Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD24502

Multimodal Fusion Target Detection Based on MF-YOLO and Its Innovative Applications in Automotive Field

Yuanze Xiao

Wuhan University of Technology, Wuhan, China
338187@whut.edu.cn

Abstract. As autonomous driving and intelligent transport systems advance, vehicles face growing demands for environmental perception. Current target detection technologies struggle with small targets and complex conditions like low light and bad weather, raising concerns about autonomous driving safety. This paper introduces an innovative design integrating MF-YOLO technology with vehicles. By fusing IR and RGB data, MF-YOLO boosts detection accuracy and robustness. It also incorporates a BRA module and an improved loss function to optimize model performance. These enhancements significantly improve detection in complex environments and enhance autopilot safety, offering a novel solution for automotive intelligence.

Keywords: MF-YOLO, multi-modal fusion, object detection, autonomous driving, automotive safety

1. Introduction

Autonomous driving and intelligent transport systems have driven a need for advanced target detection technology in vehicles. While Google's Waymo and Tesla's Autopilot use multimodal data fusion to enhance detection accuracy and robustness, these systems often require complex sensors and heavy computational resources, making lightweight and real-time performance challenging. Existing research has improved detection in complex environments like low light and bad weather through algorithmic and hardware advancements, but there remains a gap in integrating these technologies into practical automotive applications. This paper proposes integrating MF-YOLO technology with automobiles to create an efficient environment sensing system. The innovations lie in MF-YOLO's improvements over the YOLO series, including a new IR and RGB fusion method, the introduction of the BRA module, and an improved loss function [1][2]. These advancements enhance small target detection and adaptability to complex backgrounds. Additionally, the design introduces a lightweight real-time detection module and optimized multimodal fusion strategy for complex environments, significantly improving detection performance and supporting the development of automotive intelligence.

DOI: 10.54254/2755-2721/2025.LD24502

2. Core components design

2.1. MF-YOLO principle of operation

The MF-YOLO model introduces a novel image fusion method for infrared (IR) and red, green and blue (RGB) image fusion method, which makes full use of the advantages of the two modal data by using bidirectional symmetric fusion to significantly improve the accuracy and robustness of small target detection. Additionally, the model incorporates the BiLevel Routing Attention (BRA) module, which effectively boosts the model's adaptability to intricate backgrounds. Furthermore, by integrating the Intersection over Union (IoU) and Gaussian Wasserstein distance, a new loss function is designed, which serves to further optimize the performance of the model.

2.1.1. Model structure and flow

The overall model structure of MF-YOLO is based on YOLOv5s,utilizing CSPNet as the backbone for feature extraction. CSPNet comprises numerous CBS components and CSP modules [3]. CBS components involve convolution, batch normalization, and SiLu activation functions [4]. The CSP module splits the previous layer's feature map into two branches. It reduces channels via 1×1 convolution in one branch and feeds the other to a ResNet or CBS block. Eventually, the two branches' feature maps are concatenated. The SPP module employs parallel max-pooling layers of varying kernel sizes to capture multi-scale features. And the performance of the model has been significantly improved by introducing a multimodal fusion architecture, a BRA module and an improved loss function. The specific process is as follow Figure 1:

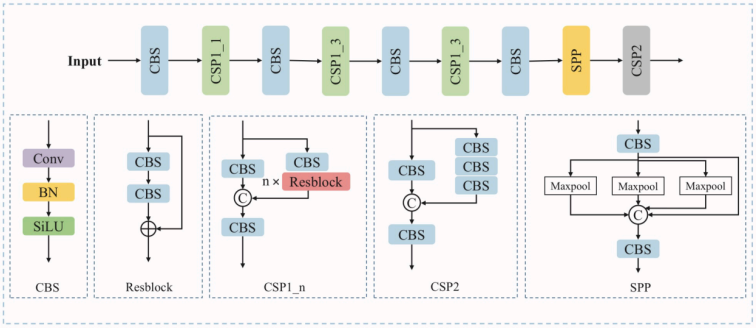

Figure 1. YOLOv5s backbone structure. Overview of CSP, CBS and SPP structures

2.1.2. Multimodal fusion architecture

One of the core innovations of MF-YOLO is its multimodal fusion architecture, which significantly improves the accuracy and robustness of target detection by fusing infrared (IR) and red, green and blue (RGB) image data [5]. Specifically, the multimodal fusion process is as follows:

