robustness (detection stability under single sensor failure).

Experimental results show that the proposed model has obvious advantages compared with the original YOLOv5 and single - modal models. The mAP@0.5 on the nuScenes dataset exceeds 92% (8% higher than the original YOLOv5), the night missing detection rate is reduced by 15% compared with the pure visual model, and the 3D IoU is more than 85%. The FPS is maintained above 30, and the end - to - end delay is less than 100ms, meeting the real - time requirements of intelligent driving. When a single sensor fails, the detection accuracy only decreases by about 3%, with strong fault tolerance.

Ablation experiments verify that the dual - stream backbone, CBAM module, and dynamic fusion strategy all effectively improve the model performance.

V. Conclusion

This paper proposes a YOLOv5-based multi-modal intelligent driving target detection method, which fuses multi-

Yuqi Huang, Chenlin Ma, Jing Zhao, Yuhuan Fang, and Wenxin Zhang, "A YOLO-Based Multimodal Object Detection Method for Intelligent Driving," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 8, Issue 10, pp. 260-262, 2026.

sensor data to address the poor performance of single-modal perception in complex scenarios. The innovations include constructing a dual-stream backbone for parallel feature extraction, inserting CBAM to enhance target feature extraction, designing dynamic fusion and adaptive NMS to improve accuracy, and adding a 3D detection head to achieve 2D/3D joint detection.

Experimental results indicate that the model exhibits high accuracy, real-time performance, and robustness. It can effectively detect targets in complex scenarios such as extreme weather and low light conditions, with a significant improvement in system fault tolerance. The results can be applied to autonomous vehicles and intelligent traffic monitoring systems, offering technical support for the development of intelligent transportation.

In the follow-up, the model will be further lightweighted for deployment on edge computing devices. Additionally, transformer-based cross-modal feature alignment technology will be introduced to optimize the multi-modal fusion strategy and further enhance detection performance.

ACKNOWLEDGMENT

This work was supported by the Undergraduate Scientific Research Innovation Fund of Anhui University of Finance and Economics (Project Approval Number: XSKY25149).

The authors would like to thank Professor Yin Shishu for her professional guidance, as well as all team members for their support and cooperation during the whole research process.

REFERENCES

[1] Zeng W S. Research on Traffic Object Detection Algorithm Based on Improved YOLOv5[D]. Guilin: Guilin University of Electronic Technology, 2023.

[2] Liu K, Song X J. Research on Driving Object Detection Algorithm Based on Improved CenterNet[J]. Automation and Instrumentation, 2024, 39(10):108-112.

[3] LI Y. Multispectral pedestrian detection in autonomous driving: A review [J]. IEIE Transactions on Smart Processing & Computing,2021,10(1):10-16.

[4] Fursa I, Fandi E, Musat V, et al. Worsening Perception: Real-time Degradation of Autonomous Vehicle Perception Performance[J]. arXiv preprint arXiv:2103.02760, 2021.

[5] Gao Q, Tang F X, Li D, et al. Research on Pedestrian Detection Method in Dense Scenes Based on Improved YOLOv5[J]. Foreign Electronic Measurement Technology, 2023,42 (4): 125-130.

[6] Apollo Team. YOLO-3D: A Unified Framework for 3D Object Detection in Autonomous Driving[R]. Baidu Research, 2023.

[7] ZHANG X, ZHOU X, LIN M, et al. ShuffleNet: An extremely efficient convolutional neural network for mobile devices[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2018:6848-6856.