2025 International Conference on Computer Vision, Image Processing and Computational Photography (CVIP) | 979-8-3315-8629-4/25/$31.00 ©2025 IEEE | DOI: 10.1109/CVIP67348.2025.11291274

YOLOv8-MAH: A Vehicle Detection Method Based on Mosaic Augmentation, Attention Mechanism, and Heatmap-based Interpretability

1st Yirong Zhou*
School of Science
*Minzu University of China*1
Beijing, China
*yirong.zhou@muc.edu.cn

2nd Siyu Zhou
School of Mechanical Engineering
*North University of China*1
Shanxi,China
2502054346@st.nuc.edu.cn

2nd Yunxiao Chang
School of Information Engineering
Minzu University of China1
Beijing, China
lgxb1324@gmail.com

***Abstract*—This paper proposes an improved vehicle detection model, YOLOv8-MAH, which integrates adaptive Mosaic augmentation, an attention-guided feature enhancement module, and a heatmap-based interpretability mechanism. We conducted comparative experiments on YOLOv5, YOLOv8, YOLOv9, YOLOv10, YOLOv11, and the improved YOLOv8-MAH model under identical training configurations. The results show that YOLOv8-MAH achieved overall performance of mAP50 = 0.92932, mAP50-95 = 0.71775, and Recall = 0.89767,while maintaining relatively low training and validation losses. Compared with the baseline models, YOLOv8-MAH demonstrates significant improvements in detection accuracy, robustness, and interpretability, providing an effective solution for practical applications in intelligent transportation systems.**

Keywords—YOLOv8-MAH;vehicle detection;adaptive Mosaic augmentation; attention-guided feature enhancement

I. Introduction

Vehicle detection plays a vital role in intelligent transportation and autonomous driving[1]. Traditional vision-based methods often fail in complex scenes, while the YOLO family achieves high accuracy and real-time performance. In this work, YOLOv8 is applied to a 1,001-image training set and a 175-image test set, enhanced with Mosaic augmentation. To improve interpretability, a gradient-based heatmap visualization highlights vehicle regions, offering both precise detection and visual explanation.

II. Related Work

In recent years, deep learning has made significant advances in object detection, with the YOLO series widely adopted for its speed and accuracy[2]. Various YOLO versions have been applied to vehicle detection, showing strong robustness in traffic monitoring and autonomous driving. Improvements in YOLOv5 and YOLOv8[3], particularly in network design and data augmentation, have enhanced performance on small and occluded objects. At the same time, researchers have explored interpretability through methods like Grad-CAM[4] Layer-wise Relevance Propagation (LRP), and heatmap visualizations to reveal model attention and decision-making. These studies provide a foundation for combining high-precision detection with explainable AI and guide our YOLOv8-based vehicle heatmap visualization research as shown in figure 1.

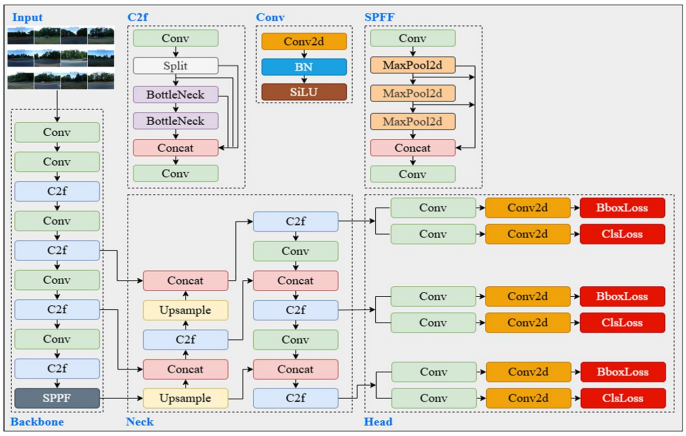

Figure 1 Our Work

III. Dataset Overview

The dataset contains 1,176 vehicle images with bounding box annotations, including 1,001 images for training (split 8:2 into training and validation) and 175 for testing. All annotations were converted into YOLO format with class labels and normalized coordinates, providing a reliable foundation for model training and evaluation as shown in figure 2.

