Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 11, No. 1, 2025

A Vehicle Re-Identification Algorithm for Long-Distance Small Targets Combining YOLOv8 Object Detection Algorithm

Chenyu Gu *, Hong Du, Xiaozheng Zhang, Ying Wang, Zhonglin Yang, Gaotian Liu, Chuan Zhang, Lei Sun

China Northern Vehicle Research Institute, Beijing, China

*** Corresponding author:** Chenyu Gu

Abstract: With the rapid development of intelligent technologies, deep learning-based object detection has been widely applied in dynamic fields, especially in vehicle management. However, challenges in accurate recognition and tracking remain, particularly due to issues like intra-class variations and inter-class similarities in vehicle re-identification across camera viewpoints. This paper proposes a method for long-range small vehicle target re-identification based on the YOLOv8 object detection algorithm, aiming to improve the detection accuracy and re-identification performance of small targets in complex environments. First, a detection and re-identification dataset containing long-range vehicle images is constructed based on a complex background and small target dataset. Secondly, the YOLOv8-EMA-RFB optimization algorithm is introduced, which combines the EMA (Exponential Moving Average) attention mechanism and RFB (Receptive Field Block) structure. The EMA module enhances feature extraction for small targets and reduces noise interference, while the RFB module increases the receptive field, improving the detection capability for small targets. Through these optimizations, the proposed method effectively improves the detection accuracy of long-range small targets in the YOLOv8 model. Additionally, by incorporating the AlignedReID method, feature matching in object re-identification is improved, further enhancing the accuracy and robustness of multi-object tracking. Experimental results demonstrate that the proposed method achieves high precision and real-time performance in multi-scene vehicle detection and re-identification, offering a novel solution for intelligent transportation and urban security surveillance.

Keywords: Vehicle Re-identification; Object Detection; Attention Mechanism.

1. Introduction

With the rapid development of intelligence, deep learning-based object detection has been applied in dynamic fields of view. Its object detection capabilities can be used to generate target trajectories, aiding target management tasks, thereby addressing vehicle control issues in different scenarios. With the emergence of large-scale vehicle datasets and deep learning-based methods, significant progress has been made in vehicle re-identification tasks. However, due to the complex viewpoint transformations across cameras and two main challenges in vehicle re-identification—intra-class variation and inter-class similarity—vehicle re-identification tasks remain challenging [1].

Generally, vehicle detection algorithms perform well in static scenes, but maintaining unique target identification numbers for the same target becomes difficult under cross-camera conditions or when affected by viewpoint changes, occlusion, and other factors.

The main challenges faced by vehicle detection and re-identification algorithms under different fields of view include the short-term disappearance of targets due to field-of-view changes, the impact of viewpoint changes on feature matching, and the difficulty of extracting features from small targets. These challenges collectively constrain the performance of vehicle detection and re-identification.

In the early stages, researchers focused on extracting global or appearance features, such as vehicle windows, headlights, interior, brand logos, and other distinguishable features. However, due to the issue of inter-class similarity, this approach has certain limitations. Subsequently, methods utilizing local features to improve the accuracy of vehicle re-identification tasks have been widely proposed. For example, reference [2] uses object detection methods to extract local regions such as headlights, inspection stickers, and decorations from vehicle images, and uses these local parts for vehicle re-identification. To better focus on local regions, references [3-5] designed methods to divide feature maps along the horizontal and vertical dimensions to extract fine-grained local features.

However, in complex traffic environments, due to factors such as camera viewpoint changes and lighting, the features extracted by the above methods are often ineffective for completing vehicle re-identification tasks. This paper aims to optimize vehicle detection and re-identification algorithms in such scenarios.

2. Optimization of Object Detectors for Long-Distance Small Vehicle Targets

2.1. Design of Vehicle Target Detection Algorithm Based on YOLOv8

YOLOv8 model is one of the latest generation of YOLO series algorithms introduced by Ultralytics, which contains five versions from small to large: YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x. The detection accuracy of the model is progressively improved with the increase of the size of the model, so that a suitable YOLOv8

model version can be selected according to the specific task requirements. version of the YOLOv8 model.YOLOv8 performs well in tasks such as target detection, instance segmentation and target classification [17].The structure of the YOLOv8 model is shown in Fig.1.

