CCW-YOLO: A Modified YOLOv5s Network for Pedestrian Detection in Complex Traffic Scenes

Zhaodi Wang *, Shuqiang Yang, Huafeng Qin, Yike Liu, Jinyan Ding

Posted Date: 30 October 2024

doi: 10.20944/preprints202410.2326.v1

Keywords: pedestrian detection; traffic scene; YOLO neural network; Coordinate Attention

Article

CCW-YOLO: A Modified YOLOv5s Network for Pedestrian Detection in Complex Traffic Scenes

Zhaodi Wang [1,*], Shuqiang Yang [1,2], Huafeng Qin [3], Yike Liu [1] and Jinyan Ding [1]

[1] College of Physical and Electronic Information, Luoyang Normal University, HeNan 471934, China
[2] School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221000, China
[3] The Chongqing Key Laboratory of Intelligent Perception and BlockChain Technology, Chongqing Technology and Business University, Chongqing 400067, China
* Correspondence: wzdzzu@126.com

Abstract: In traffic scene, pedestrian target detection faces significant issues of misdetection and omission due to factors such as crowd density and obstacle occlusion. To address these challenges and enhance detection accuracy, we propose an improved CCW-YOLO algorithm. The algorithm first introduces a lightweight convolutional layer using GhostConv and incorporates an enhanced C2f module to improve the network's detection performance. Additionally, it integrates the Coordinate Attention module to better capture key points of the targets. Next, the bounding box loss function CIoU Loss at the output of YOLOv5 is replaced with WiseIoU Loss to enhance adaptability to various detection scenarios, thereby further improving accuracy. Finally, we develop a pedestrian count detection system by using PyQt5 to enhance human-computer interaction. Experimental results on the INRIA public dataset show that our algorithm achieves a detection accuracy of 95.6%, representing a 8.7% improvement over the original YOLOv5s algorithm. This advancement significantly enhances the detection of small objects in images and effectively addresses misdetection and omission issues in complex environments. These findings have important practical implications for ensuring traffic safety and optimizing traffic flow.

Keywords: pedestrian detection; traffic scene; YOLO neural network; Coordinate Attention

1. Introduction

With the rapid acceleration of urbanization, traffic safety issues are becoming increasingly prominent. Pedestrian detection, as a crucial technology in intelligent transport systems, holds significant research importance and practical value. Effective pedestrian detection algorithms can identify and locate pedestrians in real-time traffic scenarios, thus providing essential data support for intelligent driving [1], automated parking, and monitoring systems. This significantly enhances road safety and pedestrian protection. Moreover, pedestrian detection technology plays a vital role in smart city development by aiding urban management [2], facilitating traffic flow analysis [3], and enabling event monitoring, thereby promoting the intelligence and humanization of urban traffic.

However, pedestrian detection faces numerous challenges due to the complexity of traffic environments. Background noise can lead to false positives or missed detections, while variations in illumination impact image clarity and contrast, making pedestrian features difficult to discern. Additionally, the diverse postures of pedestrians and differences in clothing colors further complicate detection, and occlusion is particularly common in busy traffic settings, often resulting in the misidentification of pedestrians. To overcome these challenges, it is essential to develop more robust detection algorithms that enhance pedestrian detection performance in complex environments, ultimately providing strong support for the safe operation of intelligent transport systems.

1.1. Related Work

Pedestrian detection algorithms in complex scenes can be broadly categorized into two main types: traditional machine vision-based methods and deep learning-based approaches. Traditional methods rely on manual feature extraction and established machine learning algorithms, boasting a

 doi:10.20944/preprints202410.2326.v1

long research history and relatively mature techniques. For example, Haar feature cascade classifier [4], which achieves fast detection through simple rectangular features, is suitable for real-time applications and performs well especially in simpler contexts. Another classic algorithm is HOG (Histogram of Oriented Gradients) [5], which achieves good detection results by calculating the orientation gradient features of the image and combining them with Support Vector Machine (SVM) [6] for classification, especially performs well in detecting pedestrians under different poses and shape changes. DPM (Deformable Part Model) [7] introduces the concept of deformable parts to capture the overall shape of a pedestrian by modeling its constituent parts and their interrelationships. Although DPM significantly improves accuracy, it comes with relatively high computational complexity [8]. In addition, pedestrian detection methods based on image segmentation have also received attention, which are usually combined with background modelling techniques to achieve pedestrian detection by separating the foreground from the background. While these traditional methods are effective in simpler scenes, they are susceptible to interference in complex environments and dynamic backgrounds, which can diminish detection efficacy.

With the rapid advancement of deep learning technology, methods based on convolutional neural networks (CNN) [9] have increasingly become the mainstream approach in pedestrian detection. The classical algorithm, Region-based Convolutional Neural Networks (R-CNN) [10], markedly enhances the accuracy of pedestrian detection by generating candidate regions and employing CNNs for feature extraction. However, the computational complexity of R-CNN limits its performance in real-time applications [11]. To address this issue, Faster R-CNN [12] was developed, incorporating a Region Proposal Network (RPN) to effectively balance detection speed and accuracy, thus becoming widely used across various application scenarios.Conversely, the YOLO (You Only Look Once) [13] algorithm achieves end-to-end detection using a single neural network, enabling pedestrian detection under real-time conditions. It is particularly well-suited for video applications. The Single Shot MultiBox Detector (SSD) [14] adopts a similar principle, enhancing adaptability to pedestrians of varying scales by detecting them across multi-scale feature maps. Additionally, Mask R-CNN [15] extends Faster R-CNN by integrating instance segmentation capabilities, allowing for simultaneous detection and segmentation, which provides higher detection accuracy and fine-grained information.These deep learning methods significantly outperform traditional approaches in complex scenes, not only improving the accuracy of pedestrian detection but also prompting technological advancements and expanding applications in this field.

1.2. Motivation

Traditional methods have made some progress in early pedestrian detection research; however, their limitations are becoming increasingly apparent. Haar feature cascade classifier is widely used due to its high computational efficiency, but it performs poorly under light changes and complex backgrounds, which can easily lead to misdetection and omission [16]. Reference [17] used HOG (Histogram of Oriented Gradients) combined with Support Vector Machines (SVM), which improves the detection accuracy but is not robust enough in dealing with different poses and scale variations, and its performance tends to degrade significantly, especially in crowded scenes. In addition, although DPM (Deformable Part Model) in the literature [18] is able to capture the shape and structure changes of pedestrians, its computational complexity is high and its real-time performance is limited.Therefore, the primary shortcomings of traditional pedestrian detection methods can be summarized as follows: limitations in feature selection, sensitivity to variations in lighting and background, inadequate real-time performance, and poor robustness against scale, pose, and occlusion [19]. These methods typically rely on hand-crafted features, making it challenging to capture the complex morphology of pedestrians fully. This often results in increased rates of misdetection and omission in complex environments. Furthermore, the high computational demands associated with these traditional techniques hinder their effectiveness in dynamic and crowded scenes.