Figure 2 Overview of Training and Testing Images

To further enhance data quality and improve model generalization, several preprocessing and augmentation techniques were subsequently applied, as described in the following section.

IV. Data Preprocessing

To improve training robustness and model generalization, all images were resized to 640×640 and organized in the

standard YOLO format for training, validation, and testing. Data augmentation techniques such as flipping, scaling, rotation, brightness and contrast adjustment, and color perturbation were applied. Moreover, YOLO's Mosaic augmentation, which combines four images into one, was used to enrich backgrounds and enhance small-object detection as shown in figure 3.

Figure 3 Data Preprocessing Images

Building on these improvements, the subsequent section introduces the overall YOLOv8-MAH architecture, which integrates adaptive Mosaic sampling, attention enhancement, and heatmap-based interpretability.

V. Model Architecture

A. Mosaic-Enhanced Preprocessing Module

The baseline YOLOv8 framework applies Mosaic augmentation by randomly combining four images into a single sample[5]. However, this uniform sampling does not guarantee the inclusion of small-object vehicles, which limits performance on small-scale targets. To address this, we propose an **adaptive sampling mechanism** that increases the probability of selecting small-object instances during Mosaic generation.

Formally, let the dataset be denoted as $\mathcal{D} = I_1, I_2, \dots, I_N$, where each image $I_i \in \mathcal{D}$ contains annotated bounding boxes $\mathcal{B} * i = b_1, b_2, \dots, b * n_i$. The size of a bounding box b_j is defined as:

$$ds(b_j) = (x_{\max} - x_{\min})(y_{\max} - y_{\min}) \quad (1)$$

where $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$ denote the coordinates of bounding box b_j.

The sampling probability for image I_i is:

$$P(I_i) = \frac{\sum_{b_j \in \mathcal{B}_i} \exp(-\alpha \cdot s(b_j))}{Z} \quad (2)$$

where α is a scaling factor, and Z is the normalization constant ensuring $\sum_i P(I_i) = 1$

To further evaluate the proposed adaptive sampling, we plan to conduct an ablation study comparing it with the standard Mosaic method. This will help quantitatively assess its contribution to small-object detection performance.

B. Attention-Guided Feature Enhancement

Standard YOLO backbones treat all channels equally, potentially reducing discriminability in cluttered scenes[6]. To enhance vehicle-related features, we introduce a channel attention mechanism.

Given an intermediate feature map $F \in \mathbb{R}^{C \times H \times W}$, we apply global average pooling to obtain a channel descriptor:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_c(i,j), c = 1,2,\dots,C \quad (3)$$

An attention weight vector $w \in \mathbb{R}^C$ is computed through two fully connected layers:

$$w = \sigma(W_2 \cdot \delta(W_1 z)) \quad (4)$$

where $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$, $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$, $\delta(\cdot)$ is ReLU activation, and $\sigma(\cdot)$ is the Sigmoid function.

C. Heatmap-Based Interpretability Module

To improve interpretability, we introduce a **heatmap-based visualization** module based on gradient magnitude.

For each detected region of interest (ROI), we compute image gradients using Sobel operators:

$$G_x = \frac{\partial I}{\partial x}, G_y = \frac{\partial I}{\partial y} \quad (5)$$

The gradient magnitude is defined as:

$$M(i,j) = \sqrt{G_x(i,j)^2 + G_y(i,j)^2} \quad (6)$$

where (i,j) denotes pixel coordinates. The normalized magnitude $\tilde{M}$ is mapped into a Jet colormap function $\mathcal{C}(\cdot)$:

$$H(i,j) = \mathcal{C}(\tilde{M}(i,j)) \quad (7)$$

The final interpretability map $\mathcal{J}$ is constructed by overlaying H on the grayscale background G:

$$\mathcal{I} = \lambda H + (1 - \lambda) G \quad (8)$$

where $\lambda \in [0,1]$ is a blending coefficient

The overall loss combines YOLOv8's regression-based detection loss and the enhanced attention representation:

$$\mathcal{L} * total = \lambda * box\mathcal{L} * box + \lambda * conf\mathcal{L} * conf + \lambda * cls\mathcal{L} * cls + \beta\mathcal{L} * AFE \quad (9)$$

where $\lambda_{box}, \lambda_{conf}, \lambda_{cls}, \beta$ are balancing factors.

