Real-Time Pedestrian and Non-Motor Vehicle Detection with Enhanced YOLOv5 Algorithm

Jinhao Xu

Henan Province Intelligent Transportation Video Image Perception and Recognition Engineering Technology Research Center

Abstract: ***This paper presents an improved YOLOv5 model for the real-time detection of pedestrians and non-motorized vehicles, addressing critical challenges in complex traffic scenarios such as small-object missed detection and occluded-object false alarms. The proposed enhancements focus on three core components: firstly, an attention mechanism is incorporated to augment the backbone network's feature extraction capabilities; secondly, the neck's feature fusion architecture is optimized for more effective multi-scale aggregation; and lastly, the loss function is improved to accelerate convergence and enhance localization accuracy. Experiments conducted on a self-collected dataset and public benchmarks demonstrate that our model achieves a higher mean Average Precision (mAP) while retaining a high inference speed (FPS), thereby providing a reliable and efficient visual perception solution for intelligent transportation and autonomous driving systems.***

Keywords: Object detection; YOLOv5; Real-time recognition; Intelligent transportation.

1. INTRODUCTION

This paper aims to make targeted improvements to the YOLOv5 model. The main contributions include: introducing the CBAM attention mechanism to enhance feature extraction capabilities and suppress background noise; designing a bidirectional feature pyramid network (BiFPN) to strengthen multi-scale feature fusion and improve small target detection performance; replacing the original loss with the Focal-EIOU loss function to accelerate convergence and improve localization accuracy. Through systematic experiments on self-built and public datasets, the improved model is verified to significantly enhance detection accuracy while maintaining a high frame rate. Zhang et al.'s (2025) deep neural network approach to public data assets and data-driven decision models [1]. Healthcare technology advances through Wei et al.'s (2025) AI-driven intelligent health management systems in telemedicine [2], while computer vision applications in industrial settings include Zheng, Zhou, and Lu's (2023) improved YOLOv5s algorithm for rebar cross-section detection [3] and Zhao, Zhang, and Hu's (2023) Res2Net-YOLACT+HSV approach for smart warehouse track identification [4]. Robotics and sensing technologies progress with Xu's (2025) machine learning-enhanced fingertip tactile sensing [5], while neural network optimization is advanced by Wu et al.'s (2023) Jump-GRS structured pruning method for neural decoding [6]. Security frameworks are strengthened by Miao et al.'s (2025) authentication protocol for AI-based IoT supply chain systems [7], and robotics research expands through Guo and Tao's (2025) modeling of robot environmental interaction [8]. Software architecture innovations include Zhou's (2025) performance monitoring in microservices architecture [9], complemented by data security through Zhang's (2025) blockchain-based medical data sharing [10]. Analytical methodologies advance through Yu's (2025) Python applications in market analysis [11] and Liu's (2025) digital marketing optimization based on 4P theory [12]. Sports technology features Ren, Ren, and Lyu's (2025) IoT-based 3D pose estimation for athletes [13], while urban management benefits from Zhou et al.'s (2024) optimized garbage recognition model [14]. Information retrieval systems are enhanced by Jin et al.'s (2025) Rankflow workflow using large language models [15], and computational efficiency advances through Xie et al.'s (2024) RTop-K selection for neural acceleration [16]. Educational technology progresses with Yang's (2024) computer-assisted communicative competence training [17], while AI architectures are refined through Chen et al.'s (2024) decoupled-head attention learning [18]. Business intelligence innovations include Tian et al.'s (2025) cross-attention multi-task learning for ad recall [19], and economic applications feature Tang, Yu, and Liu's (2025) supply chain coordination research [20]. Materials science advances through Zhang and Needleman's (2020) stress-strain response identification [21], while recruitment technology evolves with Li et al.'s (2025) GPT and graph neural networks for resume-job matching [22]. Time-series analysis progresses through Su et al.'s (2025) WaveLST-Trans model for financial anomaly detection [23] and Zhang et al.'s (2025) MamNet for network traffic forecasting [24], complemented by Zhang, Li, and Li's (2025) deep learning for carbon market forecasting [25], with domain adaptation advancing through Peng et al.'s (2023) RAIN framework for black-box adaptation [26].