The YOLOv8 model consists of three main parts: the Backbone network (Backbone), the Neck network (Neck), and the detection head (Head).

The Backbone network of YOLOv8 is responsible for extracting features from the input image and transforming the image into a feature representation with rich semantic information.YOLOv8n adopts Darknet-53 as the Backbone network and introduces the C2f module for residual learning. Compared with the C3 module in YOLOv5, the C2f module has better feature extraction capability while maintaining a smaller number of parameters. Specifically, the C2f module adopts a cross-convolutional structure combined with BottleneckBlock and SPPF (SpatialPyramidPoolingFast) modules to enhance the feature extraction capability. This structural design not only reduces redundant parameters, but also enhances the computational efficiency. In the backbone network, the Conv convolutional module and the C2f module are stacked in tandem four times, and each stack is called a stage. This design allows YOLOv8n to incorporate richer global features while remaining lightweight. Finally, the different feature layers are encoded using the SPPF module, and the processed feature maps are fed into the neck network for further fusion and processing.

The neck network is simplified compared to YOLOv5 by removing the two convolutional fully-connected layers and adopting a PAN-FPN structure, which is responsible for multi-scale feature fusion. This structure enhances the feature representation capability by fusing feature maps from different stages of the backbone network. Specifically, the neck network contains the SPPF module, the PAA module, and the PAN module, which work together to enable the YOLOv8n to effectively utilize the features extracted from the backbone network to enhance the model's performance in the target detection task. With the PAN-FPN structure, the neck network achieves efficient multi-level feature fusion and enhances the detection of multi-scale targets, thus improving the overall model accuracy and robustness. This design implements top-down and bottom-up feature pyramids to fuse different features processed by the backbone network.

In the detection head part, YOLOv8 adopts the current mainstream Decoupled-Head structure instead of the Coupled-Head structure of YOLOv5. The Decoupled-Head handles the classification and detection tasks separately, and at the same time converts from an Anchor-Frame mechanism to an Anchor-Free mechanism (Anchor-Free), which improves the detection efficiency and accuracy. The detection head part is responsible for the final prediction tasks, including bounding box regression, target classification, and confidence prediction.The detection head of YOLOv8 consists of a convolutional layer, a global average pooling layer, and a loss function. The convolutional layer generates the detection results through a series of convolutional and deconvolutional operations, and is responsible for predicting the bounding box regression value and the confidence of target presence for each anchor box. The global average pooling layer enhances the model's ability to handle multi-category classification tasks by reducing the dimensionality of the feature map and outputting a probability distribution for each category. For the loss function part, YOLOv8 employs Task-AlignedAssigner positive and negative sample matching in the detection header and DistributionFocalLoss (DFL) to further enhance the performance of the model. Through these optimizations and improvements, YOLOv8 finally and effectively improves the accuracy and robustness of target detection. The application of the decoupled head structure and the anchorless mechanism makes the model more flexible and efficient in dealing with targets of different sizes and classes.