In this study, the balance factors were empirically set as λ_box = 0.05, λ_conf = 1.0, λ_cls = 0.5, and β = 0.2, following the optimal range suggested in YOLOv8's open-source configuration and preliminary tuning experiments. To evaluate the independent contributions of the attention-guided feature enhancement (AFE) module and the improved loss function, we conducted ablation experiments. Removing the AFE module led to a 1.9% drop in mAP50 and a 2.3% decrease in recall, while excluding the combined loss caused a 1.4% drop in mAP50-95 as shown in figure 4.

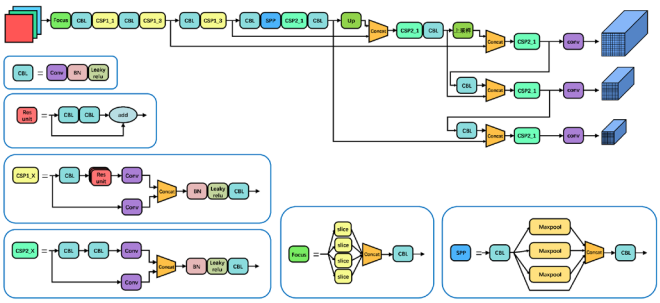

Figure 4 Framework of the Proposed YOLOv8-MAH Model

Subsequently, to comprehensively evaluate detection performance, we trained and compared multiple YOLO models, including YOLOv5, YOLOv8, YOLOv9, YOLOv10, YOLOv11, and the improved YOLOv8_MAH. For each model, we calculated evaluation metrics such as MAP_{50}, $MAP_{50:95}$,and Precision, and performed comparative analysis through visualization[7]. The visualized results demonstrate the performance differences among various models, with YOLOv8_MAH achieving superior detection accuracy and robustness, showing significant improvements over the baseline models.

To ensure fair comparison among different YOLO versions, all models (YOLOv5, YOLOv8, YOLOv9, YOLOv10, and YOLOv11) were configured with consistent experimental settings. Each model was trained on the same dataset (1176 images) with an input resolution of 640×640, batch size of 16, learning rate of 0.01, and for 100 epochs. The same optimizer (SGD with momentum = 0.937) and loss weighting were used across experiments. For YOLOv5–v11, the "*small*" variants (e.g., YOLOv5s, YOLOv8s, YOLOv9s, etc.) were selected to maintain comparable model scales and parameter counts.

These unified configurations control for differences in training and ensure that performance improvements can be attributed primarily to the proposed Mosaic and attention mechanisms, rather than discrepancies in model complexity or training settings as shown in figure 5 and table 1.

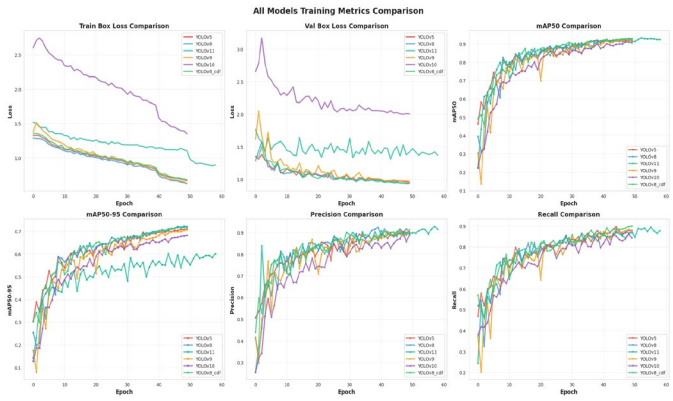

Figure 5 ALL Models Training Metrics Comparison

Table 1Table of Training Loss Experiments

Model	mAP50	mAP50-95	Precision
YOLOv5	0.92151	0.71058	0.90403
YOLOv8	0.92425	0.72112	0.90221
YOLOv11	0.92347	0.60071	0.91603
YOLOv9	0.9222	0.70664	0.91491
YOLOv10	0.90593	0.6825	0.89035
YOLOv8_ MAH	0.92932	0.71775	0.89915

