Improved Pedestrian Vehicle Detection for Small Objects Based on Attention Mechanism

Yanpeng Hao
Xi'an University of Technology
School of Computer Science and Engineering
Xi'an, China
E-mail: hnrank@163.com

Chaoyang Geng
Xi'an University of Technology
School of Computer Science and Engineering
Xi'an, China
E-mail: 541211200@qq.com

***Abstract*—This study aims to solve the low detection accuracy and susceptibility to false detection and omission in pedestrian and vehicle detection by proposing an improved YOLOv5s algorithm. Firstly, a small target detection module is added to better acquire and determine the information of pedestrians from long-range vehicles. Secondly, the multi-scale channel attention CBAM attention module is added, and the dual attention mechanism is not only flexible and convenient, but also improves the computational efficiency. Finally, the MPDIoU loss function based on minimum point distance is introduced to replace the original GIoU loss function, and this change not only enhances the regression accuracy of the model. At the same time, the convergence speed of the model is accelerated. KITTI data set was used for experiments, and the experimental results showed that the average accuracy of the model trained by the improved YOLOv5s algorithm on the data set reached 84.9%, which was 3.7% higher than that of the original YOLOv5s algorithm. It is verified that the model is suitable for high accuracy of pedestrian and vehicle recognition in complex environments, and has high value for promotion.**

Keywords-Deep Learning; Small Target Detection; CBAM; MPDIoU; Vehicle Pedestrian Detection

I. INTRODUCTION

Along with the continuous progress of computer technology, the field of computer vision is rapidly rising, and the detection of pedestrians and vehicles is becoming more and more critical in many real-life scenarios [1]. The detection of pedestrians and vehicles is a basic task in the field of computer vision, which has a wide range of applications in the industries of automatic driving, intelligent transportation, video surveillance and human flow analysis [2]. Along with the increase of urban population and the rapid prosperity of finance, the need for pedestrian detection and vehicle detection is increasing day by day. Pedestrian and vehicle detection is a key aid in reducing traffic accidents and alerting drivers. Target detection plays a key role in traffic management, surveillance security and the development of smart cities. Despite the maturity of existing technologies, missed and false detections still occur in the field of vehicle pedestrian detection, especially due to the relatively small size of the pedestrians and their frequent occlusion. Therefore, it is urgent and important to overcome this challenge by continuously optimising the performance of small target detection.

The YOLO algorithm proposed by Redmon et al. is a regression-centred algorithm, which converts the target detection problem into a regression problem by partitioning the target into a grid, an innovation that enables target detection. On the basis of YOLO, Redmon team further developed YOLOv2 version [3]. YOLOv3, which adopts Darknet-19 as a solid foundation and incorporates the residual module, significantly improving the performance of feature extraction [21]. YOLOv3, using Darknet-19 as a solid foundation and incorporating the residual module, significantly improves the performance of feature extraction. The logistic regression improvement layer is used to improve the accuracy of multi-label classification. In order to comprehensively acquire the multi-scale features of the target, the feature

pyramid (FPN) concept is added, which fuses the semantic depth of the high level of the image with the detailed visual information of the low level, which in turn improves the comprehensiveness of the target detection. While Bochkovskiy et al. inherited and developed YOLOv3, the innovative YOLOv4 is the refinement and enhancement of the previous algorithm. YOLOv4 algorithm combines the Cross-Stage Partial Network (CSPNet) with Darknet-53 as the backbone, which enhances its deep feature extraction function, and introduces the SPP (Spatial Pyramid Pooling) to deal with different scales of physical information, and the PANet.

Researchers Zhiyong Ju et. al [6]. innovatively designed Vit-YOLOv4 , a model that incorporates the Transformer architecture and deeply separable convolutional techniques to improve detection accuracy. Zihao Jia et al [22]. added a lightweight sub-pixel convolutional layer in front of the detection layer of the YOLOv5 model, which significantly improved the inference speed of the model despite sacrificing some of the detection accuracy in the lightweighting process. Although the above studies have partially contributed to the field of pedestrian detection, however, the challenge still exists when dealing with distant and dense scenes, and the problem of false and missed detections still needs to be overcome [7].

