Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.063383

ARTICLE

A Detection Algorithm for Two-Wheeled Vehicles in Complex Scenarios Based on Semi-Supervised Learning

Mingen Zhong[1], Kaibo Yang[1,*], Ziji Xiao[1], Jiawei Tan[2], Kang Fan[2], Zhiying Deng[1] and Mengli Zhou[1]

[1]Fujian Key Laboratory of Advanced Bus & Coach Design and Manufacture, Xiamen University of Technology, Xiamen, 361024, China

[2]School of Aerospace Engineering, Xiamen University, Xiamen, 361102, China

*Corresponding Author: Kaibo Yang. Email: kaiboo_yang.ty@foxmail.com

Received: 13 January 2025; Accepted: 28 March 2025; Published: 09 June 2025

ABSTRACT: With the rapid urbanization and exponential population growth in China, two-wheeled vehicles have become a popular mode of transportation, particularly for short-distance travel. However, due to a lack of safety awareness, traffic violations by two-wheeled vehicle riders have become a widespread concern, contributing to urban traffic risks. Currently, significant human and material resources are being allocated to monitor and intercept non-compliant riders to ensure safe driving behavior. To enhance the safety, efficiency, and cost-effectiveness of traffic monitoring, automated detection systems based on image processing algorithms can be employed to identify traffic violations from eye-level video footage. In this study, we propose a robust detection algorithm specifically designed for two-wheeled vehicles, which serves as a fundamental step toward intelligent traffic monitoring. Our approach integrates a novel convolutional and attention mechanism to improve detection accuracy and efficiency. Additionally, we introduce a semi-supervised training strategy that leverages a large number of unlabeled images to enhance the model's learning capability by extracting valuable background information. This method enables the model to generalize effectively to diverse urban environments and varying lighting conditions. We evaluate our proposed algorithm on a custom-built dataset, and experimental results demonstrate its superior performance, achieving an average precision (AP) of 95% and a recall (R) of 90.6%. Furthermore, the model maintains a computational efficiency of only 25.7 GFLOPs while achieving a high processing speed of 249 FPS, making it highly suitable for deployment on edge devices. Compared to existing detection methods, our approach significantly enhances the accuracy and robustness of two-wheeled vehicle identification while ensuring real-time performance.

KEYWORDS: Two wheeled vehicles; illegal behavior detection; object detection; semi supervised learning; deep learning; transformer; convolutional neural network

1 Introduction

Two-wheeled vehicles, including electric bicycles and motorcycles, serve as crucial transportation modes in both developing regions and modern cities due to their compact size, parking convenience, operational flexibility, and adaptability to complex road conditions [1–3]. However, this rapid increase in two-wheeled vehicles has led to a corresponding surge in traffic violations.

Given that riders of two-wheeled vehicles are particularly vulnerable to accidents, the consequences of traffic accidents can be severe. Statistical data [4] reveals that over half of road traffic accidents involve two-wheeled vehicles, with red light violations and speeding at intersections dramatically increasing accident risks, especially during peak hours and adverse weather conditions [5]. Consequently, preventing these

violations at intersections has become a major challenge for government agencies seeking to improve traffic conditions and enhance road safety [6–9].

While public education plays a role, administrative enforcement proves more directly effective due to its deterrent effect. Currently, two widely adopted approaches are manual interception and surveillance camera monitoring [10]. Manual interception, though highly accurate, requires substantial human resources and can lead to dangerous situations when riders attempt to evade enforcement through sudden U-turns, road crossings, acceleration, or when law enforcement officers resort to potentially unsafe intervention methods. In contrast, automated surveillance based on digital image processing technology requires minimal manpower, is cost-effective, and is highly scalable, making it a more promising technical solution [11,12]. However, its effectiveness is limited by fixed monitoring locations and camera angles, creating blind spots that riders can learn to avoid, and making it challenging to track violators through surveillance footage effectively.

Thanks to advances in digital image processing, pattern recognition, and artificial intelligence technologies, machine vision is playing an increasingly vital role in traffic management and intelligent transportation systems [13,14], offering new possibilities for addressing these challenges. By evolving beyond traditional fixed camera systems to support both stationary and mobile applications, while adapting to complex traffic environments, variable shooting modes, and random short-term occlusions [15,16], these systems can significantly reduce blind spots without alerting riders to their presence, thereby overcoming current limitations and enhancing the effectiveness of machine vision in traffic violation management.

With the rapid advancement of artificial intelligence, Convolutional Neural Network (CNN) models have been widely adopted across various fields [17], and integrating deep learning into two-wheeled vehicle detection and tracking has emerged as a prominent research direction However, conventional deep learning detection algorithms face certain limitations: their accuracy and adaptability are constrained by the scale and diversity of training samples, which require manual collection, organization, and annotation [18]. Therefore, reducing annotation costs while increasing sample size and diversity has become an urgent task for improving model performance. Semi-supervised learning presents a promising alternative approach that leverages limited labeled data in conjunction with large volumes of unlabeled data, substantially reducing annotation requirements while achieving enhanced adaptability under equivalent cost constraints. However, current semi-supervised learning strategies, whether based on consistency learning or pseudo-labeling, still face limitations as they essentially rely on pseudo-labels generated by teacher models, which can introduce noise and bias that adversely affect detection accuracy. Furthermore, current object detection methods primarily focus on two-wheeled vehicle detection from fixed camera positions and overhead views, and most algorithms have not significantly addressed computational efficiency, potentially leading to suboptimal adaptability.

To address these challenges, we have developed a comprehensive dataset and proposed a semi-supervised learning-based detection algorithm. Our key contributions include:

1. Creation of a robust dataset comprising 5200 annotated images and 15,000 unlabeled images, effectively covering common detection scenarios for two-wheeled vehicles across various environmental conditions and lighting situations.
2. Development of a novel detection algorithm based on CNN technology, incorporating our proprietary attention mechanism module, which achieves high detection accuracy while maintaining a compact model size and computational efficiency.
3. Introduction of an innovative semi-supervised training methodology that leverages our extensive collection of unlabeled data to significantly enhance the model's adaptability across diverse scenarios, while maintaining high detection precision and reliability.

The remainder of this paper is organized as follows. Section 2 provides a detailed description of our proposed approach. Section 3 encompasses the dataset construction, comparative experiments on different convolution approaches, performance evaluation of attention mechanisms, validation of semi-supervised training strategies, and comprehensive comparisons with state-of-the-art algorithms. Finally, Section 4 presents conclusions and outlines directions for future research.

2 Methods

The overall algorithm consists of two parts: a deep convolutional neural network for detecting two-wheeled vehicles and a semi-supervised learning strategy for training the network weights. The fully supervised structure is shown in Fig. 1.