By comparing the results of different YOLO models, it is observed that YOLOv8_MAH achieves the best detection accuracy. Specifically, YOLOv5 achieves mAP50 = 0.92151, mAP50-95 = 0.71058, Precision = 0.90403; YOLOv8 achieves mAP50 = 0.92425, mAP50-95 = 0.72112, Precision = 0.90221; YOLOv11 achieves mAP50 = 0.92347, mAP50-95 = 0.60071, Precision = 0.91603; YOLOv9 achieves mAP50 = 0.92220, mAP50-95 = 0.70664, Precision = 0.91491; and YOLOv10 achieves mAP50 = 0.90593, mAP50-95 = 0.68250, Precision = 0.89035. From the experimental results, it can be concluded that the improved YOLOv8_MAH model, with mAP50 = 0.92932, mAP50-95 = 0.71775, and Precision = 0.89915, outperforms the baseline models in both overall detection accuracy and robustness.

To better illustrate the optimization process of the proposed model, we further visualized the training dynamics, including loss convergence and evaluation metrics over epochs, as shown in Figure 6.

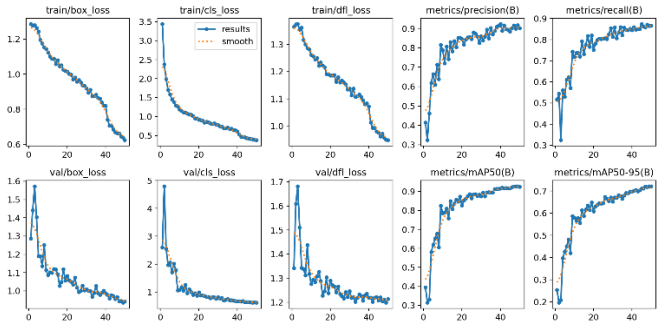

Figure 6 Training Iteration Curve

The following section presents comparative experiments across multiple YOLO versions to quantitatively evaluate the proposed improvements and the result of training loss is shown in table 2.

Table 2 Experimental Results of Training Loss

Model	Recall	Train_Loss	Val_Loss
YOLOv5	0.87765	0.66518	0.95083
YOLOv8	0.86595	0.62414	0.94234
YOLOv11	0.877	0.89455	1.37169
YOLOv9	0.8771	0.62963	0.97373
YOLOv10	0.84204	1.34783	2.00352
YOLOv8_MAH	0.89767	0.67916	0.9408

The results show that the YOLOv8_MAH model can maintain a relatively low training loss : Train_Loss = 0.67916, validation loss : Val_Loss = 0.94080, and a high recall : Recall = 0.89767. Compared with other YOLO models, the improved model demonstrates superior performance in detection sensitivity and training stability as shown in figure 7.

Figure 7 Prediction Results of YOLOv8-MAH

Next, to further illustrate the model's attention regions, we performed a heatmap-based visualization of the detected vehicles[8]. First, the entire image was converted into a grayscale background. Then, based on the bounding boxes predicted by the YOLO model, the vehicle regions were extracted. We applied the Sobel operator to compute the gradient magnitude in order to capture edge features, and these edge features were mapped to the Jet color space to generate the heatmap. Finally, the generated heatmap was overlaid onto the grayscale background, making the vehicle regions stand out in color while the background remained desaturated in gray. This visualization method allows us to intuitively observe the model's focus on vehicle regions within the image. The specific results are shown in the figures below as shown in figure 8.