In this paper, YOLOv5s is chosen as the base model, and the model is lightweighted by improving the convolutional block, introducing a small target detection layer and a multi-scale channel attention mechanism CBAM. MPDIoU loss function is used to optimize the model training, and experiments are carried out on the constructed vehicle and pedestrian data sets. The experimental data prove that the improved YOLOv5s algorithm has improved the accuracy and practicality.

II. INTRODUCTION TO THE YOLOv5 ALGORITHM

YOLOv5 provides a series of models of different sizes, which can be classified into five structures: n, s, m, l, and x, based on the network depth and model size of the network model [8]. The depth of these five models ranges from shallow to deep, and the speed of detection ranges from fast to slow, so users can choose the appropriate model according to the application scenario. In this paper, for vehicle pedestrian detection, we need to ensure the real-time detection, so we choose YOLOv5s, which has a better balance between accuracy and speed, as the base model [9]. Figure 1 below shows the algorithm flow of YOLOv5s.

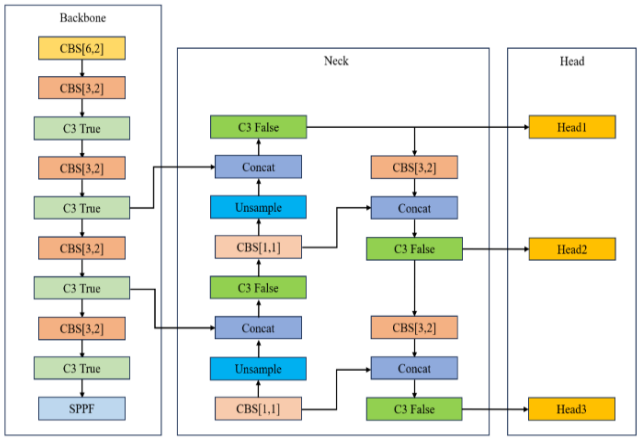

Figure 1. Flowchart of YOLOv5s algorithm

As a lightweight model of the YOLOv5 family, the YOLOv5s is designed to significantly reduce the need for computing resources while maintaining high detection accuracy, making it particularly suitable for resource-constrained environments. Its design is exquisite, thanks to the powerful function of CSPDarknet53, it can efficiently extract features and traverse the multi-level information of the input image [10]. YOLOv5s uses a multi-scale prediction strategy to process three feature maps with different resolutions, ensuring excellent detection performance regardless of the size of the target. More innovative is that it introduces an adaptive anchor frame mechanism to dynamically generate a frame that matches the target size, improving the detection accuracy [11]. In addition, the model incorporates training optimization methods such as data enhancement and more refined learning rate adjustment, and the integration of these strategies significantly improves the overall performance of YOLOv5s in various detection tasks.

YOLOv5s is designed with a hierarchical architecture, including an input preprocessing module; a powerful Backbone network, through which the core features are extracted; a feature

fusion neck network, which acts as a connecting point for feature fusion and effectively integrates the features at different levels; and finally an output prediction Head, which is responsible for generating the prediction results [12].

The core feature extraction network is a key component of the main network, and the feature extraction module is carefully designed, which combines the C3 structure module, the CBS convolution block and the SPPF spatial pyramid pooling module. The CBS module is the basic architecture, which employs traditional convolutional layers and selects the SiLU nonlinear activation function to enhance the expressive capability [13]. The C3 module introduces residual connectivity, which significantly improves the deep learning effect of the model; while the SPPF module is upgraded from the SPP module, which uses the stacked connection of small-sized pooling kernels of the same size to construct receptive fields of different sizes, which are used to obtain more detailed size of object information

The basic idea of the neck link network is to organically integrate the characteristic pyramid network and the route network, which requires effective integration of multidimensional characteristics in order to improve the accuracy and reliability of vehicle and pedestrian detection. Feature pyramid network structure transmits deep semantic content to shallow features in a top-down manner, while the path aggregation network structure transmits shallow location content to deep features. This mutually complementary structural design achieves multi-scale feature fusion by integrating semantic and location information in the feature map, through which the feature map incorporates rich semantic and precise location information, thus improving the detection accuracy and depth of feature maps at different scales [14].