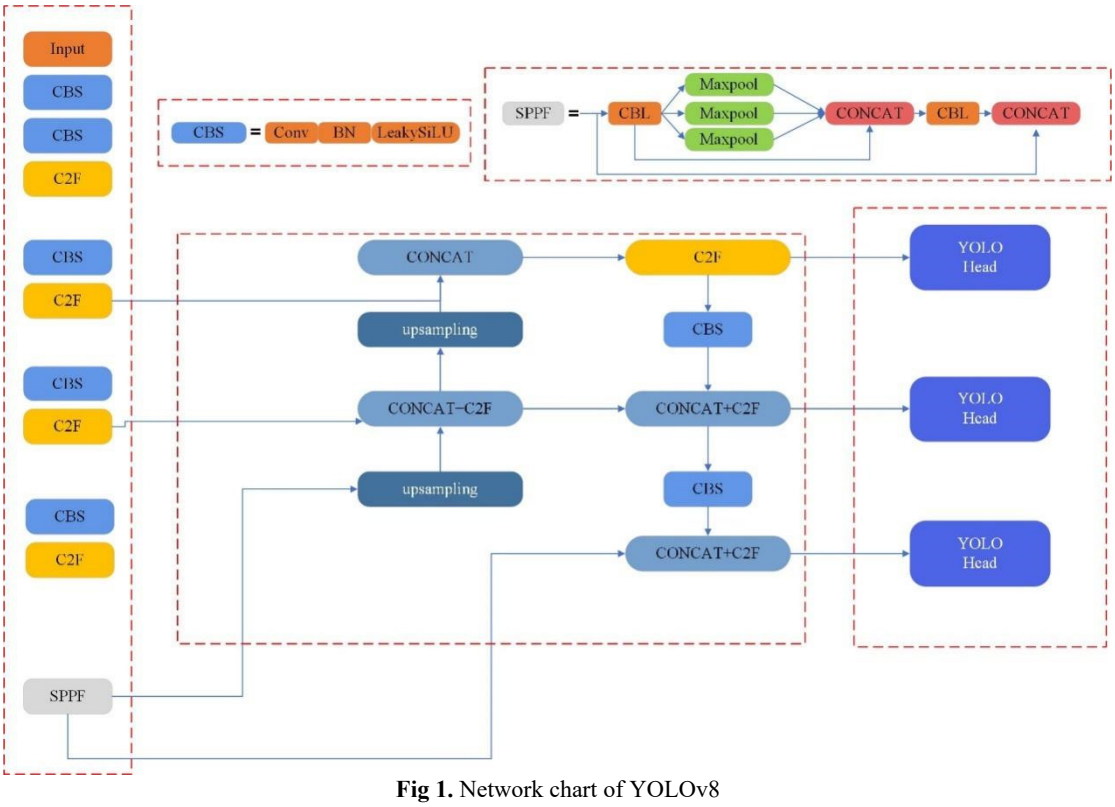

Fig 1. Network chart of YOLOv8

2.2. S Optimization of Object Detectors for Long-Distance Small Vehicle Targets

In vehicle re-identification tasks, the accuracy of object detectors significantly affects the performance of re-identification. Addressing the challenge of feature extraction for long-distance small targets, this study focuses on optimizing the YOLOv8 object detector. The proposed enhancements aim to improve the detection accuracy of small targets, thereby optimizing multi-object tracking algorithms.

The architecture of the improved YOLOv8-EMA-RFB object detection algorithm is illustrated in Fig. 2.