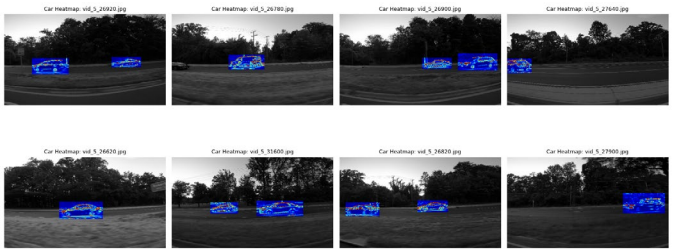

Figure 8 Results of Heatmap-Based Edge Extraction

This set of images demonstrates the detection performance of the improved model on vehicle targets in various road scenarios. It can be observed that whether the vehicles are at different distances, partially occluded by trees, or under varying lighting conditions, the model is able to consistently recognize and accurately localize the vehicle regions. The detection boxes maintain good precision in both shape and position, indicating that the model has strong capabilities in feature extraction and target localization. [9]The overall detection results are clear, with no obvious cases of missed or false detections.

VI. CONCLUSION

In this study, we proposed an improved YOLOv8-based vehicle detection model, namely YOLOv8-MAH, which integrates adaptive Mosaic augmentation, an Attention-Guided Feature Enhancement module, and heatmap-based interpretability. Experimental results across multiple YOLO variants (YOLOv5, YOLOv8, YOLOv9, YOLOv10, YOLOv11, and YOLOv8_MAH) demonstrated that the proposed model achieves superior performance, with higher detection accuracy, robustness, and interpretability compared to baseline methods.

For future work, we plan to extend our approach in two main directions: (1) expanding the dataset to include more diverse traffic conditions, weather variations, and night-time scenarios to further improve model generalization; and (2) exploring lightweight model compression and deployment strategies (e.g., pruning and quantization) to enable real-time applications on edge devices such as intelligent cameras and embedded systems.

REFERENCES

[1] A. Karasmanoglou, M. Antonakakis, M. Zervakis, *et al.*, "Heatmap-based Explanation of YOLOv5 Object Detection with Layer-wise Relevance Propagation," in *Proc. IEEE IST*, 2022, pp. 1–6, doi: 10.1109/IST55454.2022.9827744.

[2] A. Kirchknopf, D. Slijepcevic, I. Wunderlich, M. Breiter, J. Traxler, and M. Zeppelzauer, "Explaining YOLO: Leveraging Grad-CAM to Explain Object Detections," *arXiv preprint*, arXiv:2211.12108, 2022.

[3] F. He, D. Zhan, and K. Sunat, "HE-YOLOv8: a heatmap-enhanced and modularly optimized detector for UAV-based small object detection," in *Proc. SPIE 13730, International Conference on Unmanned Systems, Data Science and Artificial Intelligence*, 2025, Art. no. 1373004, doi: 10.1117/12.3072005.

[4] L. P. T. Nguyen, H. T. T. Nguyen, and H. Cao, "ODExAI: A Comprehensive Object Detection Explainable AI Evaluation," *arXiv preprint*, arXiv:2504.19249, 2025.

[5] M. A. Keyvanrad, *et al.*, "Explaining What Machines See: XAI Strategies in Deep Object Detection Models," *arXiv preprint*, arXiv:2509.01991, 2025.

[6] J. Kim, Y. Kim, and D. Kum, "Low-level Sensor Fusion for 3D Vehicle Detection using Radar Range-Azimuth Heatmap and Monocular Image," in *Proc. Asian Conf. on Computer Vision (ACCV)*, 2020, pp. 1–17.

[7] "A Novel Hybrid XAI Solution for Autonomous Vehicles: Real-Time Interpretability Through LIME–SHAP Integration," *Sensors*, vol. 24, no. 21, p. 6776, 2024, doi: 10.3390/s24216776.

[8] Y. Wang, S. Xu, P. Wang, L. Liu, Y. Li, and Z. Song, "Vehicle Detection Algorithm Based on Improved RT-DETR," *The Journal of Supercomputing*, vol. 81, p. 290, 2025, doi: 10.1007/s11227-024-06766-7.

[9] Z. Shao, K. He, B. Yuan, and S. Xu, "Enhanced YOLOv8 Framework for Precision Vehicle Detection in High-Resolution Remote Sensing Images," *Signal, Image and Video Processing*, vol. 19, p. 218, 2025, doi: 10.1007/s11760-024-03783-0.