Electronic Engineering and Informatics
G. Izat Rashed (Ed.)

doi:10.3233/ATDE240141

Improvement of YOLOv8 Detection Algorithm for Worker-Related Objectives in Construction Scenarios

Xuejing LI [a, b], Zhewei ZHANG[b1], Pengtao ZHAO [c]
[a] *Glodon Co., Ltd. , Beijing 100193, China;*
[b] *Department of Electronic Engineering, Tsinghua University, Beijing 100084, China;*
[c] *China Satellite Network Network System Research Institute Co., Ltd., Beijing 100029, China*

Abstract. In high security risk construction scenarios, the wearing of personnel safety equipment can effectively reduce safety risks. The performance of general YOLO series algorithms is still not sufficient when they are used in construction scenes to detect objects, such as pedestrians especially workers, helmets and reflective vests. This paper further optimizes the detection algorithm on the basis of YOLOv8 algorithm. The main contributions are as follows: the detection heads are modified to enhance the ability to extract small objects, a location loss function is replaced to optimize the model training process, and the attention mechanism module is added to better extract crucial features. This paper mainly conducts experiments and evaluations based on dataset SODA-C3 and dataset GLD-HR made for construction scenarios. The mean average precision of the optimal model trained by the improved algorithm reaches 83.1% and 88.5% respectively, which is 6.3% and 2.7% higher than that of the benchmark model, and the inference time is less increased.

Keywords. computer vision, deep neural networks, object detection

1. Introduction

In high-risk industries such as construction, mining, and heavy machinery manufacturing, the wearing of safety equipment such as helmets and reflective vests can effectively protect the personal safety of workers in on-site work scenarios. However, in practice, workers often fail to wear the relevant gear due to various reasons. The traditional supervision methods heavily rely on manual labor, which incurs high costs and has limited applied coverage. With the popularity of cameras and the development of machine vision algorithms, especially the gradual improvement of general object detection algorithms, the application of surveillance cameras for assisted supervision has been widely adopted.

General object detection algorithms have made significant progress in the past decade. Initially, some researchers employed manual feature extraction methods to achieve classification and detection by utilizing specific features such as color, position, and shape. However, these algorithms suffered from slow processing speed, low

[1] Corresponding author: Zhewei ZHANG, Department of Electronic Engineering, Tsinghua University, e-mail: demonmikalis@126.com

accuracy, and poor robustness. In recent years, with the breakthrough of deep learning techniques, object detection algorithms have witnessed rapid development[1,2]. The most representative algorithms in this field are the two-stage RCNN series algorithms and the one-stage YOLO series algorithms[3]. Compared to two-stage algorithms, the basic principle of one-stage algorithms is to segment the image into multiple regions and predict the position and category of the objects in one pass, which offers advantages in terms of speed. Moreover, after multiple iterative updates, the accuracy of one-stage algorithms has also been greatly improved[4]. However, these algorithms still face several challenges in practical applications: intra-class variations often occur among different instances of the same class due to lighting conditions, background variations, viewing points, person poses, occlusions, and other factors; the quality of labeled data is difficult to control due to variations in the expertise of annotators and the quality of the images; and the differences in deployment environments and device properties continuously give rise to new user requirements.

In order to further improve the performance of general object detection algorithms, especially for the tasks of safety helmet and reflective vest detection, researchers have explored various improvement methods. Wang Bing et al.[5] improved the YOLOv3 algorithm by introducing the GIoU loss. Xiao Tigan et al.[6] added shallow detection scales and increased the weight of position loss in YOLOv3. Zheng Xiao et al.[7] improved the K-means clustering algorithm in YOLOv4. Guo Shihong et al.[8] replaced the backbone network of YOLOv4 with Mobilenet-v3 and added a lightweight attention module. As for the YOLOv5 algorithm, Qiu Tianheng et al.[9] improved the path aggregation network with cross-layer weighted fusion, added an attention mechanism module with parallel mixed-domain convolution, and introduced the Ghost structure for model compression. Chen Yixiao et al.[10] utilized an enhanced visual transformer module, added a coordinate attention mechanism, and applied structural re-parameterization method. Wang Lingmin et al.[11] employed a weighted bidirectional BiFPN structure and added a coordinate attention mechanism. Yang Yongbo et al.[12] replaced the original backbone network with Mobilenet-v3 and added an attention module combining channel and space. Zhang Jin et al.[13] added a multispectral channel attention module in the backbone network and applied multi-scale training and K-means++ clustering algorithm. In addition, for pedestrian detection, Wei-Yen et al.[14] proposed a scale-aware YOLO algorithm. And Cao J. et al.[15] presented a comprehensive survey on recent advances in pedestrian detection from handcrafted to deep features.