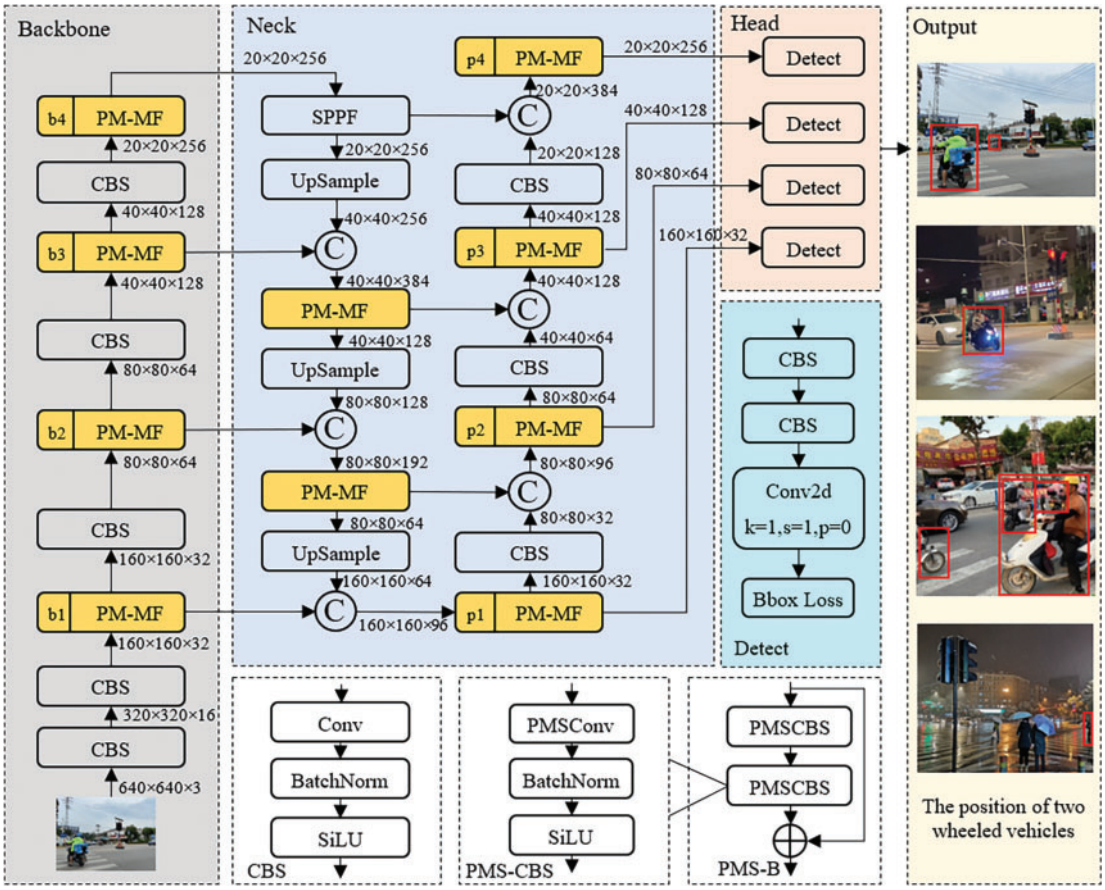

Figure 1: Network structure of two-wheeled vehicle object detection

2.1 Two-Wheeled Vehicle Object Detection Algorithm

2.1.1 Overall Network Structure

The object detection algorithm adopts our proposed Two-Wheeled Vehicle (TWV) Object Detection algorithm, which consists of three main parts: A backbone network for extracting image features; A neck network for deep fusion of image features; A head network for object detection. The detection head is divided into four different detection scales, making it suitable for object detection at different scales.

In addition, as the ultimate goal of the algorithm in this article is to deploy it to edge devices such as handheld cameras or smartphones, the algorithm strictly controls the computational complexity and FPS on different devices to ensure scalability.

2.1.2 PMSConv

The principle of PMSConv is shown in Fig. 2c. It is an improved new convolution based on partial convolution (PConv) and partial mixed convolution (PMConv). PConv is commonly used to reduce redundant computation in residual networks where multiple channels carry the same or similar information. The specific process is as follows: First, the input features are sampled into two parts. One part is directly forwarded to preserve the original gradient information, while the other part undergoes convolution to obtain deep gradient information. Then, the two parts are directly concatenated. Although this method effectively improves computational efficiency, experiments have found that performing calculations on only some of the channels leads to a lack of interaction between the uncalculated inter-layer information, which affects the effectiveness of information fusion.

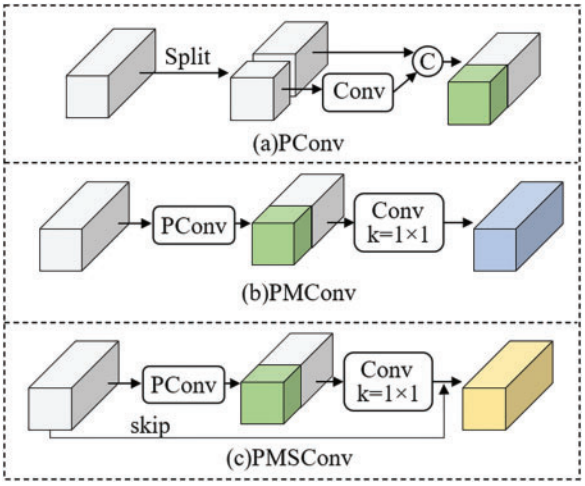

Figure 2: Different convolutional working principles

PMConv further uses 1 × 1 convolution on the original input features and intermediate filtered features for staggered computation based on PConv. This approach can reduce the computational load and extract richer gradient flows compared to ordinary convolution. However, through theoretical analysis, we believe that some significant features of the original feature layer should be preserved. Therefore, we further added a skip branch based on PMConv to obtain our final PMSConv module.

2.1.3 Partial Mix Mini Former Module

Detecting two-wheeled vehicles at long distances is crucial for subsequent efficient tracking and processing. In this state, two-wheeled vehicle targets in video images generally have the characteristics of small pixel ratios and blurred features, making the extracted features easily overwhelmed by background features and causing missed detections. Attention mechanisms are commonly used to improve the algorithm's sensitivity to such sparse features and address this issue. Among them, the Global Attention Mechanism (GAM) [19] has been proven to be a relatively effective method but comes with a huge computational burden. Therefore, this paper proposes a Mini Former Mechanism (MFM) that only calculates the most relevant regions and is sensitive to small two-wheeled vehicle targets, as shown in Fig. 3c.