2. 1 RELATED WORK AND THEORETICAL BASIS

2.1 Overview of Object Detection Algorithms

Object detection is a core task in computer vision, aiming to locate specific objects in an image and identify their categories. Mainstream algorithms can be divided into two-stage and one-stage approaches, whose core ideas and performance characteristics differ significantly (as shown in Table 1). Two-stage detectors, represented by Faster R-CNN, first generate region proposals, then classify each proposal and regress its bounding box. These methods have complex architectures and high computational costs, resulting in slower detection speeds that struggle to meet real-time requirements, but their design mechanisms usually yield higher detection accuracy. In contrast, one-stage algorithms such as the YOLO family and SSD discard the region proposal step, performing classification and localization simultaneously via dense predictions on the feature map, greatly improving inference speed and making them more suitable for real-time systems. Although early versions lagged slightly in accuracy, subsequent iterations have continuously narrowed the gap through structural improvements.

Table 1: Comparison of Two-Stage and One-Stage Detection Algorithms

Feature	Two-stage algorithm (such as Faster R-CNN)	Single-stage algorithms (such as YOLO, SSD)
Testing Process	First generate candidate regions, then perform classification and regression	Predicting targets directly on the feature map
Speed	Slow	Quick
Precision	Tall	"Higher (continuously catching up)"
Representative Model	R-CNN series	YOLO series, SSD

2.2 YOLOv5 Algorithm Principles

YOLOv5 is a key representative of the YOLO series, widely adopted for its excellent speed-accuracy trade-off and engineering friendliness. Its network architecture mainly comprises three parts: Backbone, Neck, and Head. The Backbone uses CSPDarknet; the Focus module slices the input image to reduce computation while preserving rich feature information, and the CSP (Cross Stage Partial) structure enhances gradient flow, improving the network's learning capacity and efficiency. The Neck employs PANet (Path Aggregation Network), combining FPN's top-down semantic transmission with PAN's bottom-up localization transmission to achieve multi-level feature fusion, effectively aggregating features of different scales and enhancing detection of objects of varying sizes. The Head performs the final prediction on the fused feature map, outputting object classes, confidence scores, and bounding-box coordinates.

2.3 Attention Mechanism

The attention mechanism originates from the human visual system; its core idea is to enable neural networks to autonomously focus on more important information in the input while suppressing irrelevant parts, thereby utilizing computational resources efficiently and enhancing model expressiveness. In convolutional neural networks, attention is usually embedded at different depths as plug-and-play modules to improve feature representation quality. The Convolutional Block Attention Module (CBAM) is a lightweight yet effective attention module that sequentially derives attention maps from channel and spatial dimensions. The channel attention module first performs global average pooling and max pooling on the input feature map, then processes the two results through a shared-parameter multilayer perceptron (MLP), adds them, and applies a Sigmoid activation to generate a channel weight vector, whose simplified expression is:

$$\mathrm{M_c(F)} = \sigma\Big(\mathrm{MLP}\big(\mathrm{AvgPool(F)}\big) + \mathrm{MLP}\big(\mathrm{MaxPool(F)}\big)\Big) \tag{1}$$

Spatial attention focuses on which locations in the feature map are important. It first performs average-pooling and max-pooling at the same spatial position along the channel dimension and concatenates the results, then uses a 7 × 7 convolution layer followed by a Sigmoid function to generate a spatial weight map; the computation can be simplified as:

$$M_s(F) = \sigma\left(f_{7x7}([AvgPool(F); MaxPool(F)])\right) \tag{2}$$

Finally, the input features are multiplied sequentially by the channel and spatial attention maps to achieve adaptive feature refinement. Embedding CBAM into the backbone can markedly enhance the model's ability to extract features for key objects such as pedestrians and non-motorized vehicles.

3. IMPROVED YOLOV5 ALGORITHM DESIGN

3.1 Overall Improvement Framework

To address the challenges of pedestrian and non-motorized vehicle detection in complex road scenes, this paper improves the YOLOv5s model in multiple aspects. The overall network structure after improvement is shown in Table 2; its core Backbone, Neck, and Head frameworks are retained to preserve the algorithm's main strengths. The main innovations lie in the introduction and replacement of three key modules:

(1) Embedding a CBAM attention module after every CSP block in the Backbone to strengthen the model's ability to extract critical features;

(2) Replacing the feature fusion network in the Neck from the original PANet to the more efficient weighted bidirectional feature pyramid network (BiFPN) to improve multi-scale fusion, especially for small objects;

(3) Replacing the bounding-box regression loss from CIOU Loss to Focal-EIOU Loss to optimize the training process and improve localization accuracy.