In order to enhance the supervision of construction scenes, this paper first analyzes the recently open-sourced YOLOv8 algorithm and compares it with other algorithms in the YOLO series. Subsequently, considering both detection accuracy and speed, improvements are made to the detection head, loss function, and feature attention module. Finally, the model is trained and validated using the construction scene datasets.

2. The Analysis of the YOLOv8 Algorithm

YOLOv8 is an advanced object detection algorithm that builds upon its predecessors[16,17,18]. Compared with its predecessors, YOLOv8[19] exhibits notable advantages in terms of resource utilization and speed during training. Meanwhile, it achieves the performance of high accuracy and real-time. Consequently, these characteristics make YOLOv8 gain significant attention in the field of computer vision, especially in the sub-field of object detection.

2.1 The network structure

The YOLOv8 network architecture consists of multiple components that work together to extract and fuse features, as well as locate and classify objects. As shown in the structure chart in readme.md[20], the structure is mainly composed of four component including backbone, neck, and head.

In backbone, the multi-scale feature maps generated by the backbone are passed through a series of convolutional layers. The backbone component is based on a modified version of Darknet, specifically Darknet-53. Darknet-53 is a deep neural network that contains 53 convolutional layers and serves as the feature extraction backbone.

Followed by the backbone, the neck is designed to fuse the multi-scale feature maps. This process is achieved by the feature pyramid network (FPN) and path aggregation network (PAN), which are consist of several convolutional layers, up-sample layers and concatenation operations. This structure allows the whole network to detect objects at multiple scales by combining high-level and low-level features. And it is worth mentioning that the low-level features are essential for accurate object localization and high-level features are crucial for precise object classification.

After the neck, three detection heads (P3, P4 and P5) are designed in general, which are responsible for predicting different scale objects respectively. Correspondingly, these detection heads are connected to different layers with different down-sampling rates.

In terms of the overall architecture, YOLOv8 follows a one-stage detection framework. It takes an input image and passes it through the network, producing a set of bounding box predictions and class probabilities. Non-maximum suppression is then applied to filter out redundant detections and output the final set of object detections.

2.2 The algorithm improvements

Compared with its predecessors, YOLOv8 has several notable improvement as follows.

In backbone and neck, one key improvement in YOLOv8 is the integration of a more powerful feature extraction block, named c2f, which refers to the design methods of the CSP block and the Elan block in YOLOv7. The CSP block incorporates the concept of skip connections, which allow for the direct flow of information from early layers to later layers. And the Elan block adopt the concept of split feature maps, which enables the block to capture rich and discriminative features from the front feature map. In contrast to the c3 block in YOLOv5, the c2f block can obviously enhance the feature extraction capability of the algorithm.

In neck, one minor improvement in YOLOv8 is the refine the unnecessary layers which exist in the FPN module of YOLOv5. This improvement can accelerate the algorithm inference speed without any damage to the accuracy.

In head, YOLOv8 utilize the decoupled head which separate the location branch and the classify branch. With the decoupled head, the network can predict the bounding boxes, class probabilities, object scores respectively. This improvement enables to increase the detection effect because the location problem and the classification problem has weak correlation.

Furthermore, YOLOv8 incorporates the use of anchor-free method. Different from the anchor-based method which utilizes multiple aspect ratios and scales of anchor boxes, anchor-free method is more flexible. In detail, it traverses each point in the final feature map as a reference anchor point for object detection. Subsequently, by predicting the regression values for four directions, it obtains the coordinate offsets of each predicted

object relative to the anchor point. By default, the regression maximum parameter is set to 16 and the detection headers number is set to 3, which enable the algorithm to detect the objects of various sizes and shapes.