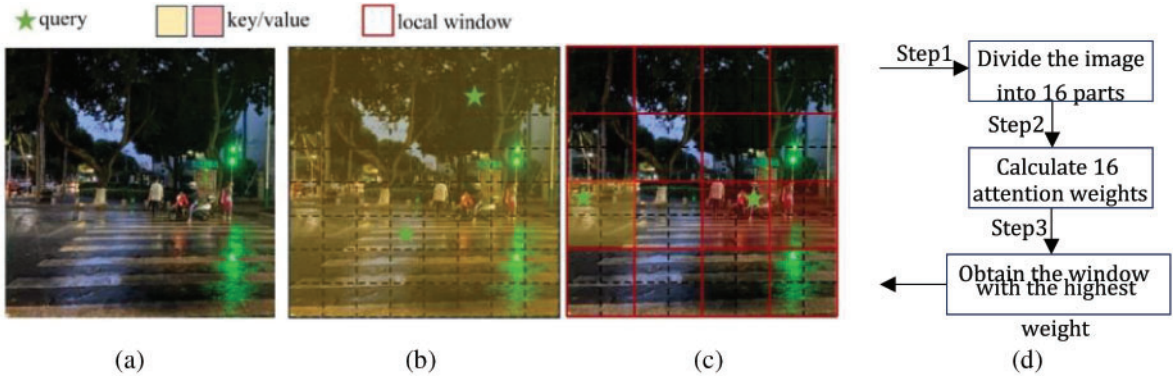

Figure 3: Comparison of the two attention mechanisms. **(a)** Original image; **(b)** GAM; **(c)** MFM; **(d)** Calculation principle

Unlike traditional attention mechanisms that process the entire feature map, our proposed Mini Former Mechanism (MFM) adopts a selective processing strategy. (1) The input image features are divided into a 4 × 4 grid structure, creating 16 independent local window regions. For each window, we independently compute its attention correlation weight to quantify its relevance to the target detection task. (2) These weights are then sorted in descending order to identify the most relevant window. Instead of processing all regions, MFM selectively collects key-value pairs only from the window with the highest attention weight, while skipping computations for the remaining 15 windows. The mechanism proves particularly effective for detecting small two-wheeled vehicles in complex scenes, efficiently managing computational resources by processing only 1/16 of the total image regions while eliminating redundant calculations in less relevant areas.

Based on the PMSConv and the MFM, we construct a plug-and-play partial mixed mini-object self-attention module (PMS-MF) to enhance the image features of special two-wheeled vehicle objects. The specific structure is shown in Fig. 4, where c = i and c = o represent the number of feature channels equal to the preset input and output channel numbers, respectively.

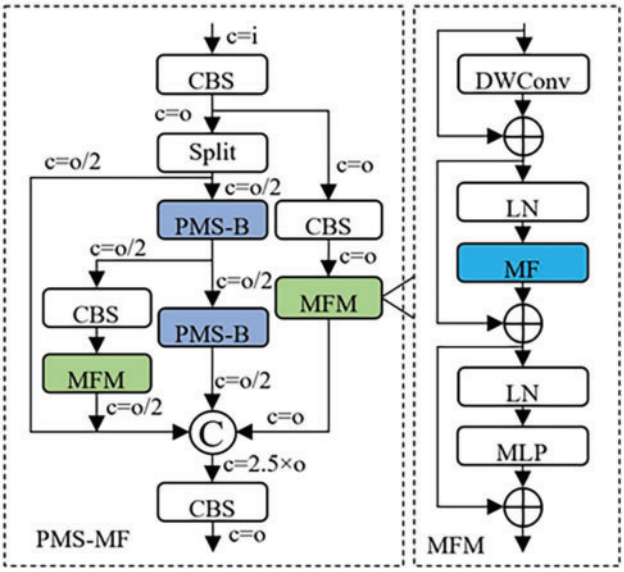

Figure 4: PMS-MF module structure

The backbone of PMS-MF first uses a CBS (Convolution-BatchNorm-SiLU) block to change the number of input feature channels and then fuses the features after passing through four feature processing branches. It does not perform TFM (Transformer) self-attention computation on all branch feature maps. This intermittent attention strategy can avoid overly highlighting important information in the feature maps and obscuring the next most important information, while effectively controlling the computational complexity.

2.1.4 Training Loss Calculation

The loss function of the TWV two-wheeled vehicles detection network consists of two parts: regression box loss and target confidence loss, as shown in Eq. (1).

$$L_s = L_s^{reg} + L_s^{obj} \tag{1}$$

CIoU (Complete-IoU) localization loss is employed, as shown in Eq. (2). CIOU comprehensively considers the overlap area between predicted and ground truth boxes, the distance between their center points, and their aspect ratios. Even when background noise is excessive or the scene is too dense, causing the predicted and ground truth boxes of two-wheeled vehicles to not intersect, CIOU can still reflect their relative distance and degree of overlap.

$$L_s^{reg} = CIOU\left(X_{(h,w)}^{reg}, Y_{(h,w)}^{reg}\right) \tag{2}$$

CE (Cross-Entropy) loss is used, as shown in Eq. (3), which can measure the degree of difference between two different probability distributions in random variables.

$$L_s^{obj} = CE\left(X_{(h,w)}^{obj}, Y_{(h,w)}^{obj}\right) \tag{3}$$

2.2 Semi Supervised Learning Strategy

In past research on two-wheeled vehicle object detection, most studies have adopted a fully supervised approach to train network models. In order to obtain models with high detection accuracy and strong adaptability, it is necessary to prepare large-scale and diverse annotated datasets in advance. To reduce the dependence on annotated datasets, a semi-supervised learning strategy (semi-supervised training plus, SSTP) is proposed, which utilizes a large number of unlabeled images to generate pseudo-labels for training models, thereby improving the detection performance of two-wheeled vehicle targets in different traffic scenarios.

2.2.1 Overall Structure

The semi-supervised learning framework generally includes two pre-trained models: a teacher model (Teacher) and a student model (Student). The teacher model automatically generates annotation information from unlabeled image samples, which is used to construct the annotated samples required for training the student model. Since these annotations may not be entirely accurate, they are referred to as pseudo-labels. After enabling the semi-supervised learning strategy, the unlabeled images are first passed through data augmentation and then fed into the teacher model TWV-T to generate pseudo-labels for the two-wheeled vehicle images. Then, the Auto-PLC (Pseudo-Label Classifier) is used to assess and classify the noise in the pseudo-labels. High-quality pseudo-labels are automatically selected to create training samples that assist in training the student model TWV-S. The already labeled samples and the corresponding two types of pseudo-labeled training samples are combined with preset weight parameters to calculate the corresponding loss functions, which are used to update the student model. Finally, the teacher model is updated based on the

student model using the Exponential Moving Average (EMA) algorithm. This process is repeated until the preset number of training rounds is reached or the total training loss no longer decreases significantly. The specific structure is shown in Fig. 5.