The primary advantage of deep learning methods over traditional approaches lies in their powerful feature learning capabilities and efficient automated processing. Utilizing architectures such as convolutional neural networks (CNNs), deep learning can automatically extract complex, high-dimensional features, significantly enhancing the recognition accuracy of pedestrians across

 doi:10.20944/preprints202410.2326.v1

varying scales, postures, and occlusions [20]. Furthermore, deep learning methods generally exhibit greater robustness, effectively managing lighting changes and background interference, making them particularly suitable for dynamic scenes and video surveillance applications. These advantages have positioned deep learning as a mainstream method in the field of pedestrian detection, garnering substantial attention, albeit accompanied by several challenges.Region-based Convolutional Neural Networks (R-CNN) generate candidate regions through selective search, subsequently extracting features and classifying them with CNNs. Despite its high detection accuracy, R-CNN's considerable computational complexity and limited real-time performance restrict its applicability [21]. In response, Faster R-CNN incorporates a Region Proposal Network (RPN) to enhance detection speed, thereby improving real-time performance, although it still falls short of meeting the demands for high frame rate detection [22].The Single Shot MultiBox Detector (SSD) conducts detection on feature maps at multiple scales, balancing speed and accuracy, and performs particularly well in detecting small targets. However, it continues to experience degradation in performance within complex backgrounds [23]. RetinaNet addresses the challenge of detection accuracy by introducing focal loss [24], which mitigates the imbalance between foreground and background samples, thus enhancing small target detection capabilities; yet, its computational complexity remains relatively high, resulting in poorer real-time performance. The YOLO (You Only Look Once) [25] algorithm has gained significant attention for its end-to-end detection approach, markedly improving detection speed by dividing images into grids and simultaneously predicting bounding boxes and categories within each grid. Nonetheless, the detection accuracy of YOLO for small targets and in crowded scenes still requires improvement [26]. Current advancements in YOLO-based algorithms primarily focus on enhancing detection accuracy and real-time performance by integrating deeper feature extraction networks and attention mechanisms, effectively improving small target detection, especially in complex and crowded environments.

1.3. Our Work

Inspired by these facts, we propose a pedestrian detection network, CCW-YOLO, based on YOLOv5, specifically designed to address issues of low accuracy, missed detections, and errors in pedestrian detection within traffic scenarios. First, our algorithm introduces a lightweight convolutional layer, GhostConv, which aims to reduce both computational load and the number of parameters while maintaining efficient feature extraction capabilities. Additionally, the enhanced C2f module further improves the overall detection performance of the network by optimizing the feature fusion process, thus enhancing adaptability across diverse scenarios.Second, we incorporate the Coordinate Attention module, an attention mechanism that enables the network to better localize pedestrians in complex backgrounds by focusing on spatial information and feature distribution. This results in more accurate detection outcomes. Finally, we adopt WiseIoU Loss as the bounding box loss function at the output of YOLOv5 to improve adaptability to various detection scenarios, ensuring that the model performs robustly in changing environments.The contributions of our work are summarized as follows:

- Introduction of Lightweight Convolution (GhostConv): We first implement lightweight convolution within the YOLOv5 pedestrian detection algorithm, significantly reducing the model's computational complexity and the number of parameters. This approach maintains excellent feature extraction capabilities while enhancing the efficiency of real-time detection.
- Design of an Improved C2f Module: By constructing an enhanced C2f module, we optimize the feature fusion process to address the low accuracy in pedestrian detection resulting from variations in environmental scales. This improvement enhances the network's adaptability across diverse scenes, thereby significantly boosting pedestrian detection performance.
- Incorporation of Coordinate Attention: We introduce a coordinate attention mechanism to enhance the model's ability to capture key target locations. This innovation facilitates more accurate localization of pedestrians in complex backgrounds, reducing the incidence of missed and false detections while improving overall detection accuracy.
- Design of the WiseIoU Loss Function: The WiseIoU Loss function is employed in bounding box calculations, integrating the overlapping region, centroid distance, and aspect ratio. This design enhances the model's adaptability to various detection scenarios, ensuring robust performance in dynamic environments.

 doi:10.20944/preprints202410.2326.v1

- Experimental Validation and Performance Enhancement: Rigorous experiments conducted on public datasets reveal that the detection accuracy of the CCW-YOLO algorithm reaches 95.6%, which is an improvement of 8.7% over the original YOLOv5s algorithm. This advancement significantly enhances the detection of small objects in images and effectively addresses the issues of misdetection and omission in complex scenes.

2. The Proposed Approach

In this study, we utilized YOLOv5s as the foundation for model improvement. Figure 1 illustrates the network structure of YOLOv5 [27], which is divided into four main components: Input, Backbone, Neck, and Output. The input image is standardized to a resolution of 640 x 640 pixels for model training.The Backbone component is responsible for feature extraction and progressively reduces the size of the feature map. The Neck component enhances the model's capability to detect targets of varying sizes through multi-scale feature fusion and upsampling operations. Finally, the output component predicts the class, bounding box coordinates, and confidence level of the target based on the feature information provided by the Neck network.

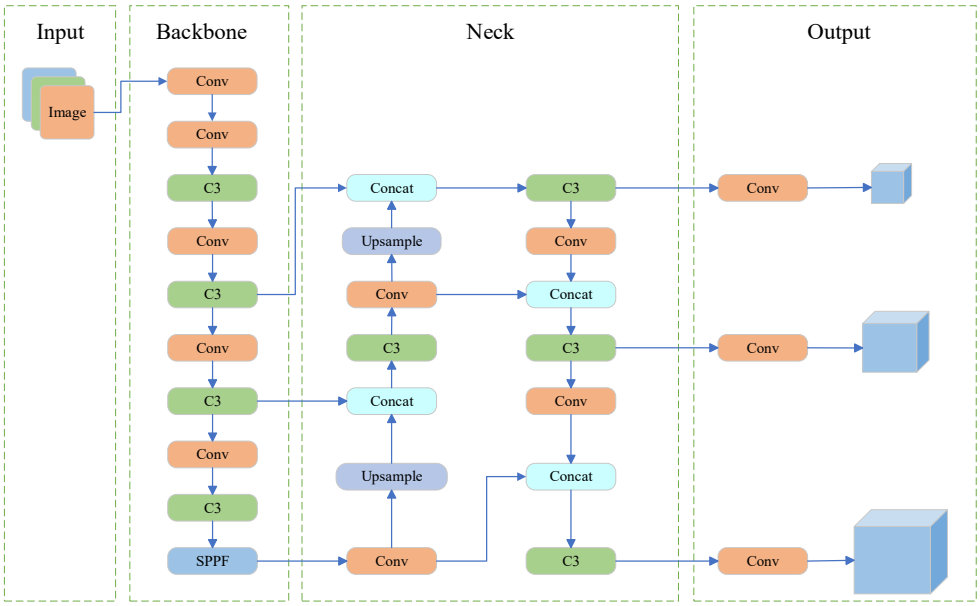


Figure 1. YOLOv5s network structure.