The head prediction network performs classification and regression operations on small, medium and large targets respectively to achieve the prediction of target category and location. YOLOv5s is a popular target detection algorithm, which has stable detection effect on many datasets [15]. However, there is still room for improvement in detection accuracy when facing small target detection and target detection in complex situations [16]. In this study, it is intended to make improvement and optimization so that YOLOv5s can better fulfill the detection task in various missions, so as to improve the detection accuracy when facing small targets and complex situations [17].

By integrating Mosaic data enhancement technology, YOLOv5s enhances the diversity of training data, thus significantly enhancing the generalization performance of the model. The adaptive anchor frame strategy optimizes the detection process and automatically adjusts to the characteristics of the training data, thereby improving the detection accuracy. At the same time, the model can flexibly adapt to the size change of the target to ensure that the target can be accurately and stably detected at different scales. Within the system, a smart combination of Mish focusing and activation functions optimizes model performance. The use of non-maximum suppression techniques can effectively eliminate redetection and ensure the accuracy of the search results. This design structure enables YOLOv5s to achieve a high level of performance and efficiency [18].

III. IMPROVEMENT OF YOLOv5s

A. Small Target Detection

The Head of the YOLOv5 network uses three detection modules at different scales. In real-world scenarios, distant pedestrian and vehicle targets often appear relatively tiny in videos and pictures, and their visible visual size is usually extremely limited, reflecting the extremely small image size. Since the features of smaller targets are not easy to identify, the situation of missed detection and false detection often occurs in the detection process. Facing the problem of missed detection and misdetection due to small target recognition, we adopt the YOLOv5 model and introduce an additional high-resolution detection layer, i.e., Level P1 160 by 160 is designed to improve the accuracy of detection of small objects. The key mission of P1 is to introduce high-resolution feature mapping, which aims to enhance the ability to accurately capture and locate remote pedestrian vehicles in detail. Pedestrians at long distances tend to present smaller sizes in images and videos, and traditional algorithms are prone to miss these

targets, leading to missed detections. The addition of the P1 layer enables the network to better handle the detection of small target pedestrians, better capture and localise the detailed information of long-distance pedestrians, and reduce the occurrence of misdetection and missed detection.

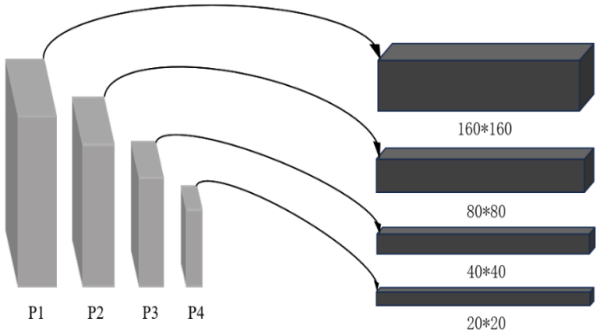

Figure 2. Multi-scale detection structure

This study proposes to introduce a strategy to achieve efficient detection of complex targets by embedding an additional fourth module at the head of the detection layer without extending the classical neck network architecture. This strategyavoids the information loss associated with increasing the depth of the network, while maintaining the savings in computational resources [19].

B. Hybrid Attention Mechanism CBAM

Inspired by human cognitive systems, the attention mechanism enables models to better understand the associations between different locations, and it can highlight important details, providing significant support for deep learning models [20]. Notably, the same attention design can be adapted to handle different data patterns and can be easily integrated into large network models. In addition to this, multiple complementary attention mechanisms can be integrated into the same network.