Table 2: Improved YOLOv5 Network Structure

Module	Improvement points	Effect	Applicable scenarios
Backbone	CSP module+CBAM	Enhance key feature extraction and suppress background interference	Complex road scene
Neck	PANet → BiFPN	Optimize multi-scale feature fusion to enhance small target detection	Small targets such as pedestrians and non-motorized vehicles
Head	CIOU → Focal-EIOU	Improve localization accuracy and alleviate sample imbalance	Dense object detection

3.2 Introduction of CBAM Attention Mechanism

The original YOLOv5 model treats all information on the feature map equally; in traffic scenes with complex backgrounds and dense targets, it is easily disturbed by a large amount of irrelevant information, resulting in insufficient feature learning for pedestrians and non-motorized vehicles. To solve this problem, the Convolutional Block Attention Module (CBAM) is introduced, motivated by enabling the network to "focus" on more important feature channels and spatial locations, adaptively emphasizing key features of pedestrians and non-motorized vehicles while suppressing background noise.

The specific integration method is: after each CSP_1, CSP_2, and CSP_3 module in the Backbone network i.e., at the point where preliminary feature extraction is completed a CBAM module is inserted in sequence. This module receives the output feature map from the previous layer and successively computes through the channel-attention submodule and the spatial-attention submodule, with the expected effect of giving the model stronger feature-discrimination capability and robustness when facing partial occlusion, illumination changes, and complex backgrounds.

3.3 Integration of Bidirectional Feature Pyramid Network (BiFPN)

The original model adopts PANet, which achieves good multi-scale feature fusion; however, when fusing input features of different scales, it does not distinguish their importance and assumes by default that all inputs contribute equally to the output. This may cause certain features that are more important for the current prediction (e.g., finer small-object features) to be diluted during fusion.

BiFPN solves this problem by introducing learnable weights, whose core is a weighted bidirectional cross-scale connection mechanism. This approach allows the network to assign different importance weights to different input features.

Replace the PANet in the original Neck entirely with the BiFPN structure. This replacement shifts the paradigm from “indiscriminate fusion” to “weighted fusion,” significantly improving detection performance for small-scale pedestrians and non-motorized vehicles while maintaining high runtime efficiency.

3.4 Optimization of the Focal-EIOU Loss Function

CIOU Loss is the native localization loss of YOLOv5. Although it comprehensively considers overlap area, center-point distance, and aspect ratio, the definition of its aspect-ratio term v can sometimes slow convergence in the later training stages or even cause inaccurate regression. Moreover, in object detection tasks, the number of easy samples far exceeds that of hard samples; CIOU Loss assigns excessive loss contribution to easy samples, thereby drowning out the gradients of hard samples and hindering model optimization.

To address these issues, Focal-EIOU Loss can be adopted as the new localization loss. This loss is first based on EIOU Loss, which decomposes the aspect-ratio term in CIOU Loss into separate regressions for the target’s width and height, formally defined as:

$$\begin{aligned} L_{EIOU} &= L_{IOU} + L_{dis} + L_{asp} \\ &= 1 - IoU + \frac{\rho^2(b,b_{gt})}{(c_w)^2+(c_h)^2} + \frac{\rho^2(w,w_{gt})}{(c_w)^2} + \frac{\rho^2(h,h_{gt})}{(c_h)^2} \end{aligned} \tag{3}$$

This definition makes the convergence process smoother and more stable, yielding higher localization accuracy.

4. EXPERIMENTS AND RESULTS ANALYSIS

4.1 Experimental Environment and Dataset

The experiments were conducted on the Ubuntu 20.04 operating system, using Python 3.8 and the PyTorch 1.10.0 deep-learning framework. The hardware platform was an NVIDIA GeForce RTX 3090 GPU (24 GB VRAM) with CUDA 11.3 acceleration. The study employed the MS COCO dataset for model pre-training and a large-scale, self-constructed dataset for the main training and testing. This custom dataset was collected from surveillance footage of multiple urban roads, comprising 10,000 high-quality images and over 35,000 annotated instances. Target classes include pedestrians, bicycles, and electric scooters. All images were meticulously labeled with the LabelImg tool and subjected to rigorous cleaning and verification to ensure data quality. During training, data-augmentation strategies such as Mosaic, random flipping, and color jittering were applied, effectively improving the model’s generalization and robustness.

4.2 Evaluation Metrics

This study employs multiple authoritative metrics for a comprehensive evaluation of model performance. For accuracy, the mean Average Precision (mAP) is used, where mAP@0.5 denotes the result at an Intersection over Union threshold of 0.5, and mAP@0.5: 0.95 represents the average precision across different IoU thresholds. Recall measures the model’s ability to detect true positives. Speed performance is quantified by FPS,... i.e., the number of image frames the model can process per second.