To address the issues of false and missed detections resulting from overlapping and occlusion of pedestrian targets in complex traffic scenarios, we have enhanced the YOLOv5s model,as shown in Figure 2.Firstly, we introduce the GhostConv and C2f modules to optimize the convolutional layer structure, replacing the C3 module in YOLOv5s. By fusing feature maps from different layers, we enhance the gradient flow information and feature representation of the model, thereby improving the detection of small targets.Secondly, the incorporation of the coordinate attention mechanism (Coordinate Attention module) strengthens the model's ability to focus on key regions within the image, further enhancing the accuracy of target detection.Finally, we improve the output loss function of YOLOv5 by adopting WiseIoU Loss in place of the original CIoU Loss. This modification enhances the model's robustness in complex backgrounds and under varying scales through more detailed pixel-level evaluations and effective handling of occlusion cases.

 doi:10.20944/preprints202410.2326.v1

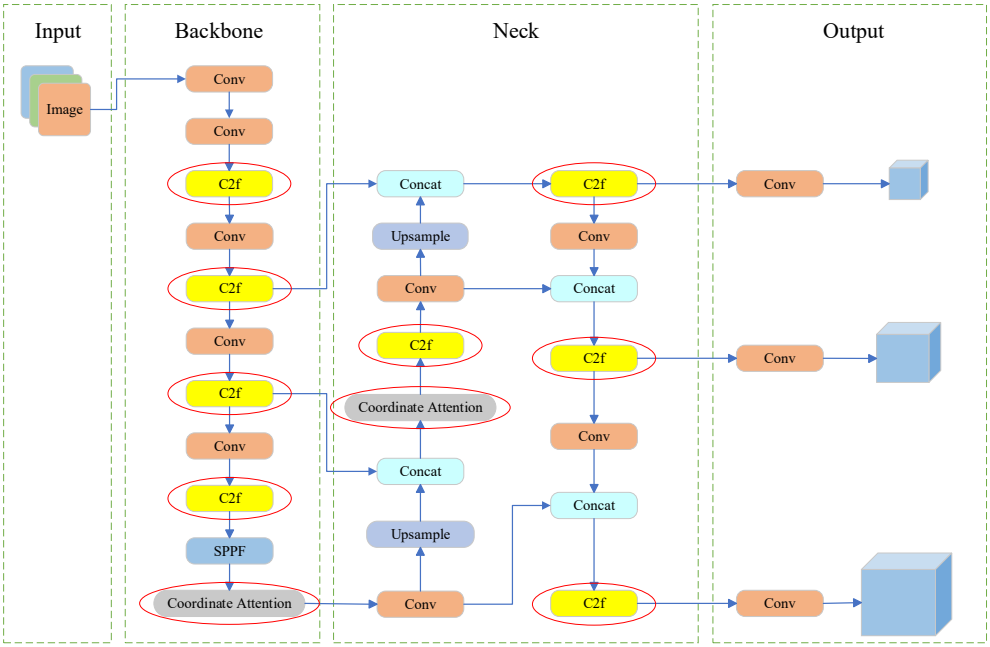


Figure 2. Improved CCW-YOLO network structure .

2.1. Improved Backbone Network

In traffic scene, the size and orientation of pedestrian targets are continuously changing, particularly for smaller targets at greater distances, whose feature representations tend to be relatively weak. This weak representation makes it challenging for the model to extract sufficient information for accurate detection. To address this issue, this study focuses on enhancing the Backbone network of the YOLO model to improve both the efficiency and accuracy of feature extraction, enabling it to better adapt to variations in target scales and shapes.

2.1.1. GhostConv Structure

The GhostConv structure [28] consists of three steps: standard convolution, Ghost generation, and feature map splicing. The operational principle of GhostConv is illustrated in Figure 3. The Ghost Module operates in three steps to obtain the same number of feature maps as a normal convolution. Firstly, GhostNet applies a standard convolution (1×1×M), followed by batch normalization and the ReLU activation function, to compress the input image in terms of channel count and generate intrinsic feature maps. The input feature maps X are convolved by Equation 1 to obtain the intrinsic feature maps $Y^{'}$

$$Y^{'} = X * f^{'} \tag{1}$$

These feature maps are then applied using a series of simple linear operations (unit mapping in parallel with linear transformation) to obtain more feature maps and increase the number of features.The feature maps of each channel of $Y^{'}$ are linearly transformed Φ_{ij} to produce the Ghost feature maps Y_{ij} by Equation 2.

$$Y_{ij} = \Phi_{ij}(Y^{'}), \quad \forall i = 1,2,...,m, \quad j = 1,2,...,s \tag{2}$$

Finally, the intrinsic feature maps obtained in the first step and the Ghost feature maps obtained in the second step are spliced (identity connection) to obtain the final result OutPut. Compared with the traditional convolutional neural network, the Ghost module significantly reduces the number of parameters required and the computational complexity, which can effectively improve the speed of training and inference.

 doi:10.20944/preprints202410.2326.v1

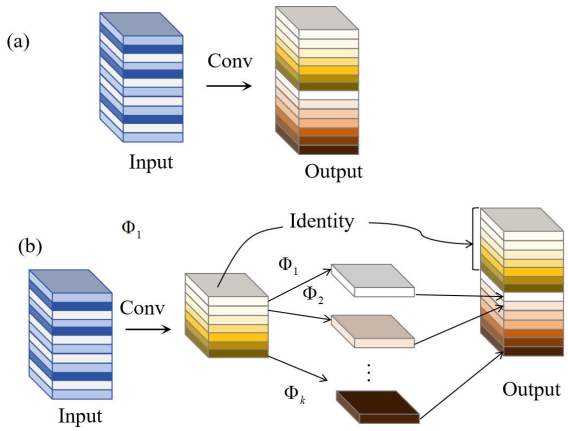


Figure 3. Comparison of Ghost convolution and Normal Convolution. (a) Normal Convolution; (b) Ghost convolution.

2.1.2. C2f Module

The C3 module in the YOLOv5s model exhibits limitations in accurately detecting targets with specific scales and aspect ratios. Its inability to adapt effectively to objects of varying scales and shapes results in suboptimal detection performance, particularly for small or irregularly shaped targets. To address this issue, we proposes replacing the C3 module in the YOLOv5s architecture with the C2f module. This enhancement significantly improves the model's capability to detect small targets by integrating feature maps from different layers, thereby providing richer gradient flow information and enhancing feature representation.