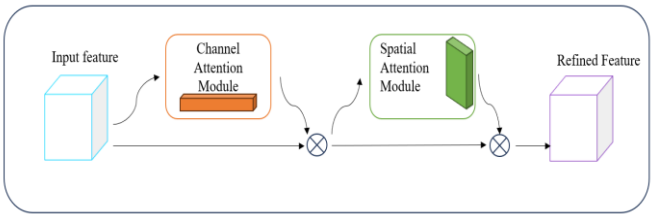

Figure 3. Structure of CBAM module

CBAM (Convolutional Block Attention Module) is an effective and relatively simple attentional mechanism that can be easily integrated into most well-known CNN frameworks, and the structure of the CBAM module is shown in Figure 3. CBAM can perform adaptive function cleaning by having specific function cards flow from two parallel directions of channel and space, respectively, and then multiply function cards to perform adaptive function cleaning. Integration of CBAM into different models has resulted in significant performance improvements on various classification and detection datasets [21]. In vehicle detection, since a large detection area usually contains multiple complex information, using CBAM allows focusing on a part of the area, which helps the network model to resist the ponderous information and focus its core attention on the beneficial objects.

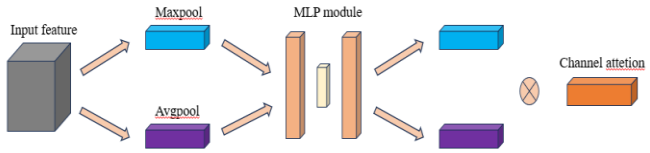

Figure 4. Channel Module

As can be seen from Fig. 4 above, in Channel Attention, the spatial information of the feature maps is integrated, and subsequently, the feature maps after fusing the features are put into it. The maximum maxpool level and the average avgpool level produce two different spatial descriptor contexts, respectively, to obtain a two-element map 1x1xc [22]. Both cards are connected to an artificial neural network, MLP (Multilayer Perceptron), to increase the number of feature channels of the original feature maps, where the MLP has one hidden layer, the number of neurons in the first stage c /r (R-reduction rate) [23]. The number of neurons in the second level is c. This design controls the model complexity to some extent and avoids overfitting. complexity, avoids overfitting, and efficiently extracts higher-order representations of the input features.

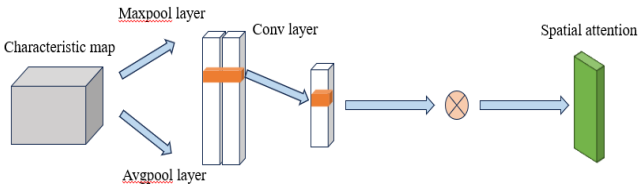

Figure 5. Space module

As shown in Figure 5, in Spatial Attention, maximum pooling on the channel dimension is performed on the feature map to find a maximum value at all positions in the feature map, which in turn produces a completely new feature map. At the same time, this study used the average pooling technique on the channel dimension to produce a new feature map containing aggregated information by averaging the channels over the feature maps at all positions. This processing can better preserve the feature information. The topic map is constrained by a maximum average pool based on channel orientation, and two output maps are superimposed on the channel to format the mapping of elements containing two pools of information.

C. Loss Function (Loss Function) Improvement

In deep learning and machine learning, the loss function plays a non-trivial role in judging the degree of difference between the model's predictions and the true labels. In the vast majority of tasks, choosing the appropriate loss function can directly affect the convergence speed and final performance of the model. By optimising the loss function, the optimal solutions of the model parameters are clearly identified, significantly improving the accuracy of the predictions. During the training phase, model variables such as neural network weights are uninterruptedly adjusted to reduce the loss function values.

In earlier versions of the YOLO series, a loss function dependent on the mean square error was constructed using the position of the centroid of the predicted frame with respect to the real frame and dimensional information (width and height).

In the case of mean square error as a loss function, the centroid coordinates and edge length information are treated as independent variables, however, in essence they are not independent and are related to each other. Therefore, the mean square error is not a good metric to represent the intersection and integration ratio (IOU) relationship between the bounding boxes [24], as in equation (1).

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

The intersection and concatenation schematic of the IoU prediction frame and the real frame is shown in Figure 6.

Figure 6. Schematic diagram of intersection and concatenation of IoU prediction and real frames.