First, the input RGB and IR images are normalised to the interval [0, 1] so as to eliminate the magnitude difference between different modal data, and then the resolution of the input image is reduced by a downsampling operation in order to reduce the amount of computation and increase the processing speed. The CBAM (Convolutional Block Attention Module) module is employed to extract the channel domain features of RGB and IR images, respectively, and generate the feature maps FRGB and FIR. Then, the attention maps mIR and mRGB are defined in the spatial domain, and the feature maps of the two modalities are fused by a 1×1 convolution operation:

DOI: 10.54254/2755-2721/2025.LD24502

$m_{IR} = f_1(F_{IR})$, $m_{RGB} = f_2(F_{RGB})$, where f1 and f2 denote 1×1 convolution operations for IR and RGB modalities, respectively, and ⊗ represents element-by-element matrix multiplication. The feature maps of the two modalities are combined with the original input image through a fusion operation: $O_1 = f_3(m_{RGB} \otimes F_{RGB} + I_{RGB})$, $O_2 = f_4(m_{IR} \otimes F_{IR} + I_{IR})$, where f3 and f4 are 1×1 convolution operations for further feature extraction. Finally, the MF-YOLO uses a bidirectional symmetric fusion method to process fused features from IR and RGB images. This method effectively uses both modalities, allowing for better target detection in low-light or night-time conditions and rich color and texture information in good lighting conditions. This bidirectional fusion method significantly improves the accuracy and robustness of small target detection, overcoming limitations of single modal data.

2.1.3. BiLevel Routing Attention (BRA) module

The input image is divided into W×S regions, and features are computed using an average pooling operation [6]. The features are then projected into small blocks to generate query, key, and value vectors. The similarity between regions is calculated, and the k regions with the highest similarity are selected for attention computation. The value vectors of these regions are weighted and summed to generate the final attention feature. Depthwise Convolution is used to enhance the expressive power of the feature. The BRA module filters noise in complex backgrounds, focusing more on the target region. This region-level routing and attention mechanism capture long-distance-dependent information, improving the model's ability to localize and classify the target [7]. The overall process is shown below Figure 2. The BRA module is able to effectively filter noise in complex backgrounds, allowing the model to focus more on target-related regions. Through the region-level routing and attention mechanism, the BRA module is able to capture long-distance dependent information, improving the model's ability to locate and classify targets.

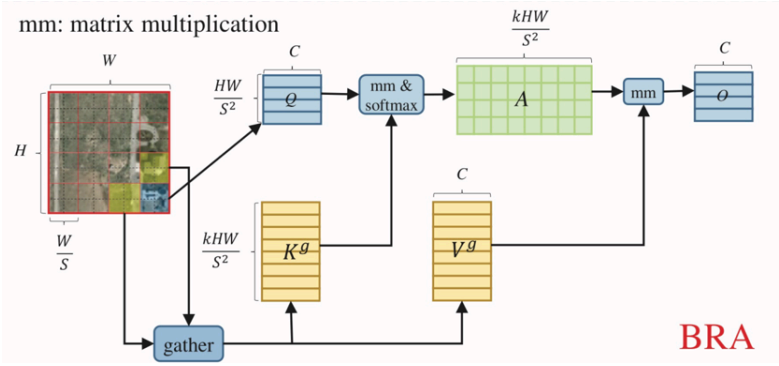

Figure 2. The composition structure of the BRA module

2.1.4. Improved loss function

In order to improve the training effect and detection accuracy of the model, MF-YOLO designs a new loss function that combines IoU (Intersection over Union) and Normalised Gaussian Wasserstein Distance (NWD) [8]. The details are as follows:

The IoU loss function is used to measure the degree of overlap between the predicted bounding box and the true bounding box, and is one of the commonly used loss functions in target detection. Its calculation formula is: $L_{IoU} = 1 - \frac{\text{Area of Overlap}}{\text{Area of Union}}$ However, the IoU loss function is deficient in some cases, such as when the prediction frame does not overlap at all with the true frame or contains