The C2f module is a hybrid neural network structure that primarily integrates high-level semantic features with low-level detail features. A comparison of the network structures of the C3 and C2f modules is presented in Figure 4. The C2f module borrows design concepts from the C3 module, removing the sub-branching convolution, which results in a more lightweight network architecture. Additionally, it incorporates a Split operation and more parallel gradient flow branches—known as skip connections—allowing the output feature map of the C2f module to contain richer gradient information and enhanced feature representation. This improvement significantly enhances the expressive capability of the entire network model, thereby boosting its performance in detecting small targets.

 doi:10.20944/preprints202410.2326.v1

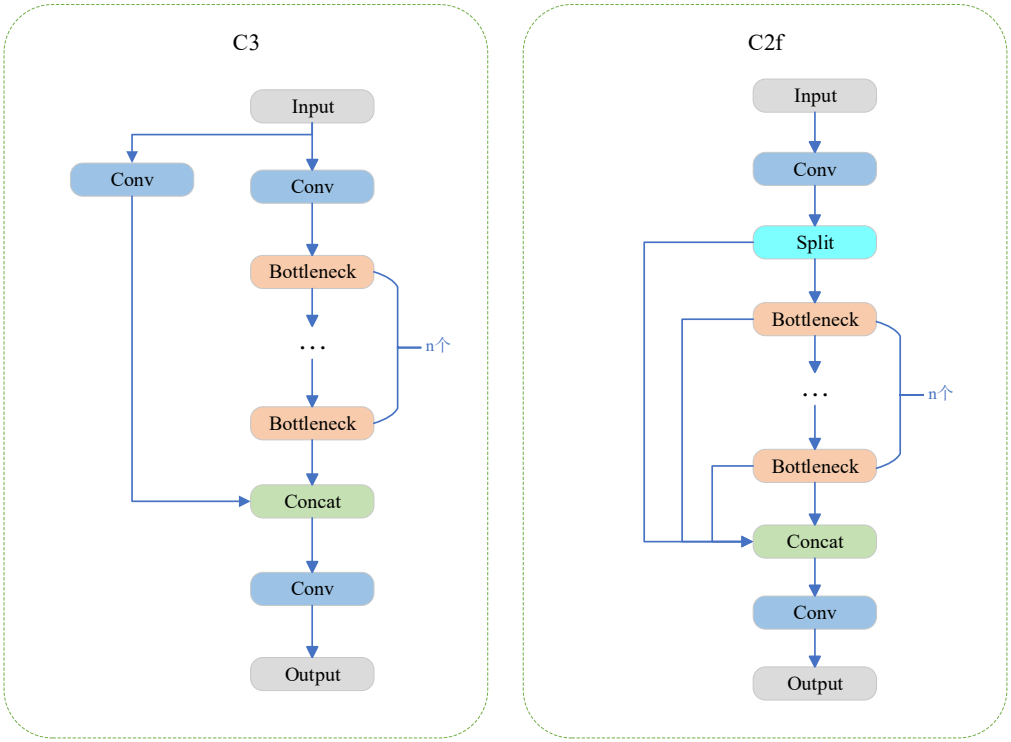


Figure 4. Comparison of C3 and C2f network structures.

In this study, we replace the C3 module in the Backbone and Neck components of the YOLOv5s model with the C2f module to enhance the model's ability to detect small targets. The C2f module [29] utilizes feature fusion, effectively stacking feature maps from various layers to create deeper feature representations by concatenating them along the channel dimension. This fusion strategy enables the C2f module to more efficiently capture the features of small targets, thereby improving detection performance. Moreover, the C2f module enhances the transfer of gradient flow information, allowing the model to concentrate on critical features of small targets, which improves both accuracy and effectiveness. Additionally, the C2f module addresses the challenge of detecting targets at varying scales by generating feature maps that are high-resolution and rich in semantic information. This is achieved through the fusion of shallow feature maps, which provide high resolution but less semantic context, with deep feature maps that offer rich semantic information but lower resolution. This approach is particularly vital for small target detection, as the semantic content from the deep feature maps significantly contributes to improved detection accuracy. The structure of the modified network is illustrated in Figure 2.

2.2. Coordinate Attention Mechanism

To address the issues of false and missed detections caused by target overlapping and occlusion, while enhancing the model's ability to focus on important regions or features within an image, we incorporate the Coordinate Attention mechanism into the Neck and Head stages of the YOLO model. This integration aims to improve the accuracy of pedestrian detection in complex scenes, and the structure of the Coordinate Attention module [30] is illustrated in Figure 2.Coordinate Attention decomposes channel attention into two one-dimensional feature encoding processes that aggregate features along two spatial directions. The resulting attention maps are applied complementarily to the input feature maps, enhancing the representation of the target object. The network structure of the Coordinate Attention module comprises four key components: an average pooling layer, concatenation and convolution operations, batch normalization and nonlinear activation, and feature separation and reweighting. This structure is illustrated in Figure 5.

 doi:10.20944/preprints202410.2326.v1

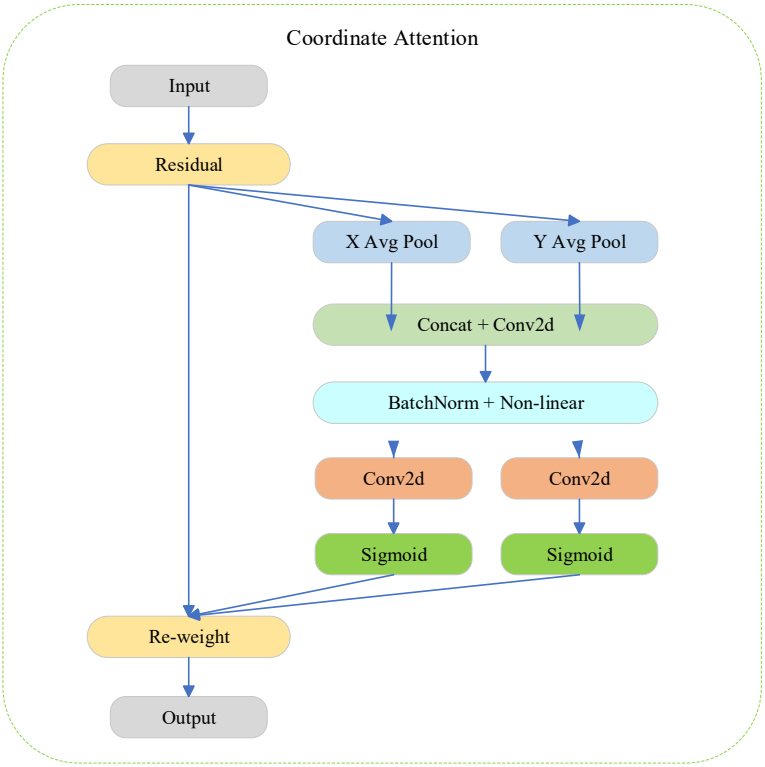


Figure 5. Coordinate Attention structure.

The Coordinate Attention module encodes channel relationships and long-term dependencies using precise positional information, allowing the network to concentrate on significant regions while maintaining a low computational cost. This process involves two main steps: Coordinate Information Embedding and Coordinate Attention Generation.