In YOLOv5, the GIoU_Loss loss function is applied instead of the traditional IoU. The GIoU takes into account the relative positional relationship between the predicted box and the real box, which not only focuses on the overlapping part, but also measures the difference in the outer join matrix between the two. The optimisation of the cost function is enhanced by adding a penalty term which calculates the difference between the concatenation and closure of the two boxes, the smaller the difference the smaller the penalty term. The GIoU loss function is calculated as in equation (2).

$$GIoU = IoU - \frac{|C / A \cup B|}{|C|} \tag{2}$$

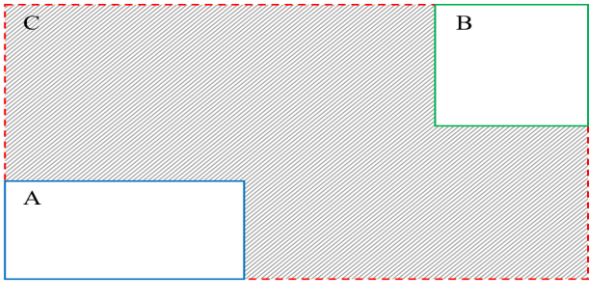

Figure 7. GIoU penalty content for minimising the area of the shaded region

Figure 7 shows the schematic diagram of the GIoU principle, where C is the area of the smallest outer rectangle, A is Ground Truth, and B is Bounding Box.

If there is an inclusion relationship between two bounding boxes, then the GIoU becomes IoU, which cannot accurately express the relative positional relationship between them. The GIoU is also heavily dependent on the IoU, and when the two boxes intersect, horizontal and vertical errors are large, convergence is slow, and do not accurately reflect the size of the overlap between the two squares [25].

The MPDIoU loss function was proposed by MA S L et al. proposed in July 2023, achieved good performance when training the target detection model on the PASCAL VOC dataset [26]. The loss function calculates similarity by calculating the minimum deviation between the desired image and the actual image, that is, taking into account the distance of the overlap area, the center distance, and the width. Through optimisation, the model can converge faster and get more accurate prediction results [27]. In order to enhance the training effect, accelerate the convergence speed of the model and improve the regression accuracy, choose MPDIoU as a new loss function. The MPDIoU calculation principle shown in figure 8 simplifies the calculation process by minimizing the distance between the upper left and lower right corners of the prediction and control cells.

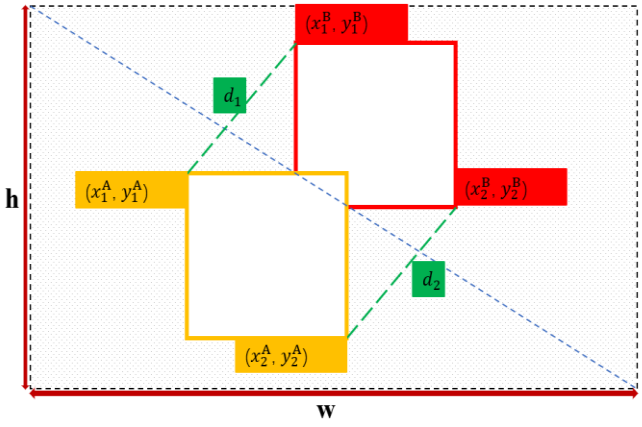

Figure 8. Computational schematic of MPDIoU

The main calculations are shown in Eqs. (3) to (5).d_1 and d_2 signify the distances that separate the upper left corner from the lower right corner of the predicted frame B and its corresponding real frame A. The $\left(x_1^A, y_1^A\right)$ and $\left(x_2^A, y_2^A\right)$ are the coordinates of the upper-left and lower-right corners of the real frame, and the boundary of the prediction frame is described by two points: the upper-left corner $\left(x_1^B, y_1^B\right)$ coordinates and the lower right corner $\left(x_2^B, y_2^B\right)$ coordinates, adopting L_{MPDIoU} as an evaluation metric, which measures the overlap between the predicted box and the real box. The width and height of the input image are denoted as w and h, respectively.