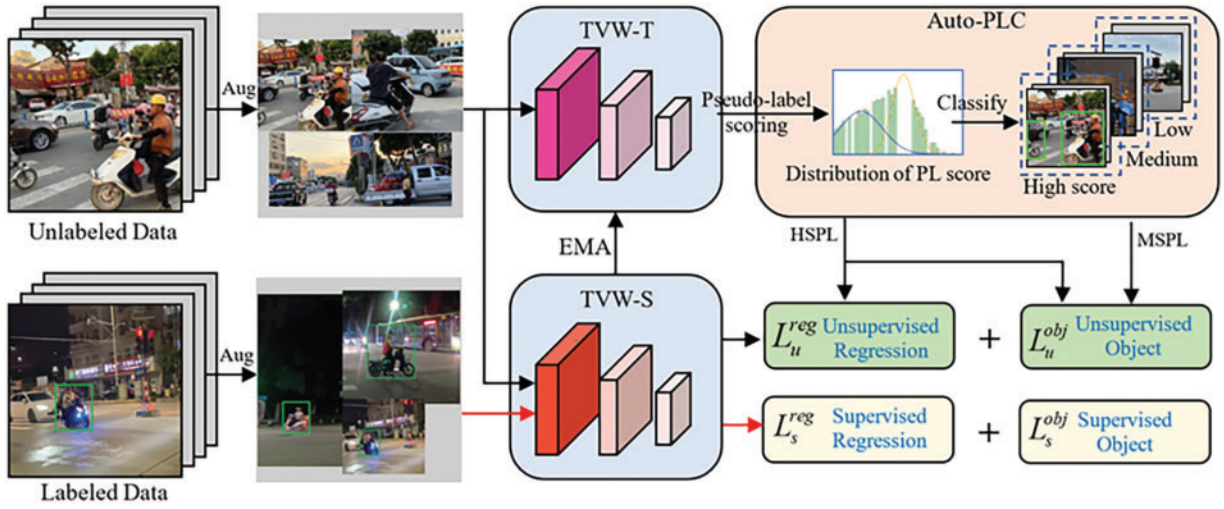

Figure 5: Semi supervised learning structure

2.2.2 Auto-PLC

Unlike conventional object detection tasks, the background information at intersections is complex, with significant interference, and the number of targets in the image can vary greatly. Conventional semi-supervised pseudo-label classification methods use quality metrics and fixed thresholds to classify pseudo-labels. Therefore, using such fixed threshold methods in this application scenario can lead to difficulties in ensuring the quality of pseudo-labels and large differences in the number of targets, which in turn affects the training effect. To address this, we propose an automatic pseudo-label classifier based on Gaussian distribution statistics: Auto-PLC. First, non-maximum suppression (NMS) is used to filter all pseudo-labels. Then, based on the weighted calculation of their confidence (C) and coordinate positions, pseudo-label scores are obtained, and statistical processing is performed on all pseudo-labels using these scores. Our statistical calculations show that the scores of these pseudo-labels generally tend to follow a Gaussian distribution. Therefore, we directly divide the pseudo-labels into high-score pseudo-labels (HSPL), medium-score pseudo-labels (MSPL), and low-score pseudo-labels (SSPL) based on their scores using the Gaussian distribution. Finally, the corresponding losses are calculated separately. For HSPL, both the regression box loss and the object confidence loss are calculated. For MSPL, only the object confidence loss is calculated. SSPL is directly discarded.

2.2.3 Training Loss Calculation

The semi-supervised training loss function consists of two parts: the supervised loss L_s from Eq. (1) and the unsupervised loss L_u, as shown in Eq. (4). The supervised loss primarily serves as a parameter correction mechanism, preventing parameter mutations caused by excessive unsupervised training errors that could affect normal model training. The unsupervised loss directly influences the training effectiveness during the semi-supervised phase and should be assigned a higher loss weight. Therefore, this paper sets the weight balance factor λ to 2.5, and subsequent experimental results validate the effectiveness of this parameter

setting.

$$L = L_s + \lambda L_u \tag{4}$$

Similar to L_s, the unsupervised loss L_u also comprises two components: regression box loss and target confidence loss, as shown in Eq. (5).

$$L_u = L_u^{reg} + L_u^{obj} \tag{5}$$

The unsupervised regression box loss L_ub is shown in Eq. (6), where: $C_{(h,w)}$ represents the pseudo-label confidence at position (h,w), $\tilde{Y}$ represents the pseudo-labels output by the teacher model, $\mathbb{I}\{\cdot\}$ is an indicator function that outputs 1 when the condition is satisfied and 0 otherwise. In this paper, the CIOU regression loss between the student model's predictions and pseudo-labels is calculated when the pseudo-label confidence $C_{(h,w)}$ exceeds the automatic threshold GaussScore2, indicating a reliable pseudo-label.

$$L_u^{reg} = \sum_{h,w} \left(\mathbb{I}_{\{C_{(h,w)};>=\text{GaussScore2}\}} \text{CIOU} \left(X_{(h,w)}^{reg}, \overline{Y}_{(h,w)}^{reg} \right) \right) \tag{6}$$

The target confidence loss L_uc is shown in Eq. (7). When the pseudo-label confidence $C_{(h,w)}$ exceeds threshold GaussScore1, indicating either a reliable pseudo-label or an offset pseudo-label, a cross-entropy loss is used to calculate the target difference between the student model's predictions and the pseudo-labels.

$$L_u^{obj} = \sum_{h,w} \left(\mathbb{I}_{\{C_{(h,w)};>=\text{GaussScore1}\}} \text{CE} \left(X_{(h,w)}^{obj}, \overline{Y}_{(h,w)}^{obj} \right) \right) \tag{7}$$

3 Experiments

3.1 Dataset

We collected datasets in 8 different cities of China. 110 short videos containing two-wheeled motor vehicles were collected at intersections under different weather conditions and time periods using action cameras. Based on the weather and time period captured in the videos, they were categorized into four classes: Daytime, Night, Rain, and Snowy day, facilitating subsequent dataset classification. Then, 5200 traffic images of two-wheeled motor vehicles with significant differences were extracted from these videos. Backlight and Traffic jam images were manually extracted from backlit and congested road segments. The number of image samples for each category is shown in Table 1. Examples of image samples for each category are shown in Fig. 6. Finally, we used the LabelImg software to annotate these images. In addition, the dataset includes 15,000 unlabeled images for semi-supervised training purposes.