Step 1: Coordinate Information Embedding. To capture attention regarding the image's width and height while encoding precise location information, the input feature map is initially globally pooled along both the width and height dimensions. Specifically, for the input feature tensor X, the features of each channel are encoded along the horizontal coordinates using a pooling kernel of size $(H,1)$. Consequently, the output for channel c at height h can be represented as follows:

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i \le W} x_c(h,i) \tag{3}$$

where z_c^h denotes the height of the c th channel as h; $x_c(h,i)$ denotes the value of the feature map with width coordinate i for the height of the c th channel as h; and W denotes the width of the feature map. Similarly, the output of the width of the c th channel as w can be written as Equation 4.

$$z_c^w(w) = \frac{1}{H} \sum_{0 \le j \le W} x_c(j,w) \tag{4}$$

where z_c^h denotes the height of the c th channel as w; $x_c(j,w)$ denotes the value of the feature map with width coordinate j for the height of the c th channel as w; and H denotes the width of the feature map.

These two transformations aggregate features from two spatial directions, resulting in a pair of direction-aware feature maps. This dual transformation enables the attention module to capture long-term dependencies along one spatial dimension while preserving precise positional information along the other. As a result, the model is better equipped to localize the target of interest effectively.

 doi:10.20944/preprints202410.2326.v1

Step 2: Coordinate Attention Generation. The aggregated feature maps generated by Equation 3 and 4 undergo a concatenation operation, after which they are transformed using the 1×1 convolutional function F_1. This process yields intermediate feature maps f, which encode spatial information in both the horizontal and vertical directions, as shown in Equation 5 :

$$f = \delta(F_1([z^h, z^w])) \tag{5}$$

Where [ ] denotes a concatenation operation along the spatial dimension, δ represents a nonlinear activation function, $f \in R^{C/r\times(H+W)}$ is an intermediate feature map that encodes spatial information in both the horizontal and vertical directions, and r is a reduction rate used to control the size of the Squeeze-and-Excitation (SE) block. The decomposition of f into two independent tensors, $f^h \in R^{C/r\times H}$ and $f^w \in R^{C/r\times W}$, along the spatial dimension, and the transformation of f^h and f^w into tensors with the same number of channels as the input X using two additional 1×1 convolutional transforms, F_h and F_w, respectively, as shown in Equation 6 and 7:

$$g^h = \delta(F_h(f^h)) \tag{6}$$

$$g^w = \delta(F_w(f^w)) \tag{7}$$

where δ is the Sigmoid activation function. To reduce model complexity and computational overhead, the number of channels in f is typically decreased using appropriate reduction ratios. Subsequently, the outputs g^h and g^w are expanded to serve as the attention weights, respectively. The final output of the Coordinate Attention (CA) module, denoted as $Y = [y_1, y_2, ..., y_c]$, can be expressed as Equation 8:

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j) \tag{8}$$

2.3. Improved WiseIoU Loss

The loss function utilized in bounding box regression (BBR) is pivotal for pedestrian target detection, as a well-defined loss function can significantly enhance model performance. In traffic scenarios, issues such as omissions or misdetections can result in incorrect filtering of pedestrian prediction frames. This problem is particularly evident when the actual frame does not overlap with the predicted frame. The traditional Intersection over Union (IoU) loss function often suffers from gradient vanishing, which hinders the model's convergence speed and ultimately compromises detection accuracy.To address this challenge, we introduces an enhancement to the loss function derived from the YOLOv5 framework, specifically by replacing the original Complete IoU (CIoU) Loss with WiseIoU Loss. This modification aims to improve detection accuracy in complex traffic environments.

WiseIoU Loss is an optimized loss function [31] designed for target detection that employs a pixel-level weighting strategy to more accurately evaluate the correspondence between the predicted frame and the actual frame, with its computational principles illustrated in Figure 6. This enhancement enables WiseIoU Loss to better capture the effectiveness of target detection, thereby improving overall detection accuracy.Compared with CIoU,, WiseIoU Loss offers a more comprehensive assessment of the relationship between the target area and its surrounding context. It not only emphasizes the overlapping regions of the two frames but also incorporates surrounding area information, resulting in a more precise evaluation of target detection. This capability provides WiseIoU Loss with a significant advantage in complex scenarios and when handling irregularly shaped targets.Furthermore, WiseIoU Loss features the flexibility to dynamically adjust weights. By modifying the weight matrix, the importance of various regions can be tailored to meet specific task requirements. This adaptability allows WiseIoU Loss to effectively accommodate diverse datasets and target detection challenges.

 doi:10.20944/preprints202410.2326.v1

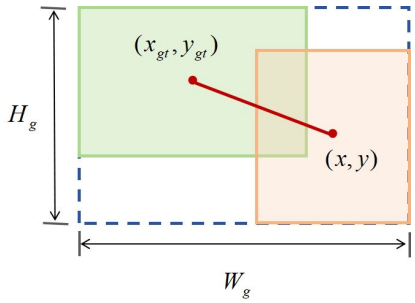


Figure 6. WiseIoU calculation block diagram.

As shown in Equation 9, WiseIoU Loss incorporates additional factors, including the length and width of the target frame, as well as the distance from the center point, during the calculation process. These factors enable a more precise reflection of the similarity between the predicted and actual frames, thereby yielding a more accurate loss value. By optimizing this loss value, the model can learn a more effective method for regressing the target frame and accurately assess the discrepancies between the predicted and actual frames. This leads to more effective optimization and improved detection performance.

$$WiseIoU = exp\left(\frac{(x-x_{gt})^2 + (y-y_{gt})^2}{(w_g^2 + H_g^2)^*}\right) \tag{9}$$

3. Experimental Results

To evaluate our approach, we conducted algorithm validation on the INRIA pedestrian dataset and performed comparative experiments with the pre-improvement algorithm. All experiments were executed on a server running a Windows operating system, utilizing Python 3.8.18 as the programming language, PyTorch 2.2.1 as the deep learning framework, and CUDA 11.8 as the parallel computing platform. The experiments were carried out on NVIDIA GeForce RTX 4060 GPUs. During the training process, we set the epoch size to 300 and the batch size to 16, with an initial learning rate of 0.01. The image resolution was configured to 640×640 pixels, while all other parameters were maintained at their default values.

3.1. Dataset Description

The INRIA Pedestrian Dataset [32] is a standard dataset, widely used in computer vision, dedicated to pedestrian detection tasks. Created by the French National Institute for Information and Automation Research (INRIA), the dataset aims to provide researchers with a unified benchmark for evaluating and comparing the performance of pedestrian detection algorithms.The images in the INRIA Pedestrian Dataset typically have a resolution of 640x480 pixels and encompass a diverse range of scenes, including urban streets and public areas, reflecting the variety of pedestrian appearances under different lighting and background conditions. The dataset consists of a total of 2,475 images, with 1,218 images designated for training and 1,257 images for testing.