$$d_1^2 = \left(x_1^B - x_1^A\right)^2 + \left(y_1^B - y_1^A\right)^2 \tag{3}$$

$$d_2^2 = \left(x_2^B - x_2^A\right)^2 + \left(y_2^B - y_2^A\right)^2 \tag{4}$$

$$L_{MPDIoU} = 1 - \left(\frac{A \cap B}{A \cup B} - \frac{d_1^2}{w^2 + h^2} - \frac{d_2^2}{w^2 + h^2} \right) \tag{5}$$

IV. Experimental Configuration and Analysis of Results

A. Experimental environment

The operating system trial environment is Windows 11 (64-bit), running on 32GB of RAM, the graphics driver is RTX2070 SUPER, and it is powered by a 12th generation Intel processor i5-12400F.

B. Data sets and evaluation indicators

This study relies on the KITTI dataset, which was created jointly by the Karlsruhe Institute of Technology, Germany, and the Toyota Technological Institute, Chicago, USA. The dataset scenarios are collected from a variety of complex road environments such as rural, motorway and urban from different perspectives and time periods [28]. Especially in vehicle and pedestrian detection tasks are considered important benchmarks for measuring the performance of the technology. 9200 images were selected from the preprocessed dataset for the experiments, dividing the training, validation and test sets in a ratio of 7:1:2.

When evaluating the performance of a model, it is common to rely on two core metrics, precision (P)

and recall (R), which together reveal the accuracy and completeness of the model. (6) to (8) are the calculation formulas used in this study.

$$P = \frac{T_P}{T_P + F_P} \tag{6}$$

$$R = \frac{T_P}{T_P + F_N} \tag{7}$$

$$mAP = \frac{\sum_{i=1}^{N} P_{A_i}}{N} \tag{8}$$

T_p represents the number of instances that are accurately identified as positive; F_p represents the number of instances that are actually negative but are judged to be positive; F_N represents those instances that were actually positive but were misidentified as negative; N is the number of categories of samples in the data set.

C. Impact of different attention mechanisms on the model

In order to investigate the effect of multiple attention mechanisms on the model performance, the C3 module was replaced with ECA, SE, CCA, SA-Net, MS-CAM, CBAM The results of the experiments with six attention mechanisms are shown in Tables 1 and 2.

TABLE I. EFFECT OF INTERNAL PARAMETERS ON THE MODEL

Batch Size	Average accuracy Mean average precision mAP percent	accuracy Precision P/%	recall rate Rcall R/%	confidence level (math.) Confidence/%
10	87.4	92.0	96.1	86.0
13	88.3	93.4	96.0	84.0
16	88.7	93.7	95.0	84.0
18	89.4	94.3	95.9	82.0
20	90.3	95.2	95.7	86.0

TABLE II. INCORPORATION OF MULTIPLE ATTENTION MECHANISMS

Attention Mechanism	mAP%	P/%	R/%
+SE	85.3	91.8	95.0
+ECA	86.6	92.3	94.1
+CCA	86.4	92.0	94.6
+SA-Net	87.8	92.5	94.7
+MS-CAM	87.5	92.7	95.2
+CBAM	88.5	93.2	95.8

The loss function is improved by adding a small target detection layer and introducing an attention mechanism. From the experimental data in Table 2, it can be seen that after the introduction of the CBAM attention module, the precision and recall of the model show an improvement, which effectively improves the performance metrics compared to other attention mechanisms.

D. Experimental result

The model of this article has been compared with other universal algorithms on Kitty recyclable data sets, which count as shown in table 3.

Compared with the benchmark model YOLOv5s, mAP@0.5 increases by 4.0%. Although the volume of the improved algorithm increases by 1.7MB, and the detection speed is reduced, but it still meets the real-time requirement.