DOI: 10.54254/2755-2721/2025.LD24502

it completely, and does not provide effective gradient information. To address the shortcomings of the IoU loss function, MF-YOLO introduces the Gaussian Wasserstein distance. This distance calculates the Wasserstein distance between two Gaussian distributions by converting the position and size of the bounding box into the form of a Gaussian distribution. Its calculation formula is: $L_{NWD} = 1 - \frac{W_2^2(\mu_1,\mu_2)}{\sqrt{2\pi|\Sigma|}}$ where \mu_1 and \mu_2 denote the mean vectors of the Gaussian distribution of the predicted and true frames, respectively, and \Sigma denotes the covariance matrix.

The total loss function of MF-YOLO consists of the IoU loss and the Gaussian Wasserstein distance loss as given in Eq: $L_{total} = c_1 L_{IoU} + c_2 L_{NWD}$ where c_1 and c_2 are weighting coefficients to balance the two losses. By combining IoU loss and Gaussian Wasserstein distance loss, MF-YOLO is able to more accurately measure the difference between the predicted frame and the real frame, improving model training and detection accuracy.

2.1.5. Effectiveness prediction

Based on experimental data and graphical estimates, MF-YOLO is expected to achieve the following effects when combined with automobiles: Improved Detection Accuracy: With the introduction of multimodal fusion and BRA module, MF-YOLO has significantly improved its accuracy in small target detection. For example in table 1, on the VEDAI dataset ,the mAP@0.5 of MF-YOLO reaches 76.62%, which is a significant improvement compared with other methods [9].

Table 1: Performance of MF- YOLO on VEDAI dataset

Method	Car	Pickup	Camping	Truck	Other	Tractor	Boat	Van	Params	mAP@0.5 (%)
YOLOv3	84.57	72.68	67.13	61.96	43.04	65.24	37.10	58.29	61.54M	61.26
YOLOv4	85.46	72.84	72.38	62.82	48.94	68.99	34.28	54.66	52.51M	62.55
YOLOv5s	80.81	68.48	69.06	54.71	46.76	64.29	24.25	45.96	7.07M	56.79
YOLOv5m	82.53	72.32	68.41	59.25	46.20	66.23	33.51	57.11	21.07M	60.69
YOLOv5l	82.83	72.32	69.92	63.94	48.48	63.07	40.12	56.46	46.64M	62.16
YOLOv5x	84.33	72.95	70.09	61.15	49.94	67.35	38.71	56.65	87.25M	62.65
YOLOR	84.15	78.27	68.81	52.60	46.75	67.88	21.47	57.91	—	59.73
SuperYOLO	91.13	85.66	79.30	70.18	57.33	80.41	60.24	76.50	4.85M	75.09
MF-YOLO	92.03	86.61	78.19	72.58	57.36	82.88	64.64	78.66	4.78M	76.62

Real-time enhancement: MF-YOLO's lightweight design enables it to run efficiently on embedded on-board hardware to meet real-time requirements. For example in table 2, on the NWPU VHR-10 dataset, MF-YOLO achieves a high level of detection speed and is able to process video streams in real-time [10].

DOI: 10.54254/2755-2721/2025.LD24502

Table 2: Performance of MF-YOLO on NWPU VHR-10 dataset

Method	Params	mAP@0.5 (%)
YOLOv3	61.57M	88.30
FCOS [26]	31.86M	89.65
ShuffleNet [27]	12.10M	83.00
SuperYOLO	7.68M	93.30
MF-YOLO	6.52M	91.63

Improved adaptability to complex environments: By optimising the multimodal data fusion strategy, MF-YOLO's detection performance in complex environments such as low light and bad weather is significantly improved, providing more reliable environment sensing support for autonomous driving.

2.2. MF-YOLO design for automotive integration

The integration of MF-YOLO with automobiles is achieved through three key innovations that collectively improve the system's detection performance and robustness, meeting the stringent requirements of real-time and accuracy for autonomous driving.