3.2. Evaluation Metrics

The evaluation metrics provide a comprehensive assessment of the performance and effectiveness of the improved YOLO model from various perspectives. This paper primarily evaluates the algorithm using the following metrics:

(1) Precision: Precision refers to the proportion of true positive samples among all samples predicted as positive by the model. It serves as a measure of the model's accuracy in predicting the target. The calculation formula for precision is presented in Equation 10.

$$Presicion = TP/(TP + FP) \tag{10}$$

 doi:10.20944/preprints202410.2326.v1

In our research, targets are classified as positive examples, while non-targets are deemed negative examples. Specifically, we define the following terms:

- TP (True Positives): These are true cases, meaning positive instances that are accurately identified as such by the model.
- FP (False Positives): These refer to pseudo-positive cases, which are instances that the model incorrectly identifies as positive, although they are actually negative.

(2) Recall: Recall measures the proportion of all actual positive samples that are correctly identified by the model. This metric assesses the model's ability to accurately detect all positive instances. The calculation formula for recall is presented in Equation 11.

$$Recall = TP/(TP + FN) \tag{11}$$

In this context, we define the following terms:

- FN (False Negatives): These are pseudo-negative cases, meaning instances that are actually positive but have been incorrectly identified as negative by the model.
- TN (True Negatives): These are true negative cases, referring to instances that are accurately identified as negative by the model.

(3) Mean Average Precision (mAP): mAP is a crucial evaluation metric in target detection tasks. It calculates the average accuracy across different thresholds of Average Precision (AP), which represents the area under the Precision-Recall curve. This metric provides a comprehensive assessment of the model's performance across various categories. A higher mAP value indicates that the model is more effective at detecting targets across different classes. The relevant formulas are presented in Equations 12 and 13.

$$AP = \int_0^1 P(r)dr \tag{12}$$

$$mAP = \frac{AP\_1 + AP\_2 + AP\_3 + \ldots + AP\_n}{n} \tag{13}$$

where $P(r)$ is the precision when the recall is r.

3.3. Comparison Experiment

We conduct a comparative analysis between the original YOLOv5s model featuring the C3 module and the improved YOLOv5s model incorporating the C2f module. We calculate their precision and recall metrics. Additionally, we evaluate the mean average precision (mAP) at an IoU threshold of 0.5 (mAP_0.5) and the mean average precision across IoU thresholds ranging from 0.5 to 0.95 (mAP_0.5:0.95). The results of this analysis are presented in Table 1.Data analysis indicates that replacing the C3 module with the C2f module in the Backbone and Neck sections of the YOLOv5s model results in a significant enhancement of model performance. The original YOLOv5s model achieves a training accuracy of 88.3%, a recall of 75.8%, an average precision of mAP_0.5 at 86.9%, and a mAP_0.5:0.95 of 56.8%. Following the replacement, accuracy improves by 3.2%, while mAP_0.5 and mAP_0.5:0.95 increase by 3.3% and 1.9%, respectively. Although there is a slight decrease in recall by 0.4%, these results clearly demonstrate that the C2f module effectively enhances the overall performance of the YOLOv5s model.

Table 1. Comparison table of network structure optimization experiments.

Network Models	Precision(%)	Recall(%)	mAP_0.5(%)	mAP_0.5:0.95(%)
YOLOv5s	88.3	75.8	86.9	56.8
YOLOv5s+C2f	91.5	75.4	90.2	58.7

To assess the detection performance of the algorithm incorporating the Coordinate Attention module, we compared the improved model against both the original YOLOv5 model and the version utilizing the SE attention mechanism. The results of this comparison are presented in Table 2. By contrast, the YOLOv5s fusion network model incorporating the Coordinate Attention module

 doi:10.20944/preprints202410.2326.v1

demonstrates improvements of 4.0% in accuracy, 5.6% in recall, and 5.5% in both mAP_0.5 and mAP_0.5:0.95. These results indicate that the Coordinate Attention module not only surpasses the performance of the original YOLOv5s model but also outperforms the model utilizing the SE module.

Table 2. Comparison of different attention mechanisms.

Network Models	Precision(%)	Recall(%)	mAP_0.5(%)	mAP_0.5:0.95(%)
YOLOv5s	88.3	75.8	86.2	56.8
YOLOv5s+SE	89.7	80.2	90.2	56.4
YOLOv5s+CA	92.3	81.4	92.4	62.3

To evaluate the impact of WiseIoU Loss on enhancing the detection capabilities of the YOLOv5s algorithm, we conducted comparative experiments using three bounding box loss functions: CIoU, EIoU, and WiseIoU. The results of this comparison are presented in Table 3.Experimental results demonstrate that replacing CIoU Loss with EIoU Loss and WiseIoU Loss in the YOLOv5s model significantly enhances detection performance. The original CIoU Loss model achieves an accuracy of 88.3%, a recall of 75.8%, an average precision of mAP_0.5 at 86.9%, and an mAP_0.5:0.95 of 56.8% on the self-constructed pedestrian dataset. When using EIoU Loss, accuracy increases by 0.4%, recall improves by 2.7%, and mAP_0.5 rises by 1.5%, although mAP_0.5:0.95 experiences a slight decline of 1.2%. In contrast, the introduction of WiseIoU Loss yields more pronounced improvements: accuracy increases by 1.5%, recall by 4.5%, and both mAP_0.5 and mAP_0.5:0.95 rise by 2.7% and 1.7%, respectively. These findings indicate that WiseIoU Loss significantly outperforms both CIoU Loss and EIoU Loss in enhancing the performance of the YOLOv5s algorithm, highlighting its superior effectiveness in bounding box regression tasks.

Table 3. Comparison of different loss functions.

Network Models	Precision(%)	Recall(%)	mAP_0.5(%)	mAP_0.5:0.95(%)
YOLOv5s+CIoU	88.3	75.8	86.9	56.8
YOLOv5s+EIoU	88.6	78.5	88.4	55.9
YOLOv5s+WiseIoU	89.8	80.3	89.6	58.5

We conducted eight sets of ablation experiments and recorded the training results on INRIA public dataset, as presented in Table 4. The findings indicate that the improved CCW-YOLO algorithm demonstrates a significant enhancement in detection accuracy. Specifically, when employing the C2f module and the Coordinate Attention module, the average precision (mAP_0.5) reaches 93.1%, while mAP_0.5:0.95 is 63.6%. These results represent improvements of 6.2% and 6.8%, respectively, compared to the original YOLOv5s algorithm. Additionally, when compared to the results obtained using the C2f module alone, there are improvements of 2.9% and 4.9%, and improvements of 0.7% and 1.3% when compared to the Coordinate Attention module alone.In contrast, the combination of the C2f module and WiseIoU Loss yields an mAP_0.5 of 90.1% and an mAP_0.5:0.95 of 56.9%, representing enhancements of 3.2% and 0.1% over the original YOLOv5s algorithm. Ultimately, when the three improvements—C2f module, Coordinate Attention module, and WiseIoU Loss—are applied simultaneously, the average precision reaches 95.6%, with mAP_0.5:0.95 at 65.9%, marking improvements of 8.7% and 9.1% compared to the original YOLOv5s algorithm, respectively. These results clearly demonstrate that the simultaneous application of all three improvements yields the most significant performance enhancement for the YOLOv5s algorithm model.