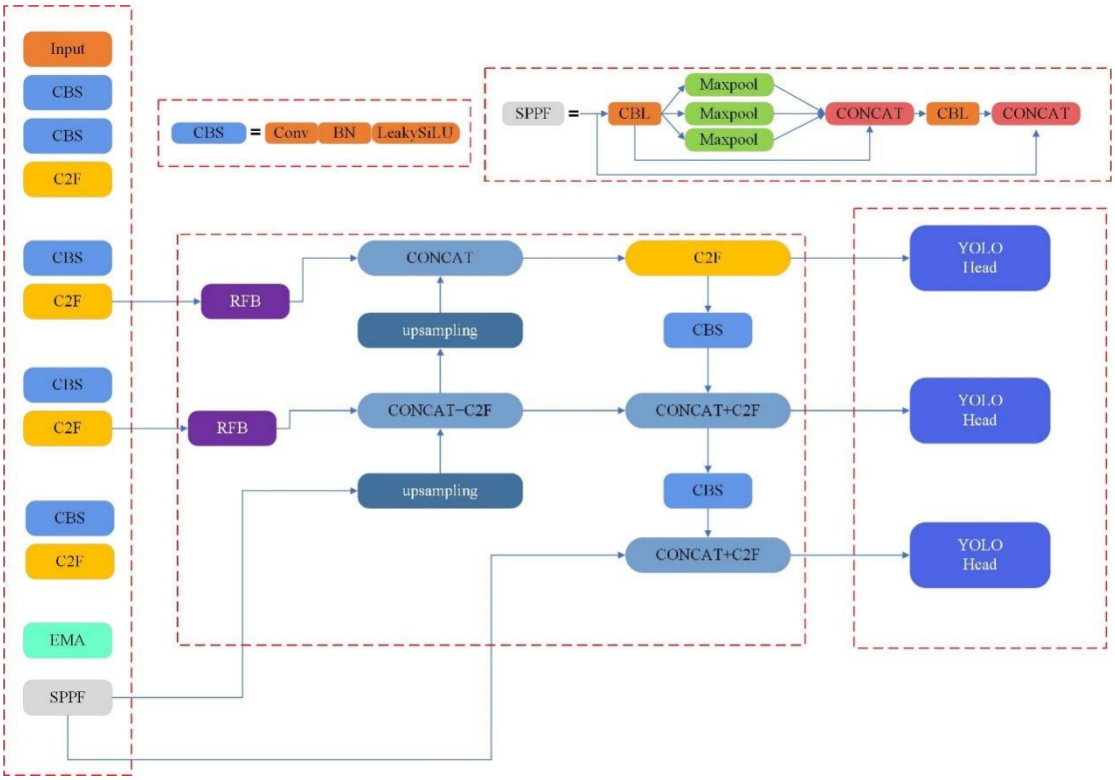

Fig 2. The Architecture Diagram of the YOLOv8-EMA-RFB Object Detection Algorithm

2.2.1. Integration of the Attention Module

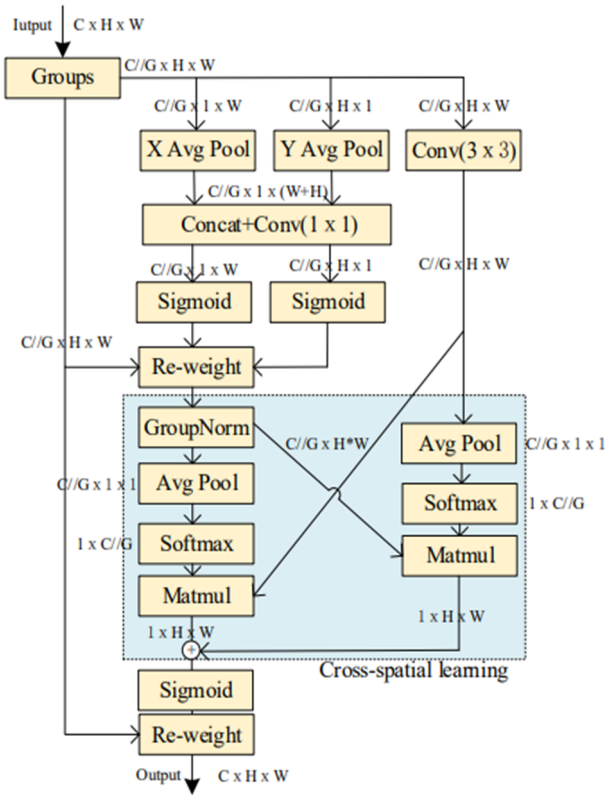

Fig 3. The overall structure of the EMA module

The EMA (Exponential Moving Average) attention module is added to the backbone of YOLOv8 to enhance feature representation and improve overall performance.

EMA utilizes exponential weighted moving averages to effectively focus on important features while smoothing noise, which enhances the ability to extract key features. Compared to other complex attention mechanisms, EMA has lower computational complexity, maintaining high computational efficiency while improving performance. EMA is adaptable to various visual tasks and enhances the model's ability to handle different tasks effectively. The overall structure of the EMA module is shown in Fig.3. In this section, we will discuss how EMA learns effective channel descriptions in convolutional operations without compressing the channel dimension, thus generating superior pixel-level attention for high-level feature maps.

Specifically, we extract the shared 1x1 convolution component from the CA (Coordinate Attention) module and refer to it as the 1x1 branch in EMA. To aggregate multi-scale spatial structure information, a 3x3 convolution kernel is introduced in the parallel path of the 1x1 branch for fast responses, referred to as the 3x3 branch. Considering feature grouping and multi-scale structure, this design helps efficiently establish both short-range and long-range dependencies, thereby improving performance.

For a given input feature map $X \in RC \times H \times W$, EMA divides X along the channel dimension into G sub-features to learn different semantics. This grouping can be expressed as $X=[X0,X1,\ldots,XG-1]$, where $Xi \in RC/G \times H \times W$. Without loss

of generality, we assume that G≪C, and the learned attention weight descriptors are used to enhance the feature representation of the regions of interest in each sub-feature.