Regarding the loss function, YOLOv8 utilizes both the Complete Intersection over Union (CIoU) loss and the Distribution Focal Loss (DFL) for location, and utilize the Binary Cross-Entropy (BCE) for classify, which can be represented by Eq. (1). The CIoU loss measures the center distance between the predicted bounding box and the ground truth bounding box, and compares the length-width ratio of the two box. In addition to the CIoU loss, the DFL loss is designed to accelerate the convergence speed for multiple objects in different position of one image.

$$Loss = \lambda_1 L_{CIoU} + \lambda_2 L_{DFL} + \lambda_3 L_{Cls}, \quad \lambda_1 = 7.5, \quad \lambda_2 = 1.5, \quad \lambda_3 = 0.5 \tag{1}$$

Moreover, the algorithm introduces a new box matching strategy called Task-Aligned Assigner. This strategy replaces the traditional static matching strategy and aims to assign object proposals to ground truth objects in a more adaptive and context-aware manner. The Task-Aligned Assigner takes into consideration various factors, such as the size, aspect ratio, and location of objects, to ensure a more accurate and task-specific assignment of object proposals.

Additionally, another notable characteristic of YOLOv8 is its utilization of the multi-scale training tricks. This involves training the algorithm on images of various resolutions, which allows for better generalization and robustness.

3. Our Improvement and Optimization for YOLOV8 Algorithm

In recent years, there has been a significant focus on improving the network architecture and optimizing algorithms in object detection fields, which aim to enhance the performance, efficiency, and accuracy of existing models or algorithms. After some investigation, we have found that there is still improvement room with the open-source YOLOv8[19] algorithm in accomplishing specific tasks. Our improvement YOLOv8 network is shown in Figure 1 for better illustration.

3.1 The addition of attention mechanism

To enhance the ability of feature extraction, we integrate the attention mechanism in the YOLOv8 network. Compared by multiple attention mechanisms, Global Attention Mechanism (GAM)[21] is chosen in our approach. The GAM is a prior method that has demonstrated strong performance in various computer vision tasks, including the detection task. It enables the network to gain global context information while maintaining its ability to focus on local details. Similar to Convolutional Block Attention module (CBAM), GAM is designed to capture both spatial and channel-wise attention to enhance the cross-dimension interactions. Particularly, spatial attention enables the network to focus on important spatial regions within the feature maps, while channel-wise attention allows the network to emphasize on relevant channels that contain valuable object-related information.

3.2 The modification of detection heads

In order to improve the detection performance of small-sized objects while reducing the parameter and computational complexity, we optimized the design of the detection heads. Specifically, an additional detection head (P2) with a down-sampling rate of 4 is introduced to detect small-sized objects, since it can better capture fine-grained details. Additionally, in DFL loss function as mentioned in section 2.2, the regression parameters of the detection heads are set to 16 by default, hence the functionality of two detection heads (p3 and p4) with down-sampling rates of 8 and 16 can be replaced by the detection heads (p2 and p5) with down-sampling rates of 4 and 32. Therefore, in order to refine the original network structure, the replaceable detection heads (P3 and p4) are removed, while the detection head (p5) for detect large-sized objects is retained. This algorithm modification theory allows for a better trade-off between accuracy and speed.