Table 1: Performance comparison experiment of attention mechanism

Daytime	Night	Rain	Snowy day	Backlight	Traffic jam	All
1023	1131	632	825	732	857	5200

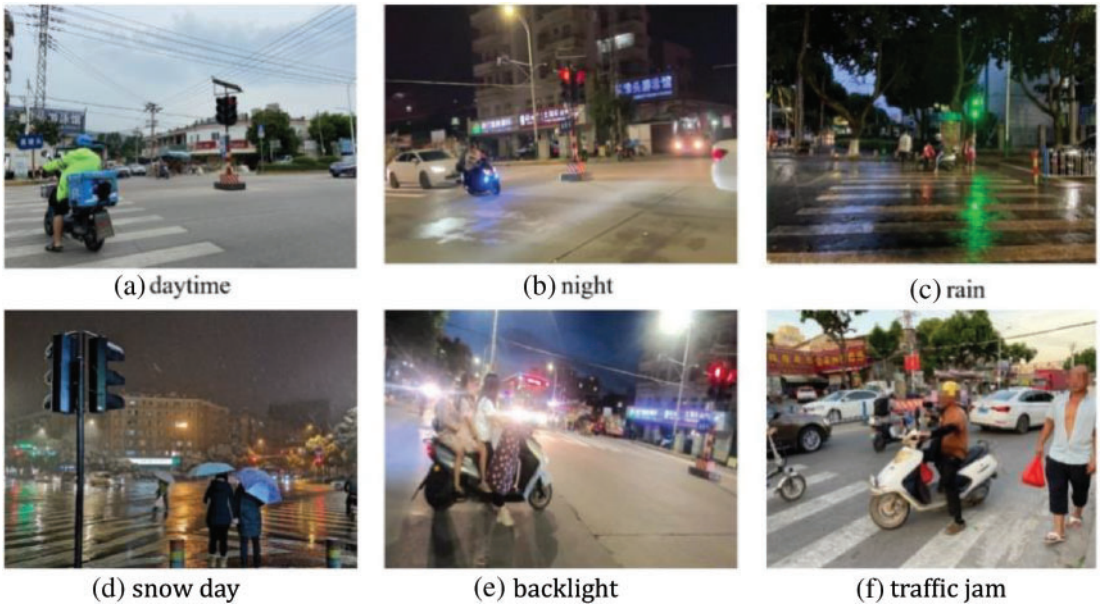

Figure 6: Examples of two-wheeled vehicle images in 6 different scenarios

3.2 Experimental Condition

The experiments were conducted on a 64-bit Windows 10 operating system using an NVIDIA GeForce RTX 3090 graphics card for model training. The initial learning rate was set to 0.001 with exponential decay, a weight decay rate of 0.0005, using the Adam optimizer for 200 epochs. Input images were resized to 640 × 640 pixels and augmented using geometric distortions including random flipping, rotation, translation, and perspective transformation, along with random adjustments to brightness, contrast, saturation, and hue.

To ensure reproducibility, we provide the following crucial implementation details:

1. Training Strategy: Batch size set to 8; Learning rate warm-up period of 10 epochs; EMA (Exponential Moving Average) for model weight updates; Cosine annealing scheduler for learning rate adjustment.
2. Semi-supervised Implementation: Teacher model updates every 2 epochs; Consistency regularization loss with weight coefficient of 0.5; Each iteration uses 50% labeled and 50% unlabeled data.
3. Hyperparameter Optimization: Bayesian optimization for parameter tuning; Label smoothing technique with coefficient 0.1.

3.3 Evaluation Metrics

In the two-wheeled vehicles detection task, four metrics are selected to evaluate the model performance: Recall (R), AP (Average Precision), GFLOPs (Giga Floating Point Operations Per Second), and frames per second (FPS). Among them, R and AP are used to assess the algorithm's accuracy. The expressions are as follows.

$$\mathrm{R} = \frac{TP}{TP + FN} \tag{8}$$

$$\mathrm{P} = \frac{TP}{TP + FP} \tag{9}$$

$$\mathrm{AP} = \int_0^1 P\left(R\right) \mathrm{dR} \tag{10}$$

In the formula, *TP* is a true positive, which means the number of targets correctly detected by the algorithm; *FP* is false positive, which refers to the number of false targets detected; *FN* is false negative, meaning the number of undetected targets $P(R)$ represents the accuracy value at a recall rate of R.

3.4 Experiment

3.4.1 Comparative Experiments of Different Convolutions

To verify the impact of our designed PMSConv on the detection accuracy of two-wheeled vehicles object detection models compared to existing convolutions, we additionally selected Conv, PConv, and PMConv mentioned earlier to replace the bottleneck module in PMS-MF. The experimental results are shown in the Table 2.

Table 2: Comparative experiments of different convolutions

ID	Convolution type	AP/%	R/%	F1/%	GFLOPs	FPS
1	Conv	93.2	88.7	90.9	26.3	231
2	PConv	92.1	86.3	89.1	23.4	252
3	PMConv	92.8	87.2	89.9	25.4	243
4	PMSConv	**93.3**	**88.9**	**91.1**	25.7	249

It can be seen that there are certain differences in the performance and computational efficiency of the TWV network when using different convolution modules: when using standard Conv, the average two-wheeled vehicle AP of the TWV network is 93.2%, the recall rate (R) is 88.7%, and the GFLOPs are 26.3. This module performs stably in terms of detection accuracy and recall rate, but its floating-point operation is relatively high, indicating high computational overhead; After using the PConv module, the floating-point computational load of the model was significantly reduced to 23.4 GFLOPs, which is a decrease of 2.9 GFLOPs compared to the Conv module, or about 11.0% of the computational load. However, this reduction in computational complexity comes at the cost of decreased detection accuracy and recall, with AP at 92.1% and R at 86.3%, resulting in slightly weaker performance compared to Conv. It can be seen that PConv is more suitable for scenarios that are sensitive to computing resources, but it may sacrifice some detection performance; The PMConv module performs well in balancing performance and computational efficiency, with a detection accuracy (AP) of 92.8%, a recall rate of 87.2%, and a floating-point computational load of 25.4 GFLOPs. Compared with the Conv module, PMConv reduces the computational load by 0.9 GFLOPs (approximately 3.4%) while maintaining high detection performance, making it a more balanced choice; The PMSConv we ultimately designed has the best overall performance in terms of performance and efficiency, with a detection accuracy (AP) of 93.3% and a recall rate of 88.9%, which is the highest among all modules. At the same time, its floating-point computational complexity is 25.7 GFLOPs. Compared with the Conv module, although the reduction in floating-point operations is limited, only by 0.6 GFLOPs, about 2.3%, it has been optimized in terms of detection performance, demonstrating strong comprehensive capabilities. These experimental results demonstrate that the final designed PMSConv convolution can effectively balance the performance and computational efficiency of the model, thereby meeting the needs of different scenarios.