4.3 Ablation Study

To verify the effectiveness of each proposed module, we conducted systematic ablation experiments. The results show that the baseline YOLOv5s achieves an mAP@0.5 of 86.2%. Adding the CBAM module alone raises mAP to 87.8%, demonstrating that the attention mechanism effectively enhances feature selectivity. Introducing BiFPN yields the most significant performance gain, with mAP reaching 88.5%, highlighting the importance of weighted feature fusion. Adopting the Focal-EIOU loss function further improves mAP by 0.8%. When all modules are integrated, the complete model attains 90.5% mAP, a 4.3% improvement over the baseline, while maintaining a high FPS of 115, fully satisfying real-time detection requirements.

4.4 Comparative Experiments

Comparative experiments between the improved model proposed in this paper and current mainstream object detection algorithms show that the traditional two-stage algorithm Faster R-CNN achieves high accuracy but is slow, reaching only 15 FPS. Among single-stage algorithms, SSD300 and YOLOv3-tiny have speed advantages but insufficient detection accuracy. YOLOv4 attains 88.9% mAP, yet its computational cost is high. The improved YOLOv5s model in this paper reaches 90.5% on the mAP@0.5 metric, significantly outperforming all comparison models, while maintaining a real-time processing speed of 115FPS , achieving the best balance between accuracy and speed with only a slight increase in parameter count.

4.5 Visualization Results Analysis

Through visual comparative analysis, the effectiveness of the improved model is demonstrated intuitively. In complex-scene test images, the original YOLOv5 model exhibits obvious missed detections of small objects and difficulty recognizing occluded targets. The improved model not only accurately detects distant small-scale pedestrians and non-motorized vehicles but also shows stronger recognition capability for partially occluded targets. Meanwhile, the improved model significantly reduces false detections of background interference as targets, yielding more reliable detection results.

5. CONCLUSIONS AND OUTLOOK

5.1 Summary of Research Work

This study systematically improves the YOLOv5 model to meet the practical needs of pedestrian and non-motorized vehicle detection in complex road scenes. By introducing the CBAM attention mechanism, BiFPN feature fusion network, and Focal-EIOU loss function, the model’s detection performance is effectively enhanced. Experimental results show that the improved model achieves an mAP@0.5 of 90.5% on the self-built dataset, a significant 4.3% improvement over the original YOLOv5s baseline. It also maintains a high processing speed of 115 FPS, achieving a good balance between accuracy and speed, thereby validating the effectiveness of the proposed improvement strategies.

5.2 Main Innovations

The main innovations of this paper are reflected in three aspects: First, combining the CBAM attention mechanism with the YOLOv5 model enhances the model’s ability to focus on key features and effectively suppresses complex background interference. Second, replacing the original feature fusion network with the BiFPN structure significantly improves detection performance for small-scale targets through a weighted bidirectional fusion mechanism. Third, introducing the Focal-EIOU loss function to optimize the bounding-box regression process not only accelerates model convergence but also improves localization accuracy.

5.3 Future Work Outlook

Future research will proceed in the following directions: First, exploring lightweight techniques such as model pruning and quantization to further improve model efficiency and meet deployment requirements on embedded devices like in-vehicle terminals. Second, extending the current 2D detection to 3D spatial perception and multi-object tracking to obtain richer environmental information. Third, conducting in-depth research on enhancing model robustness under adverse conditions such as extreme weather and low illumination. Finally, exploring domain-adaptive methods to enable the model to quickly adapt to new environments.

PROJECT SUPPORT

Pedestrian and Non-motorized Vehicle Recognition Algorithm Based on Deep Learning (ETRC-240204).

REFERENCES

[1] Zhang, Shiwen, et al. "Optimizing the Operation Mechanism of Public Data Assets and Data-Driven Decision Models in the Digital Economy With Deep Neural Networks." Journal of Organizational and End User Computing (JOEUC) 37.1 (2025): 1-22.

[2] Wei, Xiangang, et al. "AI driven intelligent health management systems in telemedicine: An applied research study." Journal of Computer Science and Frontier Technologies 1.2 (2025): 78-86.

[3] Zheng, Y., Zhou, G., & Lu, B. (2023). Rebar Cross-section Detection Based on Improved YOLOv5s Algorithm. Innovation & Technology Advances, 1(1), 1–6. https://doi.org/10.61187/ita.v1i1.1

[4] Zhao, X., Zhang, L., & Hu, Z. (2023). Smart warehouse track identification based on Res2Net-YOLACT+HSV. Innovation & Technology Advances, 1(1), 7–11. https://doi.org/10.61187/ita.v1i1.2

[5] Xu, Zhongjin. "Machine Learning-Enhanced Fingertip Tactile Sensing: From Contact Estimation to Reconstruction." Journal of Intelligence Technology and Innovation (JITI) 3.2 (2025): 20-39.