The larger local receptive field of neurons allows them to collect multi-scale spatial information. Based on this, EMA designs three parallel paths to extract the attention weight descriptors for the grouped feature maps. Two of the parallel paths belong to the 1x1 branch, and the third path belongs to the 3x3 branch. To capture dependencies across all channels and reduce computational overhead, EMA models the cross-channel information interaction along the channel direction.

Specifically, in the 1x1 branch, two 1D global average pooling operations are used to encode the channels along two spatial directions. In the 3x3 branch, only one 3x3 convolution kernel is stacked to capture multi-scale feature representations.

Since the convolution function does not involve batch coefficients in its dimensions, the number of convolution kernels is independent of the batch coefficients in the forward pass. Therefore, EMA reshapes and arranges the GG groups into the batch dimension, redefining the input tensor as shape C//G×H×W.

On one hand, similar to CA, EMA concatenates the two encoded features along the height of the image and shares the same 1x1 convolution operation without performing dimensionality compression in the 1x1 branch. After decomposing the output of the 1x1 convolution into two vectors, two non-linear Sigmoid functions are applied to fit a two-dimensional binomial distribution based on linear convolution. To achieve different cross-channel interaction features between the two parallel paths of the 1x1 branch, EMA aggregates the two channel attention maps within each group by simple multiplication.

On the other hand, the 3x3 branch captures local cross-channel interactions via a 3x3 convolution to expand the feature space. In this way, EMA not only encodes cross-channel information to adjust the importance of different channels but also preserves precise spatial structure information within the channels.

Advantages of EMA Attention Mechanism for Small Object Detection:

1. Enhances sensitivity to small objects, improving the detection capability for tiny targets.

2. Smoothens the feature information of small objects, reducing noise interference.

2.2.2. YOLOv8 Model with Integrated RFB Structure

RFB (Receptive Field Block) is a network module designed to enhance feature extraction capabilities by increasing the receptive field, which improves the performance of convolutional neural networks (CNNs) in object detection tasks, further enhancing detection accuracy and speed.

RFB is a novel feature extraction module that simulates the receptive field of human vision to strengthen the network's feature extraction capability. Structurally, RFB is inspired by the multi-branch network architecture of Inception. It builds on Inception by incorporating dilated convolutions, where the dilation rate increases with the size of the convolution kernel, effectively expanding the receptive field. The RFB network structure, as shown in Fig.4, uses 1×1, 3×3, and 5×5 convolution kernels for feature extraction across three channels. It then applies dilation rates of 1, 3, and 5 to the corresponding 3×3 convolutions, respectively. The effective feature layers at different scales from the three branches are concatenated. Finally, a 1×1 convolution layer is applied across channels to perform residual connections with the input's effective feature layers, integrating different-sized receptive fields. This process expands feature information extraction, which further enhances detection accuracy and speed.

Advantages of the RFB Network for Small Object Detection:

1. Multi-Scale Feature Fusion: RFB extracts features using convolution kernels of different scales, allowing for better identification of small objects.

2. Enhanced Detail Capturing: By expanding the receptive field, RFB captures fine details of small objects, reducing the likelihood of missing detections.

3. Expanded Receptive Field: The use of dilated convolution increases the receptive field, enabling the network to better capture long-distance spatial semantic information.

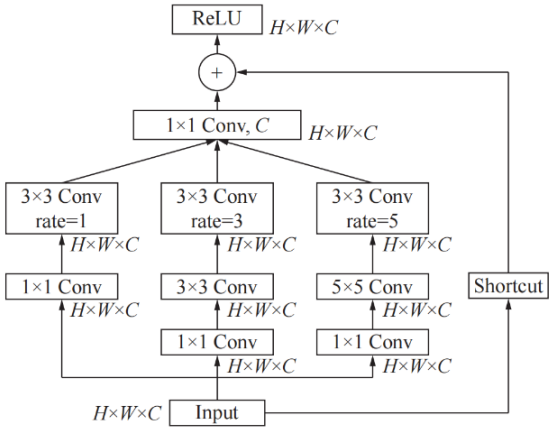

Fig 4. RFB Network Structure

3. Improvement and Optimization of the AlignedReID-based Object Re-identification Algorithm

To improve the accuracy of multi-target tracking algorithms integrating target re-identification on mobile platforms, this study addresses the challenges of feature matching in target re-identification by selecting AlignedReID as the target re-identification algorithm. Optimization of this algorithm is carried out to enhance the accuracy of re-identification for small targets at long distances.