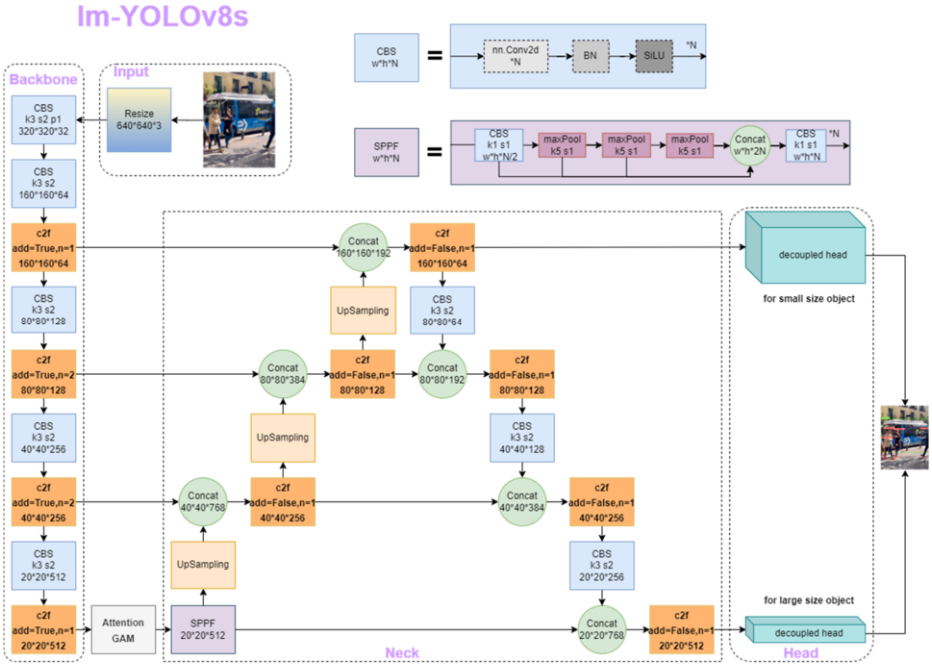


Figure 1. The network structure of improvement YOLOv8 algorithm

3.3 The refinement of location loss function

The location loss and object box regression is crucial for convergence process of object detection models. The goal is to accurately predict the coordinates of the bounding boxes surrounding the objects in one image. However, it can be challenging due to object variations such as in sizes, poses, and occlusions. To address this challenge, there has been continuous attention and research on improving the loss function of location regression.

Based on the overlap area between the two boxes of the ground truth box B^{gt} and the predicted box B, researchers primarily designed a loss function called Intersection

over Union (IoU) loss which can be represented by Eq. (2). The intersection is the area shared by the two boxes, expressed as $\left|B \cap B^{gt}\right|$, while the union is the total area covered by both boxes, expressed as $\left|B \cup B^{gt}\right|$. By dividing the intersection by the union, the IoU provides a value between 0 and 1, where 1 indicates a perfect match between the predicted and ground truth boxes.

$$L_{IoU} = 1 - IoU = 1 - \left|B \cap B^{gt}\right| / \left|B \cup B^{gt}\right| \tag{2}$$

However, it fails to effectively measure the difference between two boxes when they do not overlap, resulting in a constant value of 0, and it does not take into account the width and height information of the two boxes. Followed by Generalized IoU (GIoU) loss and Distance-IoU (DIoU) loss, CIoU loss represented by Eq. (3) is proposed to further optimize the penalty term of the IoU loss. CIoU loss introduces the normalized distance between the centers of the two boxes, which is the ratio between the Euclidean distance of the center points and the diagonal length of the minimum external bounding rectangle of the two boxes, expressed as $\rho(\cdot)$ and c respectively. Additionally, CIoU loss incorporates the width-hight ratio of the boxes into the calculation by considering the difference between the two boxes. By doing so, it provides a more comprehensive evaluation of box dissimilarity, considering both spatial and aspect ratio differences.

$$\begin{aligned} &L_{CIoU} = 1 - IoU + \rho^2(b, b^{gt}) / c^2 + \alpha\upsilon, \\ &\alpha = \upsilon / (1 - IoU + \upsilon), \\ &\upsilon = 4 / \pi^2 (\arctan w^{gt} / h^{gt} - \arctan w / h)^2 \end{aligned} \tag{3}$$

In addition, the WIoU (Weighted IoU) loss introduces a non-monotonic dynamic focusing mechanism, which can be represented by Eqs. (4), (5) and (6). It can further explore the potential information of the dissimilarity of the two boxes and address the balance issue between good and bad quality samples. In Eq. (4), R in the range [1:e] represents the distance penalty term, which is introduced as a focusing attention coefficient for amplifying the loss value of normal quality predicted boxes. Then it can be inferred that the range of WIoUv1 loss values is [0:e]. In Eq. (5), β represents the outlier degree, and the default values of δ and α are 3 and 1.9 respectively. The coefficient r is the non-monotonic dynamic gradient gain coefficient, which increases earlier and decreases later with the increase of β. When β=0, r=0. When β=x(0<x<δ), r reaches its maximum value and is greater than 1, which leads to the maximum gradient gain. When β=δ, r=1. When β>δ, r<1. To sum up, it can be inferred that the range of the WIoUv3 loss is [0:r*e]. The parameter m is the momentum coefficient, determined by the training rounds t and the training batches n. It is used to delay the time for $\overline{L_{IoU}}$ to reach its true value. At the beginning of training, $\overline{L_{IoU}}$ is initialized as 1, and the gradient gain increases with the increase of L_{IoU}. When L_{IoU} is 1, the gradient gain is maximized. In the early stage of training, due to the existence of the momentum coefficient m, $\overline{L_{IoU}}$ gradually approaches its true value, i.e. gradually decreasing. In the

later stage of training, $\overline{L_{IoU}}$ becomes small, and WIoUv3 assigns small gradient gains to low-quality predicted boxes to reduce harmful gradients, while focusing on normal quality predicted boxes to improve the localization ability. It can be speculated that this loss function effectively avoids the harmful effects of label errors on model training.