[6] Wu, Xiaomin, et al. "Jump-GRS: a multi-phase approach to structured pruning of neural networks for neural decoding." Journal of neural engineering 20.4 (2023): 046020.

[7] Miao, Junfeng, et al. "Secure and efficient authentication protocol for supply chain systems in artificial intelligence-based Internet of Things." IEEE Internet of Things Journal (2025).

[8] Guo, Y., & Tao, D. (2025). Modeling and Simulation Analysis of Robot Environmental Interaction. Artificial Intelligence Technology Research, 2(8).

[9] Zhou, Z. (2025). Research on Software Performance Monitoring and Optimization Strategies in Microservices Architecture. Artificial Intelligence Technology Research, 2(9).

[10] Zhang, T. (2025). Research and Application of Blockchain-Based Medical Data Security Sharing Technology. Artificial Intelligence Technology Research, 2(9).

[11] Yu, Z. (2025). Advanced Applications of Python in Market Trend Analysis Research. MODERN ECONOMICS, 6(1), 115.

[12] Liu, Huanyu. "Research on Digital Marketing Strategy Optimization Based on 4P Theory and Its Empirical Analysis."

[13] Ren, Fei, Chao Ren, and Tianyi Lyu. "Iot-based 3d pose estimation and motion optimization for athletes: Application of c3d and openpose." Alexandria Engineering Journal 115 (2025): 210-221.

[14] Zhou, Y., Wang, Z., Zheng, S., Zhou, L., Dai, L., Luo, H., ... & Sui, M. (2024). Optimization of automated garbage recognition model based on resnet-50 and weakly supervised cnn for sustainable urban development. Alexandria Engineering Journal, 108, 415-427.

[15] Jin, Can, et al. "Rankflow: A multi-role collaborative reranking workflow utilizing large language models." Companion Proceedings of the ACM on Web Conference 2025. 2025.

[16] Xie, Xi, et al. "RTop-K: Ultra-Fast Row-Wise Top-K Selection for Neural Network Acceleration on GPUs." The Thirteenth International Conference on Learning Representations. 2024.

[17] Yang, C. (2024). A Study of Computer-Assisted Communicative Competence Training Methods in Cross-Cultural English Teaching. Applied Mathematics and Nonlinear Sciences, 9(1). Scopus. https://doi.org/10.2478/amns-2024-2895

[18] Chen, Yilong, et al. "Dha: Learning decoupled-head attention from transformer checkpoints via adaptive heads fusion." Advances in Neural Information Processing Systems 37 (2024): 45879-45913.

[19] Q. Tian, D. Zou, Y. Han and X. Li, "A Business Intelligence Innovative Approach to Ad Recall: Cross-Attention Multi-Task Learning for Digital Advertising," 2025 IEEE 6th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT), Shenzhen, China, 2025, pp. 1249-1253, doi: 10.1109/AINIT65432.2025.11035473.

[20] Tang, H., Yu, Z., & Liu, H. (2025). Supply Chain Coordination with Dynamic Pricing Advertising and Consumer Welfare An Economic Application. Journal of Industrial Engineering and Applied Science, 3(5), 1–6.

[21] Zhang, Yupeng, and Alan Needleman. "Influence of assumed strain hardening relation on plastic stress-strain response identification from conical indentation." Journal of Engineering Materials and Technology 142.3 (2020): 031002.

[22] Li, Huaxu, et al. "Enhancing Intelligent Recruitment With Generative Pretrained Transformer and Hierarchical Graph Neural Networks: Optimizing Resume-Job Matching With Deep Learning and Graph-Based Modeling." Journal of Organizational and End User Computing (JOEUC) 37.1 (2025): 1-24.

[23] Su, Tian, et al. "Anomaly Detection and Risk Early Warning System for Financial Time Series Based on the WaveLST-Trans Model." (2025).

[24] Zhang, Yujun, et al. "MamNet: A Novel Hybrid Model for Time-Series Forecasting and Frequency Pattern Analysis in Network Traffic." arXiv preprint arXiv:2507.00304 (2025).

[25] Zhang, Zongzhen, Qianwei Li, and Runlong Li. "Leveraging Deep Learning for Carbon Market Price Forecasting and Risk Evaluation in Green Finance Under Climate Change." Journal of Organizational and End User Computing (JOEUC) 37.1 (2025): 1-27.

[26] Peng, Qucheng, et al. "RAIN: regularization on input and network for black-box domain adaptation." Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. 2023.