AlignedReID is a deep learning-based target re-identification method designed to improve matching accuracy across different viewpoints. The core of this method lies in aligning feature maps from different images to capture critical discriminative details, thereby achieving stronger re-identification performance. Unlike traditional re-identification methods, AlignedReID dynamically aligns local features, making the identity features of an individual more accurate and thus improving the matching precision for small targets at long distances.

The Bottleneck Attention Mechanism (BAM) is a simple yet effective attention-based bottleneck structure that dynamically weights features based on their importance, enabling the bottleneck attention block to focus on significant discriminative features while ignoring redundant or irrelevant ones. The structure of BAM is shown in Fig. 5. In this study, the Bottleneck Attention Mechanism module is introduced into the AlignedReID re-identification network, enhancing the feature extraction capability through the incorporation of

channel attention.

The introduction of the Bottleneck Attention Mechanism into the AlignedReID algorithm brings several benefits:

Enhanced Feature Discriminability Small targets at long distances often suffer from low resolution, making their features blurry. By introducing the Bottleneck Attention Mechanism into the bottleneck blocks of the ResNet model, this module assigns different weights to each channel, automatically focusing on more discriminative features.

Suppression of Redundant Information: Background or irrelevant information in images of small targets at long distances can interfere with the extraction of target features. The Bottleneck Attention Mechanism dynamically adjusts attention weights across channels to suppress features that are unrelated to target recognition.

Improved Capturing of Small Target Details: The Bottleneck Attention Mechanism uses global pooling operations to generate channel descriptors, allowing the model to learn global context in the image and focus on detailed features. This is particularly important for small targets at long distances, as they often lack significant large-scale features and rely on finer details for differentiation.

Adaptation to Multi-Scale Feature Extraction: Small targets at long distances vary greatly in size, necessitating multi-scale feature extraction for effective recognition. The introduction of the Bottleneck Attention Mechanism enhances the model's ability to adapt to multi-scale information by automatically adjusting the attention weights across feature map channels, allowing the model to extract useful information from various scales.

Improved Robustness and Accuracy of the Network: The Bottleneck Attention Mechanism not only focuses on important features but also adaptively adjusts gradients during backpropagation to strengthen learning from critical channels, thus maintaining high precision in target recognition in complex environments.

In summary, introducing the Bottleneck Attention Mechanism into AlignedReID significantly improves the re-identification performance of small targets at long distances. This is achieved through enhanced feature discriminability, suppression of redundant information, better capture of detailed features, improved multi-scale adaptability, and increased robustness. These improvements allow the network to effectively perform small target re-identification even in the presence of challenges such as long distances, low resolution, and complex backgrounds.

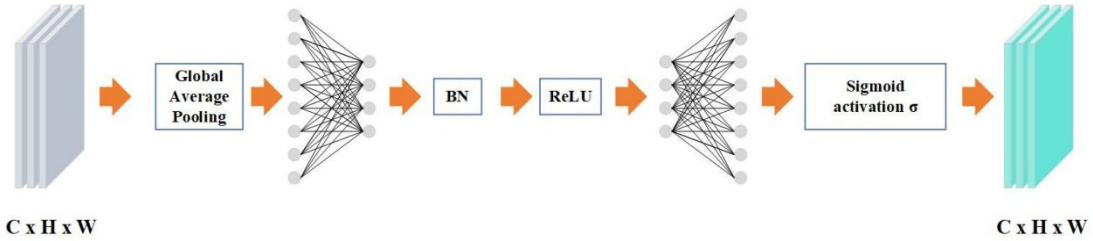

Fig 5. Bottleneck Attention Mechanism Module

4. Application Example Comparison

In terms of the experimental environment and configuration, we utilized a Windows-based software and hardware development environment, featuring a 13th-generation Intel Core i9-13900HX CPU, NVIDIA GeForce RTX 4090 GPU, and the PyTorch deep learning framework. Key experimental configurations included learning rate, batch size, and the number of iterations, all of which significantly influenced the performance of the algorithm.