3.4.2 Comparative Experiments on Different Attention Mechanisms

To verify the effectiveness of MFM, all MFMs in the network were removed and replaced with: no attention mechanism, Channel Attention Mechanism (CAM), Convolutional Block Attention Module

(CBAM), Global Attention Mechanism (GAM), BiFormer Module, and the proposed MFM. Comparative experiments were conducted, and the experimental results are shown in Table 3.

Table 3: Comparison of different attention mechanisms

Attention type	Daytime		Night		Rain		Snowy day		GFLOPs
	AP/%	R/%	AP/%	R/%	AP/%	R/%	AP/%	R/%	
None	91.8	87.3	90.5	85.2	89.7	84.1	88.9	83.5	24.5
CAM	92.6	88.4	91.2	86.5	90.3	85.7	89.5	84.8	25.1
CBAM	93.1	89	91.8	87.2	90.9	86.3	90.1	85.5	25.3
GAM	93.3	89.2	92	87.5	91.2	86.8	90.4	85.8	25.4
BFM	93.5	89.5	92.5	88	91.8	87.2	90.9	86.3	25.7
MFM	**93.8**	**89.8**	**92.9**	**88.4**	**92.3**	**87.8**	**91.5**	**86.9**	25.7
Attention type	**Backlight**		**Traffic Jam**		**ALL**		**-**		**GFLOPs**
	AP/%	R/%	AP/%	R/%	AP/%	R/%	–	–	
None	89.3	85	88.7	84.5	89.8	84.9	–	–	24.5
CAM	90.5	86.4	90	85.8	91.2	86.1	–	–	25.1
CBAM	91.2	87.2	90.7	86.6	91.9	87	–	–	25.3
GAM	91.6	87.5	91	86.9	92.2	87.4	–	–	25.4
BFM	92	88	91.6	87.4	92.7	87.8	–	–	25.7
MFM	**92.8**	**88.6**	**92.4**	**88**	**93.3**	**88.9**	–	–	25.7

From the data in the Table 3, it can be seen that:

1. No attention mechanism: The model has the lowest AP and R across all scenarios, performing especially poorly in complex conditions like rain and snow due to weak feature extraction.
2. Introducing classic attention mechanisms (CAM, CBAM, GAM) CAM: After introducing the channel attention mechanism, the model has a certain degree of improvement in AP and R in all scenarios, especially in simpler scenarios such as Daytime and Backlight, showing significant optimization. However, the global feature extraction capability of CAM is limited, and its improvement effect in complex scenes is relatively small. CBAM: Compared to CAM, CBAM significantly improves the model's adaptability to complex scenes such as Rain and Traffic jam by combining channel and spatial attention. The improvement in AP and R is greater, but the computational cost slightly increases. GAM: The addition of the Global Attention Mechanism (GAM) further improves the performance of the model in all scenarios, especially in Night and Snowy day, but the computational complexity also increases.
3. MFM: Compared with other attention mechanisms, MFM has the highest AP and R in all scenes, especially in complex scenes such as Rain and Snowy day, showing significant advantages. In scenarios with significant interference features such as Rain, Snowy day, Backlight, and Traffic jam. In simple scenarios such as Daytime, the difference in AP and R between MFM and GAM is relatively moderate, with AP improving by 0.5% and R by 0.6%. In complex scenarios such as the Traffic jam, the performance improvements become more substantial, with AP improving by up to 1.4% and R by 1.1% compared to GAM. This indicates that MFM can more effectively focus on key regions, especially when dealing with sparse features and complex backgrounds, with significant advantages.

In terms of computational overhead, MFM's GFLOPs are 25.7, which is only an increase of 1.2 GFLOPs compared to 24.5 without the attention mechanism. At the same time, it significantly improves the model performance, fully reflecting MFM's good balance between computational efficiency and model performance.

Overall, MFM has become the best attention mechanism choice in complex scenarios due to its precise attention to key areas and efficient computation, while also demonstrating its outstanding advantages in balancing performance and efficiency.

3.4.3 The Ablation Experiment of PMSConv and MFM

In order to demonstrate the performance advantages of our semi supervised learning strategy and the advantages of auto PLC, we conducted a series of ablation experiments, where SST represents the use of semi supervised training strategy, Auto PLC is the automated pseudo label classification method we ultimately adopted, and PLC is the traditional pseudo label classification method based on label threshold for pseudo label classification.

As shown in Table 4, the following analysis can be obtained:

1. In Experiment 1, without introducing semi supervised learning strategy (SST), pseudo label classifier (PLC), or automatic pseudo label classifier (Auto PLC), the AP of the model was 93.3%, and the R was 88.9%. This represents the performance of the TWV model under the basic fully supervised learning strategy. Although it performs well in detection accuracy and recall, it has not fully utilized unlabeled data, leaving some room for optimization.
2. In Experiment 2, only SST was introduced, and the AP and R of the model increased to 94.1% and 89.5%, respectively, by 0.8% and 0.6%. This confirms that semi supervised learning can significantly improve model performance by utilizing unlabeled samples, especially in rare or diverse data scenarios.
3. Experiment 3 further combined SST with a pseudo label classifier (PLC), and the AP and R of the model reached 94.5% and 90.0%, respectively, which increased by 0.4% and 0.5% compared to Experiment 2. This indicates that PLC has improved the utilization efficiency of unlabeled data through a simple and efficient pseudo label filtering mechanism, thereby further optimizing detection performance.
4. Experiment 4 introduced SST and Auto PLC. When SST was combined with Auto PLC, the AP of the model increased to 95.0% and R increased to 90.6%. Compared with Experiment No. 3, AP and R increased by 0.5% and 0.6%, respectively. Auto PLC significantly reduces error propagation in pseudo labels through adaptive pseudo label classification and optimization strategies, demonstrating better detection performance compared to traditional PLCs. This mechanism is particularly significant for object detection of two two-wheeled vehicle in complex traffic scenarios.

Table 4: The ablation experiment of PMSConv and MFM

ID	PMSConv	MFM	AP/%	R/%	F1/%	GFLOPs
1	×	×	90.0	84.0	86.9	25.1
2	√	×	89.9	84.9	87.3	24.5
3	×	√	93.2	88.7	90.9	26.3
4	√	√	93.3	89.9	91.6	25.7

The comprehensive four sets of ablation experiments have demonstrated that (1) with the gradual introduction of semi supervised learning strategy (SST), pseudo label classifier (PLC), and automatic pseudo label classifier (Auto PLC), the AP and R of the TWV two-wheeled vehicle detection model continue to

improve, fully verifying the effectiveness of these strategies. (2) Compared to traditional PLCs, Auto PLC can generate and screen pseudo labels more intelligently, and it performs particularly well in complex scenes, ultimately achieving the best experimental levels of AP and R. (3) Based on the experimental results, it can be seen that the combination of SST and Auto PLC is an efficient optimization strategy that not only significantly improves model performance, but also does not significantly increase computational overhead, which is one of the core innovations of this study.