Table 4. Comparative experimental results on INRIA database.

YOLOv5s	C2f	CA	WiseIoU	mAP_0.5(%)	mAP_0.5:0.95(%)
√				86.9	56.8
√	√			90.2	58.7

 doi:10.20944/preprints202410.2326.v1

√		√		92.4	62.3
√			√	89.6	58.5
√	√	√		93.1	63.6
√	√		√	90.1	58.9
√		√	√	92.6	62.7
√	√	√	√	95.6	65.9

Finally, we present the experimental results for all comparison methods alongside the CCW-YOLO method on the dataset, as shown in Table 5. The results indicate that CCW-YOLO outperforms existing methods, achieving the highest accuracy of 95.6%, which represents a 8.7% improvement over the original YOLOv5s algorithm. This finding further validates the effectiveness of our proposed method in the pedestrian detection task.

Table 5. Comparative experimental results on INRIA database.

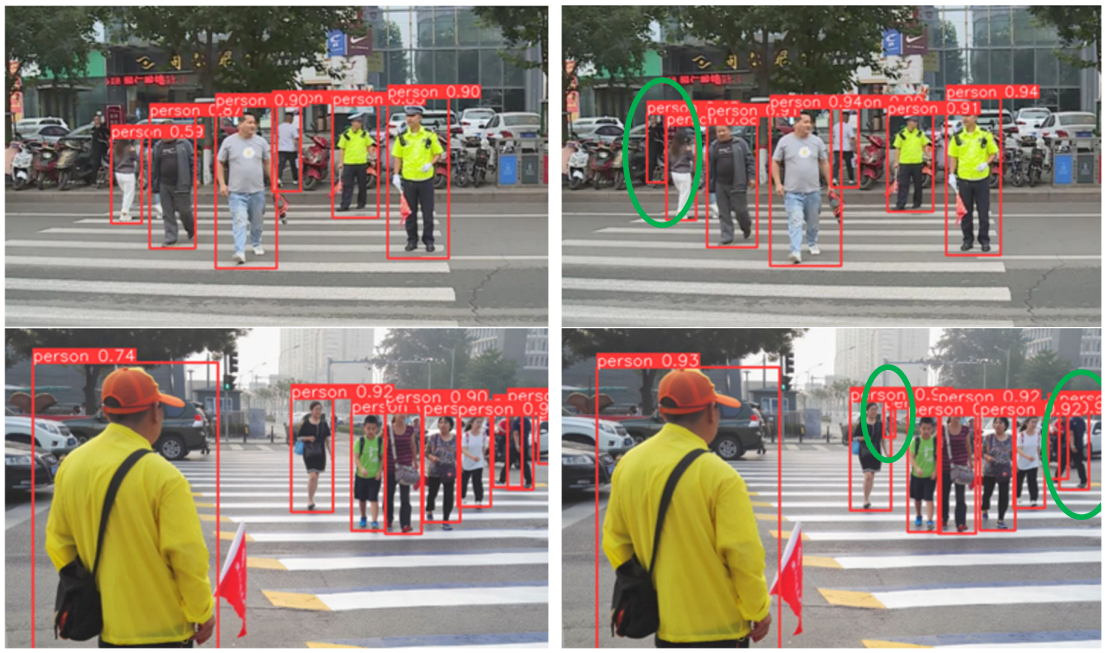
Models	Precision(%)	Recall(%)	mAP_0.5(%)	mAP_0.5:0.95(%)
Faster R-CNN	90.5	76.6	84.9	57.7
YOLOv5s	88.3	75.8	86.9	56.8
YOLOv7	91.5	75.4	90.2	58.7
YOLOv8	92.3	81.4	92.4	62.3
CCW-YOLO	95.6	83.5	94.2	65.9

3.4. Visual Assessment

We evaluated the performance of the improved CCW-YOLO model against the original YOLOv5s model on a test dataset captured during two different times of day: daytime and evening, within a complex traffic scene. The comparative results of the detection are illustrated in Figure 7. Specifically, Figure 7(a) presents the detection results from the original YOLOv5s algorithm, while Figure 7(b) displays the detection outcomes of the three enhanced CCW-YOLO algorithms.

As illustrated in Figures 7(a), the results obtained using the original YOLOv5s algorithm for detection are not satisfactory. The presence of occlusion between vehicles and pedestrians leads to false detections, while the detection accuracy for some pedestrian targets is relatively low. Additionally, the similarity between pedestrians and the background contributes to instances of missed detection. Consequently, the overall performance of the original YOLOv5s algorithm is rather limited.In contrast, Figures 7(b) demonstrates that the improved CCW-YOLO algorithm exhibits significant enhancements in detection results. The modified algorithm shows greater accuracy in frame localization and is capable of comprehensively detecting all pedestrian targets within the image. Furthermore, compared to the original algorithm, the improved CCW-YOLO algorithm efficiently recognizes pedestrian targets that closely resemble the background and effectively corrects previous false detection issues. Additionally, the enhanced CCW-YOLO algorithm shows a marked improvement in detection accuracy.

 doi:10.20944/preprints202410.2326.v1



(**a**) (**b**)

Figure 2. Comparison of pedestrian detection results between the original and improved algorithms: (**a**) YOLOv5s; (**b**) CCW-YOLO.

4. Conclusions

In this paper, we propose an improved YOLO algorithm, termed CCW-YOLO, designed to address the challenges of false and missed detections in pedestrian detection within complex traffic scenarios. Our algorithm significantly reduces the model's computational complexity by incorporating a lightweight convolutional layer, GhostConv, while preserving robust feature extraction capabilities. Additionally, we enhance the feature fusion process through an optimized C2f module, which increases the network's adaptability across diverse environments.Moreover, we integrate a Coordinate Attention mechanism to enable more precise target localization, thereby further improving detection accuracy. To refine the bounding box loss function, we employ WiseIoU Loss, which enhances the model's adaptability to various detection scenarios. Experimental results on the INRIA public dataset demonstrate that CCW-YOLO achieves a detection accuracy of 94.2%, surpassing the original YOLOv5s by 8.7%. This advancement effectively addresses pedestrian detection challenges in complex environments and offers technical support for enhancing traffic safety.

Looking ahead, future research will focus on further enhancing the performance and applicability of the CCW-YOLO algorithm. We plan to develop larger, specialized datasets to improve the model's generalization capabilities, particularly its adaptability to different traffic conditions. Additionally, we aim to optimize data preprocessing techniques and model architecture to bolster feature representation. Exploring the integration of CCW-YOLO with other deep learning frameworks is also a priority, as this could strengthen its robustness in intricate scenarios. Through these research avenues, we aspire to deliver more efficient pedestrian detection solutions that contribute to the advancement of intelligent transportation systems and related technologies.