TABLE III. COMPARISON OF DIFFERENT ALGORITHMS

Model	Volume	mAP@0.5%
YOLOv5s	14.0	86.9
YOLOv8s	22.4	86.5
YOLOv6s	37.4	83.0
YOLOv4	245.9	79.5
SSD	100.3	71.0
YOLOv5l	93.7	89.7
ours	15.7	90.9

E. Ablation experiments

In the study, some cutting-edge technologies were added to the YOLOv5s model, and three improvement points were proposed, namely, improvement of small target detection, denoted by X in Table 4, addition of the multi-scale channel attention mechanism CBAM, denoted by C in Table 4, and improvement of the MPDIoU loss function, denoted by L in Table 4, and denoted by ticks for whether it is added or not, resulting in an advanced detector, naming it XCL-YOLO. the improvement points are added to the YOLOv5s model to carry out the experiments, respectively.

According to the data in Table 4, the accuracy of each optimisation point for the detection of small targets has been improved to a different degree, and the improvement is significant. After the introduction of the small target detection module, there is an improvement of 3.1% on Person class and 1.4% on mAP. After adding the multi-scale channel attention mechanism CBAM, the improvement is 2.6% on Person class and 1.6% on mAP. Finally, improving the loss function from GIoU to MPDIoU improved 3.3% on Person class and 2.1% on mAP. A comparison of the algorithm before and after improvement is shown in Figure 9.

TABLE IV. ABLATION EXPERIMENTS

Methods	X	C	L	Person/%	Car/%	mAP%
YOLOv5s				79.2	94.3	86.9
X-YOLOv5s	√			82.3	94.4	88.3
C-YOLOv5s		√		81.8	95.2	88.5
L-YOLOv5s			√	82.5	95.5	89.0
XC-YOLOv5s	√	√		83.1	96.2	89.6
XL-YOLOv5s		√	√	83.4	96.6	90.0
CL-YOLOv5s	√		√	83.9	96.3	90.1
XCL-YOLOv5s	√	√	√	84.7	97.0	90.9

Figure 9. Comparison between before and after algorithm improvement

V. CONCLUSIONS

In this paper, an improved XCL-YOLO model algorithm is proposed, especially for the occlusion and small object detection problems in complex environments, and focuses on solving the object and missing detection problems often faced by existing models. The improved algorithm adds a small target detection layer, increases the attention mechanism, replaces the loss function, and adds the multi-scale channel attention mechanism CBAM. These changes adapt to the relationships and elements at different levels and on different channels, thus enhancing its ability to recognise targets and improve tracking accuracy. Recognition rate of vehicles and pedestrians in complex background is improved, and the approximation degree between boundary boxes can be accurately considered. The model can converge better in the training process, and the average accuracy mAP is increased by 1.4%, 1.6% and 2.1% respectively.

The new XCL-YOLO has been trained and tested on the KITTI dataset with an improved overall detection accuracy of 4.0% compared to the original model. The detection speed decreases slightly, but it still meets the detection requirements. This improved algorithm is more suitable for vehicle and pedestrian detection tasks. There will be overfitting risk in the training process, and future research can start from how to reduce the overfitting phenomenon and improve the generalization ability of the algorithm on various data sets. In addition, with the advancement of related technologies, we expect that such algorithms can be more widely used in the field of vision.

REFERENCES

[1] XU Yanwei, LI Jun, DONG Yuanfang, et al. A review of YOLO series target detection algorithms[J/OL]. Computer Science and Exploration, 1-19[2024-07-30].

[2] LI Yunpeng, HOU Lingyan,WANG Chao. Motion target detection in autonomous driving based on YOLOv3 [J]. Computer Engineering and Design, 2019, 40(04):1139-1144.

[3] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI,.USA, 2017, pp. 6517-6525

[4] Lv Wo-Feng, Lu Hua-Cai. Research on traffic sign recognition technology based on YOLOv5 algorithm [J]. Journal of Electronic Measurement and Instrumentation, 2021, 35(10):137-144.

[5] ZHAO H, FENG Y B. Research on traffic sign detection based on CGS-Ghost YOLO [J]. Computer Engineering, 2023, 49(12):194-204.