$$L_{WIoUv3} = rL_{WIoUv1} = rR_{WIoUv1}L_{IoU}, \quad R_{WIoUv1} = \exp(\rho^2(b, b^{gt}) / c^2) \tag{4}$$

$$r = \beta / (\delta\alpha^{\beta-\delta}), \quad \beta = L_{IoU}^{*} / \overline{L_{IoU}}, \quad m = 1 - \sqrt[tn]{0.05} \tag{5}$$

4. The Experiments of Improvement Algorithm

The experiments in this paper are conducted on a server with the Ubuntu 18.04 operating system. The server is equipped with an Intel(R) Xeon(R) Gold 5222 CPU @ 3.80GHz, which has 4 cards, each with 8 cores and 32GB memory. The GPU is the NVIDIA GeForce RTX 3090 with 4 cards, each with 24GB of memory. During experiments, a conda virtual environment is created with Python version 3.7.16, cudnn version 11.7, and torch version 1.13.1.

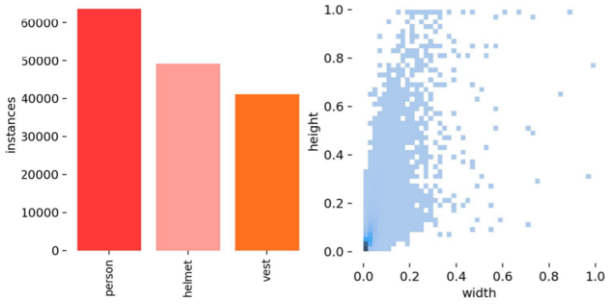
4.1 The dataset analysis

Through investigation, we find publicly available construction scene datasets include SODA[22], CHV[23], GDUT_HWD[24], MOCS[25], etc. Considering comprehensive factors in terms of image, instance, and class, other datasets are inferior to the SODA dataset. For example, the GDUT_HWD dataset only contains approximately 4,000 images and 3,000 instances for each class, which are relatively small in quantity. Mover, it merely annotated helmet with five classes: blue, white, red, yellow and none. In contrast, the SODA dataset consists of a total of around 20,000 images, with approximately 280,000 common object instances in construction scenes. It covers four major classes: workers, machinery, materials, and environment, with a total of 15 subclasses representing different kinds of instances. To ensure realistic and diversity, the dataset captured from various angles and resolutions, and the devices types consist of handheld cameras, drones, smartphones, surveillance cameras, and tower crane hook cameras. Regarding instances related to workers, there are around 70,000 person instances, 50,000 safety helmet instances, and 40,000 reflective vest instances. Considering the size and quality of the dataset, this paper primarily conducts experiments based on the SODA dataset. Since this paper focuses on worker-related objects, we created a subset of the SODA dataset, called SODA-C3. It retains only three classes of worker-related labels, named person, reflective vest, safety helmet, which represented as “P, R, S” respectively. By the way, the original dataset is randomly divided into training and validation sets in the ratio of 9:1.

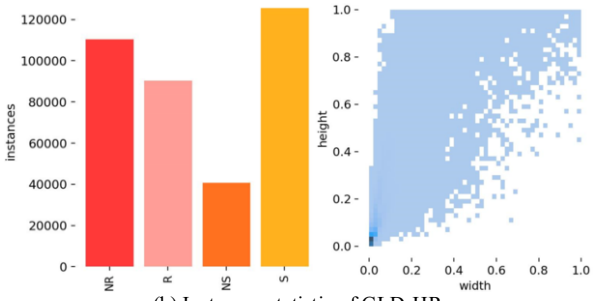
In addition, to distinguish whether workers are wearing safety helmet and reflective vest, and to monitor violations at construction sites, this paper collects a large number of images from various real construction projects of our company and creates a proprietary dataset called GLD-HR. This dataset contains a total of approximately 60,000 images, with around 110,000 instances of pedestrians without reflective vest, 90,000 instances of pedestrians with reflective vest, 40,000 instances of heads without safety helmet, and

120,000 instances of heads with safety helmet. For convenience, we use the "NR, R, NS, S" to represent them respectively.