Author Contributions: Conceptualization, Z.W. and J.D.; methodology, Z.W. and S.Y.; software, Y.L.; validation, J.D.; formal analysis, H.Q.; investigation, Z.W. and J.D.; resources, Z.W.; data curation, J.D.; writing—original draft preparation, Z.W. and J.D.; writing—review and editing, Z.W.、S.Y. and J.D.; visualization, Z.W. and J.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by National Natural Science Foundation of China, grant number 62301241, in part by the Key Research Program of Higher Education Institutions in Henan Province, grant number 25A510017.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors would like to acknowledge the anonymous reviewers and editors whose thoughtful comments helped to improve this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kesting, Arne, Martin Treiber, and Dirk Helbing. "Enhanced intelligent driver model to access the impact of driving strategies on traffic capacity." Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 368.1928 (2010): 4585-4605.
2. Xia, Qin, et al. "Test scenario design for intelligent driving system ensuring coverage and effectiveness." International Journal of Automotive Technology 19 (2018): 751-758.
3. Gui, Guan, et al. "Machine learning aided air traffic flow analysis based on aviation big data." IEEE Transactions on Vehicular Technology 69.5 (2020): 4817-4826.
4. Cui, Xinyi, et al. "3d haar-like features for pedestrian detection." 2007 IEEE International Conference on Multimedia and Expo. IEEE, 2007.
5. Wei, Yun, Qing Tian, and Teng Guo. "An improved pedestrian detection algorithm integrating haar-like features and hog descriptors." Advances in Mechanical Engineering 5 (2013): 546206.
6. Zhou, Hongzhi, and Gan Yu. "Research on pedestrian detection technology based on the SVM classifier trained by HOG and LTP features." Future Generation Computer Systems 125 (2021): 604-615.
7. Cai, Yingfeng, et al. "Research on pedestrian detection technology based on improved DPM model." 2017 IEEE 7th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER). IEEE, 2017.
8. Khemmar, Redouane, et al. "Real time pedestrian detection-based faster hog/dpm and deep learning approach." SITIS-International Conference on Signal Image Technology & Internet Based Systems. 2019.
9. Szarvas, Mate, et al. "Pedestrian detection with convolutional neural networks." IEEE Proceedings. Intelligent Vehicles Symposium, 2005.. IEEE, 2005.
10. Masita, Katleho L., Ali N. Hasan, and Satyakama Paul. "Pedestrian detection using R-CNN object detector." 2018 IEEE Latin American Conference on Computational Intelligence (LA-CCI). IEEE, 2018.
11. Zhang, Shifeng, et al. "Occlusion-aware R-CNN: Detecting pedestrians in a crowd." Proceedings of the European conference on computer vision (ECCV). 2018.
12. Hung, Goon Li, et al. "Faster R-CNN deep learning model for pedestrian detection from drone images." SN Computer Science 1 (2020): 1-9.
13. Hsu, Wei-Yen, and Wen-Yen Lin. "Ratio-and-scale-aware YOLO for pedestrian detection." IEEE transactions on image processing 30 (2020): 934-947.
14. Fan, Di, et al. "Improved ssd-based multi-scale pedestrian detection algorithm." Advances in 3D Image and Graphics Representation, Analysis, Computing and Information Technology: Algorithms and Applications, Proceedings of IC3DIT 2019, Volume 2. Springer Singapore, 2020.
15. Liu, Congqiang, Haosen Wang, and Chunjian Liu. "Double Mask R-CNN for Pedestrian Detection in a Crowd." Mobile Information Systems 2022.1 (2022): 4012252.
16. Zhang, Shanshan, Christian Bauckhage, and Armin B. Cremers. "Informed haar-like features improve pedestrian detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.
17. Gawande, Ujwalla, Kamal Hajari, and Yogesh Golhar. "Pedestrian detection and tracking in video surveillance system: issues, comprehensive review, and challenges." Recent Trends in Computational Intelligence (2020): 1-24.
18. Mao, Xiao-Jiao, et al. "Enhanced deformable part model for pedestrian detection via joint state inference." 2015 IEEE International Conference on Image Processing (ICIP). IEEE, 2015.
19. Cao, Jiale, et al. "From handcrafted to deep features for pedestrian detection: A survey." IEEE transactions on pattern analysis and machine intelligence 44.9 (2021): 4913-4934.
20. Sha, Mingzhi, and Azzedine Boukerche. "Performance evaluation of CNN-based pedestrian detectors for autonomous vehicles." Ad Hoc Networks 128 (2022): 102784.
21. Xie, Jin, et al. "Count-and similarity-aware R-CNN for pedestrian detection." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16. Springer International Publishing, 2020.

 doi:10.20944/preprints202410.2326.v1

22. Maity, Madhusri, Sriparna Banerjee, and Sheli Sinha Chaudhuri. "Faster r-cnn and yolo based vehicle detection: A survey." 2021 5th international conference on computing methodologies and communication (ICCMC). IEEE, 2021.
23. Murthy, Chintakindi Balaram, Mohammad Farukh Hashmi, and Avinash G. Keskar. "Optimized MobileNet+ SSD: a real-time pedestrian detection on a low-end edge device." International Journal of Multimedia Information Retrieval 10.3 (2021): 171-184.
24. Huang, Lincai, Zhiwen Wang, and Xiaobiao Fu. "Pedestrian detection using RetinaNet with multi-branch structure and double pooling attention mechanism." Multimedia Tools and Applications 83.2 (2024): 6051-6075.
25. Shao, Yifan, et al. "Aero-YOLO: An Efficient Vehicle and Pedestrian Detection Algorithm Based on Unmanned Aerial Imagery." Electronics 13.7 (2024): 1190.
26. Gao, Fei, et al. "Improved YOLOX for pedestrian detection in crowded scenes." Journal of Real-Time Image Processing 20.2 (2023): 24.
27. Zhao, Siqi, et al. "Improved YOLOv5 Algorithm for Intensive Pedestrian Detection." International Conference on Computational & Experimental Engineering and Sciences. Cham: Springer Nature Switzerland, 2023.
28. Cao, Jinshan, et al. "GCL-YOLO: A GhostConv-based lightweight yolo network for UAV small object detection." Remote Sensing 15.20 (2023): 4932.
29. Pan, Jingmin, et al. "C2F-YOLO: A Coarse-to-Fine Object Detection Framework Based on YOLO." Proceedings of the 2024 3rd Asia Conference on Algorithms, Computing and Machine Learning. 2024.
30. Xie, Chao, Hongyu Zhu, and Yeqi Fei. "Deep coordinate attention network for single image super-resolution." IET Image Processing 16.1 (2022): 273-284.
31. Tong, Zanjia, et al. "Wise-IoU: bounding box regression loss with dynamic focusing mechanism." arXiv preprint arXiv:2301.10051 (2023).
32. Taiana, Matteo, Jacinto C. Nascimento, and Alexandre Bernardino. "An improved labelling for the INRIA person data set for pedestrian detection." Pattern Recognition and Image Analysis: 6th Iberian Conference, IbPRIA 2013, Funchal, Madeira, Portugal, June 5-7, 2013. Proceedings 6. Springer Berlin Heidelberg, 2013.