[6] JU Zhiyong, LI Yuming,XUE Yongjie,et al. Pedestrian detection algorithm based on improved YOLOv4 model [J]. Control Engineering, 2023, 30(10):1912-1926.

[7] SHAO Yanhua, ZHANG Duo, CHU Hongyu,et al. A review of YOLO target detection based on deep learning [J]. Journal of Electronics and Information, 2022, 44(10):3697-3708.

[8] ZHOU Miaosen, TANG Quanwu, SHI Sweet,et al. Crack detection algorithm for railway track surface based on improved YOLOv5s [J]. Liquid Crystal and Display, 2023, 38(05):666-679.

[9] TIAN Z, SHEN C, CHEN H, et al. Fcos: Fully convolutional one-stage object detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 9627-9636.

[10] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis & Machine Intelligen ce, 2017, 39(6):1137-1149.

[11] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. the

pas- cal visual object classes (VOC) challenge. int. j. Comput. Vis., 88(2):303-338, 2010.

[12] Sung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Doll ar. "Focal loss for dense object detection," in Proc. ICCV, 2017, pp. 2980-2988.

[13] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. Yolov4: Optimal speed and accuracy of object detection [J]. arXiv preprint arXiv:2004.10934, 2020.

[14] ZHANG Y F, REN W Q, ZHANG Z, et al. Focal and efficient IOU loss for accurate bounding box regression [J]. Neurocomputing, 2022, 506(9):146-157.

[15] LIU Hui, LIU Xinman, LIU Dadong. Optimisation of YOLOv5 algorithm for complex road target detection [J]. Computer Engineering and Applications, 2023, 59(18):207-217.

[16] YANG Feng, DING Zhitong, XING Mengmeng, et al. A review of improved target detection algorithms for deep learning [J]. Computer Engineering and Applications, 2023, 59(11):1-15.

[17] Heng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430, 2021.

[18] Yang G, Feng W, Jin J, et al. Face mask recognition system with YOLOV5 based on image recognition[C]//2020 IEEE 6th International Conference on Computer and Communications (ICCC). IEEE, 2020: 1398-1404.

[19] LI Xiang, HE Miao, LUO Haibo. An improved YOLOv3 algorithm for occluded pedestrian detection [J]. Journal of Optics, 2022, 42(14):160-169.

[20] ZHANG Y F, REN W Q, ZHANG Z, et al. Focal and efficient IOU loss for accurate bounding box regression [J]. Neurocomputing, 2022, 506(9):146-157.

[21] CHEN Jianzhu, WANG Yue, ZHU Xiaofei, et al. Wildlife video target detection method by fusing multi-feature maps [J]. Computer Engineering and Application, 2020, 56(07):221-227.

[22] JIA Zihao, WANG Wenqing,Liu Guangcan. Improved Light-weight Traffic Sign Detection Algorithm of YOLOv5 [J]. Journal of Data Acquisition and Processing, 2023, 38(6):1434-1444.

[23] MA S L, XU Y M. MPDIoU: A Loss for efficient and accurate bounding box regression[J]. arXiv preprint arXiv:2307.07662, 2023.

[24] XU X, ZHANG X, ZHANG T. Lite-YOLOv5:A lightweight deep learning detector for on-board ship detection in large-scene sentinel-1 sar images [J]. Remote Sensing, 2022, 14(4):1018

[25] ZHAO Lulu, WANG Xueying, ZHANG Yi, et al. Research on vehicle target detection technology based on YOLOv5s fusion SENet [J]. Journal of Graphics, 2022, 43(05):776-782.

[26] LI R, WU Y. Improved YOLOv5 wheat ear detection algorithm based on attention mechanism [J]. Electronics, 2022, 11(11):1673

[27] Sun Z, Li P, Meng Q, et al. An improved YOLOv5 method to detect tailings ponds from high-resolution remote sensing images [J]. Remote Sensing, 2023, 15(7): 1796.

[28] LIU Jiaze, WANG Chao, SHENG Long. Research on pedestrian detection method based on YOLOv5 [J]. Computer and Information Technology, 2024, 32(01):37-41.