3.4.4 Semi Supervised Learning Strategy Performance Experiment

In order to demonstrate the performance advantages of our semi supervised learning strategy and the advantages of Auto-PLC, we conducted a series of ablation experiments, where SST represents the use of semi supervised training strategy, PLC is the traditional pseudo label classification method based on label threshold for pseudo label classification, and Auto-PLC is the automated pseudo label classification method we ultimately adopted,

The comprehensive four sets of ablation experiments have demonstrated that (1) with the gradual introduction of semi supervised learning strategy (SST), pseudo label classifier (PLC), and automatic pseudo label classifier (Auto-PLC), the AP and R of the TWV detection model continue to improve, fully verifying the effectiveness of these strategies. (2) Compared to traditional PLCs, Auto-PLC can generate and screen pseudo labels more intelligently, and it performs particularly well in complex scenes, ultimately achieving the best experimental levels of AP and R. (3) Based on the experimental results, it can be seen that the combination of SST and Auto PLC is an efficient optimization strategy that not only significantly improves model performance, but also does not significantly increase computational overhead, which is one of the core innovations of this study (Table 5).

Table 5: Semi supervised ablation experiment

ID	SST	PLC	Auto-PLC	AP/%	R/%	F1/%
1	×	×	×	93.3	88.9	91.0
2	√	×	×	94.1	89.5	91.7
3	√	√	×	94.5	90.0	92.2
4	√	×	√	**95.0**	**90.6**	**92.7**

3.4.5 Comparative Experiments with Other Models

To verify the superiority of the proposed algorithm, it was trained from scratch and compared with mainstream algorithms on a two-wheeled vehicle dataset. The results, shown in Table 6, reveal the following. Among fully supervised algorithms, TWV achieved the highest accuracy, followed by YOLOv11. A detailed comparison highlights:

1. Detection Performance: TWV outperforms YOLOv11 comprehensively. Under standard daytime conditions, TWV achieved AP 93.8% and R 89.8%, improving by 1.7% and 4.0%, respectively. In complex environments, especially rainy conditions, TWV's AP and R reached 92.3% and 87.8%, surpassing YOLOv11 by 7.8% and 8.8%, demonstrating superior feature extraction and adaptability.
2. Stability: TWV maintains consistent performance across scenarios, with only 0.9% and 1.4% variation in AP and R between daytime and nighttime. In contrast, YOLOv11 shows greater fluctuations, indicating TWV's robustness.

3. Efficiency: Despite a slightly higher theoretical cost (25.7 GFLOPs vs. 23.0 GFLOPs for YOLOv11), TWV achieves 249 FPS, surpassing YOLOv11 by 19 FPS, highlighting its optimized computational design.

Comparing fully supervised TWV with the semi-supervised TWV-SSTP:

1. Performance Boost: TWV-SSTP achieved state-of-the-art results, improving AP and R over TWV by 1.2% (daytime), 1.3%–1.7% (adverse conditions), proving its enhanced feature extraction in complex scenarios.
2. Robustness in Challenging Cases: TWV-SSTP excels in backlighting (AP 94.0%, R 90.0%) and traffic congestion (AP 93.8%, R 89.6%), demonstrating superior adaptability.
3. Computational Efficiency: TWV-SSTP maintains the same 25.7 GFLOPs while leveraging richer samples for better accuracy without added complexity. It achieves 249 FPS, ensuring real-time performance.

These results confirm TWV-SSTP's effectiveness in improving detection accuracy and robustness across diverse environments.

Table 6: Comparison experiment with other models

Models	Daytime		Night		Rain		Snowy day		GFLOPs	FPS
	AP/%	R/%	AP/%	R/%	AP/%	R/%	AP/%	R/%		
Faster R-CNN	80.2	70.1	70.6	62.5	74.8	66.1	76.3	67.5	40.3	125.3
SSD	72.5	61	65.8	55.8	67.1	57.3	69.5	58.6	37.6	151.1
YOLOv5	87	78.4	76.3	69.8	80.1	73.4	82.5	74.9	2.7	277.9
YOLOv8	89.5	83.2	85.6	79.5	81.3	76.5	87.2	80.6	18.2	265.3
YOLOv10	90.8	84.3	87.2	81.5	82.7	77.6	88.9	82.3	20	251
YOLOv11	92.1	85.8	89	83	84.5	79	90.6	83.5	23	230
StreamYOLO	91.0	84.5	88.0	82.0	83.0	78.0	89.5	82.5	20.0	270
DETR	74.5	82.3	68.7	76.3	70.2	77.8	71.1	79.2	18.7	252.1
RT-DETR	88.0	83.5	85.5	81.0	82.0	78.5	88.0	82.0	21.0	255
TWV	93.8	89.8	92.9	88.4	92.3	87.8	91.5	86.9	25.7	249
TWV-SSTP	**95**	**91**	**94.2**	**90**	**93.6**	**89.5**	**93**	**88.5**	25.7	249
Models	**Backlight**		**Traffic jam**		**ALL**		–		**GFLOPs**	**FPS**
	AP/%	**R/%**	**AP/%**	**R/%**	**AP/%**	**R/%**	**–**	**–**		
Faster R-CNN	80.1	69.2	76.5	67.8	75	66.3	–	–	40.3	125.3
SSD	71.3	59.7	68.4	58.2	67.2	56.5	–	–	37.6	151.1
YOLOv5	84.5	77.2	81.6	74.3	80.2	73	–	–	2.7	277.9
YOLOv8	89.2	83	87	81.2	82.8	77.5	–	–	18.2	265.3
YOLOv10	91	84.7	88.3	82	83.9	78.5	–	–	20	251
YOLOv11	92	86	89.6	83.7	85	79.8	–	–	23	230
StreamYOLO	90.0	83.5	88.5	82.0	82.5	77.0			20.0	270
DETR	72.8	81.2	71.4	78.5	70.5	78.3	–	–	18.7	252.1
RT-DETR	87.5	82.5	86.0	81.1	84.0	79.5			21.0	255
TWV	92.8	88.6	92.4	88	93.3	88.9	–	–	25.7	249
TWV-SSTP	**94**	**90**	**93.8**	**89.6**	**95**	**90.6**	–	–	25.7	249

3.4.6 Comparison of Visual Inspection Results

To highlight our algorithm's advantages, we conducted a visual comparison between a sub-optimal algorithm and TWV-SSTP, as shown in Fig. 7: (1) In common scenarios (daytime, night, rain, backlight, traffic jam), both performed similarly, though TWV-SSTP had slightly higher confidence scores. (2) When detecting smaller targets such as two-wheeled vehicles in daytime and nighttime conditions, YOLOv11 missed some detections. However, TWV-SSTP maintained good detection performance thanks to its proprietary attention mechanism. (3) In challenging scenarios like snowy days and traffic jams, where objects were either partially occluded or incompletely captured by the camera, YOLOv11 showed detection failures due to limited training samples for such cases. In contrast, TWV-SSTP, leveraging our proposed semi-supervised training method, successfully detected these challenging objects.