2.2.1. Multimodal sensor integration

Deploying infrared (IR) and red, green and blue (RGB) cameras in vehicles enables real-time multimodal data acquisition. This type of sensor integration makes full use of the advantages of both modalities and overcomes the limitations of a single modality in certain scenarios [11]. IR cameras are sensitive to thermal radiation and are suitable for detecting targets in low-light or nighttime conditions, while RGB cameras provide rich colour and texture information and are suitable for detecting targets during the day or in good lighting conditions.

2.2.2. Real-time target detection module design

Based upon MF-YOLO technology, a lightweight real-time target detection module is meticulously crafted. The module can run efficiently on the embedded on-board hardware to meet the real-time requirements of automatic driving. By introducing the multimodal fusion architecture and the BRA module, the model structure and the algorithmic flow are systematically optimized. Such optimization serves to diminish the computational complexity and curtail resource consumption, enabling the model to execute efficiently even on in-vehicle hardware that has limited resources at its disposal.

2.2.3. Optimisation of adaptation to complex environments

In the face of complex environments including nighttime, low-light conditions, and adverse weather, the multi-modal data fusion strategy is refined to enhance the detection performance and robustness of the system within complex scenarios. Through the integration of infrared and RGB image data, the system can more accurately detect pedestrians, vehicles, and other targets in low-light or bad weather conditions, which significantly improves the safety of autonomous driving.

DOI: 10.54254/2755-2721/2025.LD24502

3. Conclusion

In this paper, a multimodal fusion target detection technique based on MF-YOLO is proposed and applied to the automotive field, which significantly improves the performance of target detection in complex environments and the safety of autonomous driving. With the multimodal fusion architecture, BRA module, and improved loss function, MF-YOLO excels in small target detection, complex background adaptation, and real-time performance. Experimental results show that the technique achieves excellent performance on multiple datasets, validating its effectiveness in practical applications. MF-YOLO is set to offer robust technical support for autonomous driving and intelligent transportation initiatives. This will not only drive advancements in these domains but also stimulate the development of related technologies, fostering innovation and progress in the broader technological landscape.

References

[1] J. Redmon and A. Farhadi, 2018, “Yolov3: An incremental improvement, ” arXiv preprint arXiv: 1804.02767.
[2] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, 2020, “Yolov4: Op timal speed and accuracy of object detection, ” arXiv preprint arXiv: 2004.10934.
[3] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, 2020, “Cspnet: A new backbone that can enhance learning capability of cnn, ” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp. 390-391.
[4] S. Elfwing, E. Uchibe, and K. Doya, 2018, “Sigmoid-weighted linear units for neural network function approximation in reinforcement learning, ” Neural networks, vol. 107, pp. 3-11.
[5] W. Han, J. Chen, L. Wang, R. Feng, F. Li, L. Wu, T. Tian, and J. Yan, 2021, “Methods for small, weak object detection in optical high-resolution remote sensing images: A survey of advances and challenges, ” IEEE Geoscience and Remote Sensing Magazine, vol. 9, no. 4, pp. 8-34.
[6] L. Zhu, X. Wang, Z. Ke, W. Zhang, and R. W. Lau, 2023, “Biformer: Vision transformer with bi-level routing attention, ” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10 323-10 333.
[7] F. Chollet, 2017, “Xception: Deep learning with depthwise separable convo- lutions, ” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1251-1258.
[8] J. Wang, C. Xu, W. Yang, and L. Yu, 2021, “A normalized gaus- sian wasserstein distance for tiny object detection, ” arXiv preprint arXiv: 2110.13389.
[9] S. Razakarivony and F. Jurie, 2016, “Vehicle detection in aerial imagery: A small target detection benchmark, ” Journal of Visual Communication and Image Representation, vol. 34, pp. 187-203.
[10] Li, W., Li, A., Kong, X., Zhang, Y., & Li, Z. (2024). MF-YOLO: Multimodal Fusion for Remote Sensing Object Detection Based on YOLOv5s. In: 2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD). Jinan, China. pp. 897-903.
[11] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, 2019, “Learning roi transformer for oriented object detection in aerial images, ” in Proceed- ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2849-2858.