Experimental Dataset and Evaluation Metrics

This study used publicly available object detection datasets for experimental validation. These datasets encompass images of targets in diverse scenarios, providing high representativeness and challenges. The main evaluation metrics included Mean Average Precision (mAP), Recall, F1 Score, Rank-1, and Rank-5, offering a comprehensive evaluation of the algorithm's performance.

4.1. Optimization Results of Object Detector Algorithm

As shown in the Table 1, the experimental results summarize the optimized YOLOv8-based object detection algorithm with integrated enhancements. The results demonstrate that the YOLOv8-EMA-RFB algorithm achieves higher mAP compared to other algorithms, with minimal changes in parameter count and frames per second (FPS). This indicates the effectiveness of the YOLOv8 optimization, contributing to improved accuracy in multi-target tracking.

Table 1. Optimized Results for the YOLOv8-Based Object Detection Algorithm Integration

YOLOv8	YOLOv8-EMA	YOLOv8-EMA-RFB	Params /M	Map@0.5	FPS/ frame. s-1	GFLOPS	Recall	F1
√			6.02	84.3	74	13.2	0.87	0.85
	√		8.99	86.8	71	15.6	0.89	0.88
		√	9.39	88.2	69	20.1	0.91	0.91

4.2. Optimization Results of the Object Re-Identification Algorithm

As shown in the table, a comparison is made between the original AlignedReID algorithm and the optimized AlignedReID-veh algorithm. By comparing the mean average precision (mAP) and key re-identification parameters, such as Rank-1 and Rank-5, before and after the addition of the Bottleneck Attention Mechanism, it is evident that the incorporation of the Bottleneck Attention Mechanism module significantly enhances the detection accuracy of the re-identification algorithm.

Table 2. Three Scheme comparing

	maP	Rank-1	Rank-5
AligenedReID	88.4	92.8	97.9
AligenedReID-veh	89.3	93.9	98.1

Therefore, regarding the optimization of the YOLOv8 object detection algorithm, experimental results validate that the YOLOv8-EMA-RFB object detection algorithm achieves higher mean average precision (mAP) with fewer parameters and lower frames per second (FPS), indicating that the algorithm optimization for YOLOv8 is effective.

At the same time, for object re-identification, experimental data also demonstrate that the introduction of the Bottleneck Attention Mechanism significantly improves the performance and efficiency of the ReID model. Specifically, the addition of the Bottleneck Attention Mechanism module enhances the mAP and Rank-1 accuracy across different datasets, accelerates the model's convergence speed, and reduces the model's parameter count. Additionally, the Bottleneck Attention Mechanism shows versatility and advantages across various backbone networks by dynamically adjusting channel weights, which enhances feature discrimination. Overall, the Bottleneck Attention Mechanism module strikes an excellent balance between accuracy and efficiency, making it a superior solution for attention mechanisms.

Acknowledgments

National Natural Science Foundation of China; National Science Foundation of China, Xingxin Li, 62206257.

References

[1] Jin Zifeng. Research on pedestrian detection and re-identification technology in surveillance video [D]. University of Chinese Academy of Sciences (National Center for Space Science, Chinese Academy of Sciences), 2021.DOI : 10.27562 / d.cnki.gkyyz.2021.000029.

[2] ZHANG X,ZHANG R,CAO J,et al.Part-guided attention learning for vehicle instance retrieval[J].IEEE Transactions on Intelligent Transportation Systems,2020,23(4):3048-3060.

[3] CHEN H,LAGADEC B,BREMOND F.Partition and reunion:A two-branch neural network for vehicle re-identification [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),June 16-20, 2019,Long Beach,USA.New York:IEEE,2019:184-192.

[4] QIAN J,JIANG W,LUO H,et al.Stripe-based and attribute-aware network:A two-branch deep model for vehicle re-identificationJ.Measurement Science and Technology, 2020,31 (9): 095401.

[5] WANG H,PENG J,JIANG G,et al.Discriminative feature and dictionary learning with part-aware model for vehicle re-identificationLJ.Neurocomputing,2021,438:55-62.