In addition. As observed in Fig. 7f, our algorithm is capable of detecting two-wheeled vehicles even in cases where only a single wheel is visible or the vehicle appears blurred. However, in practical applications, the algorithm can be configured to ignore such targets by adjusting the confidence threshold. The detection of these challenging targets in this study was intentional, designed to highlight the strong adaptability of our algorithm and its superior performance compared to other methods in handling such scenarios. Furthermore, it can be noted that the images used in our research were captured from random perspectives. This approach offers the advantage of enabling the use of handheld devices for image capture at specific times and locations, eliminating the need to fix the equipment at a particular site.

Figure 7: (Continued)

Figure 7: Visual comparison between YOLOv11 and the final algorithm in this article

4 Conclusion

We propose a method for two-wheeled vehicle object detection. In developing this algorithm, we first designed our own convolution operation called PMSConv, which both reduces computational complexity and effectively fuses features from different image layers to achieve superior detection performance. We then developed an attention mechanism called MF that is specifically optimized for detecting two-wheeled vehicles in complex scenes. Using PMSConv and MF as building blocks, we created the plug-and-play PMS-MF module. Based on these components, we constructed our TVW two-wheeled vehicle detection model. We then trained the model using a semi-supervised learning approach based on Auto-PLC, utilizing limited labeled images alongside a large number of unlabeled images to achieve strong detection performance. Experimental results demonstrate that our algorithm performs well across various scenarios. In future work, we plan to expand our dataset and enhance object detection capabilities in rural and suburban environments.

Acknowledgement: The authors would like to express our sincere gratitude and appreciation to each other for our combined efforts and contributions throughout the course of this research paper.

Funding Statement: This project was supported by the Natural Science Foundation Project of Fujian Province, China (Grant No. 2023J011439 and No. 2019J01859).

Author Contributions: The authors confirm their contribution to the paper as follows: Study conception and design: Kaibo Yang, Mingen Zhong; data collection: Kaibo Yang, Ziji Xiao; analysis and interpretation of results: Kaibo Yang, Jiawei Tan, Kang Fan, Zhiying Deng, Mengli Zhou; draft manuscript preparation: Kaibo Yang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data used to support the findings of this study are available from the corresponding author upon request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Kumar P, Ranganath S, Weimin H, Sengupta K. Framework for real-time behavior interpretation from traffic video. IEEE Trans Intell Transport Syst. 2005;6(1):43–53. doi:10.1109/TITS.2004.838219.
2. Haworth N, Debnath AK. How similar are two-unit bicycle and motorcycle crashes? Accid Anal Prev. 2013;58(6):15–25. doi:10.1016/j.aap.2013.04.014.
3. Zhong M, Yang K, Xiao Z. Detection of red light running behavior of two wheeled motor vehicles at intersections. In: Eighth International Conference on Traffic Engineering and Transportation System (ICTETS 2024); 2024 Sep 20–22; Dalian, China. 33 p. doi:10.1117/12.3054510.
4. Hughes BP, Newstead S, Anund A, Shu CC, Falkmer T. A review of models relevant to road safety. Accid Anal Prev. 2015;74(3–4):250–70. doi:10.1016/j.aap.2014.06.003.
5. Pai CW. Motorcycle right-of-way accidents—a literature review. Accid Anal Prev. 2011;43(3):971–82. doi:10.1016/j.aap.2010.11.024.
6. Kumar A, Guru Prasad MS, Soni D, Rani V, Arora K, Ganesh DR. Real-time helmet detection system with vehicle number extraction for two-wheeler vehicles using YOLOv8. In: 2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT); 2024 May 2–3; Gharuan, India. p. 672–6. doi:10.1109/InCACCT61598.2024.10551102.
7. Wang L, Chen F, Yin H. Detecting and tracking vehicles in traffic by unmanned aerial vehicles. Autom Constr. 2016;72(11):294–308. doi:10.1016/j.autcon.2016.05.008.
8. Guo H, Zhang Y, Chen L, Ahmad KA. Research on vehicle detection based on improved YOLOv8 network. arXiv: 2501.00300. 2024.
9. Berwo MA, Fang Y, Sarwar N, Mahmood J, Aljohani M, Elhosseini M. YOLOv8n-CGW: a novel approach to multi-oriented vehicle detection in intelligent transportation systems. Multimed Tools Appl. 2025;84(7):3809–40. doi:10.1007/s11042-024-19145-4.
10. Zhou D, Zhao Z, Yang R, Huang S, Wu Z. Mining the micro-trajectory of two-wheeled non-motorized vehicles based on the improved YOLOx. Sensors. 2024;24(3):759. doi:10.3390/s24030759.
11. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: single shot multibox detector. In: Computer Vision—ECCV 2016: 14th European Conference; 2016 Oct 11–14; Amsterdam, The Netherlands.
12. Ganapathy S, Ajmera D. An intelligent video surveillance system for detecting the vehicles on road using refined YOLOV4. Comput Electr Eng. 2024;113(4):109036. doi:10.1016/j.compeleceng.2023.109036.
13. Long X, Deng K, Wang G, Zhang Y, Dang Q, Gao Y, et al. PP-YOLO: an effective and efficient implementation of object detector. arXiv:2007.12099. 2020.
14. Redmon J, Farhadi A. YOLOv3: an incremental improvement. arXiv:1804.02767. 2018.
15. Bochkovskiy A, Wang CY, Liao HM. YOLOv4: optimal speed and accuracy of object detection. arXiv:2004.10934. 2020.
16. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017 Jul 21–26; Honolulu, HI, USA. p. 2117–25.
17. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020.
18. Bakirci M. Utilizing YOLOv8 for enhanced traffic monitoring in intelligent transportation systems (ITS) applications. Digit Signal Process. 2024;152(2):104594. doi:10.1016/j.dsp.2024.104594.
19. Liu Y, Shao Z, Hoffmann N. Global attention mechanism: retain information to enhance channel-spatial interactions. arXiv:2112.05561. 2021.