For better illustration, the number of instances and the aspect ratios of bounding boxes for these instances in datasets SODA-C3 and GLD-HR are shown in Figure 2.



(a) Instances statistic of SODA-C3



(b) Instances statistic of GLD-HR

Figure 2. The instances statistic of datasets SODA-C3 and GLD-HR

4.2 The benchmark experiments

During model training, the default input image size is set to 640*640. The batch size is set to 64, and the number of epochs is set to 100. The pre-trained model is not loaded by default. The model configuration file is set to yolov8s.yaml, and the training framework is set to YOLOv8. During model evaluation, the default batch size is 1. The IoU threshold for non-maximum suppression is set to 0.45, and the confidence threshold is set to 0.25. The main evaluation metrics include average precision (AP), mean average precision (mAP), "parameters" indicating the model parameter count (in millions), "GFLOPs" indicating the model floating-point operations per second (in billions), and "time" indicating the average inference time per image (in milliseconds).

To compare the accuracy impact of whether to use pre-trained models, we conduct four experiments and the experimental results are presented in Table 1. In the experiments, "w" represents loading pre-trained weights, and "c" represents using only the network configuration file. It shows there is little performance difference in the trained model with and without loading pre-trained models. It indicates that the dataset is large enough and the training results are sufficiently optimized. Therefore, in the subsequent experiments, pre-trained models will not be loaded uniformly. In addition, “s” represents that the model is trained by small-sized network. Considering the factors of fast inference and lightweight deployment, the application of small-sized network

models is more widespread. Therefore, in this paper, the model trained based on the YOLOv8s network without pre-trained models is selected as the baseline model.

Table 1. The comparison of experimental results based on SODA-C3 and GLD-HR dataset.

	SODA-C3				GLD-HR				
model	mAP	AP(P)	AP(R)	AP(S)	mAP	AP(NR)	AP(R)	AP(NS)	AP(S)
YOLOv8s_w	0.787	0.878	0.777	0.707	0.867	0.877	0.848	0.896	0.847
YOLOv8s_c	0.768	0.858	0.759	0.688	0.858	0.869	0.843	0.882	0.84

4.3 The contrast experiments

In order to better evaluate the advantages of the improved models, multiple detection models are trained based on multiple network size and different algorithm frameworks. The experimental results are shown in Table 2.

It can be observed that as the size increases, the mAP and AP increase while the inference speed decreases. Among them, the model trained based on the YOLOv8s network has the smallest average inference time and the least number of parameters, but its mAP and AP are lower compared to models trained with the networks of large size. Additionally, as the model size increases, the incremental improvements in mAP and AP become smaller.

Besides, compared to models trained by other algorithm frameworks, the model trained based on the YOLOv8s network has advantages in terms of speed, parameter count, computational complexity, etc. Furthermore, it achieves competitive mAP and AP under the evaluation parameters specified in this paper.

Table 2. The comparison of experimental results based on multiple network sizes and different algorithms.

	SODA-C3					GLD-HR							
model	mAP	AP (P)	AP (R)	AP (S)	time	mAP	AP (NR)	AP (R)	AP (NS)	AP (S)	time	parameters	GFLOPs
YOLOv8s	0.768	0.858	0.759	0.688	7.7	0.858	0.869	0.843	0.882	0.84	7.9	11.1	28.4
YOLOv8m	0.779	0.871	0.767	0.7	9.6	0.87	0.882	0.852	0.895	0.852	9.9	25.8	78.7
YOLOv8l	0.784	0.876	0.77	0.705	11.3	0.878	0.891	0.863	0.899	0.857	12.1	43.6	164.8
YOLOv8x	0.789	0.88	0.775	0.712	13.0	0.881	0.893	0.867	0.906	0.859	13.1	68.1	257.4
YOLOv5s	0.793	0.852	0.777	0.752	7.5	0.871	0.871	0.848	0.898	0.867	7.6	7.1	15.9
YOLOv6s	0.786	0.866	0.768	0.739	7.7	0.876	0.878	0.854	0.897	0.871	7.8	18.3	45.1
YOLOv7s	0.777	0.853	0.785	0.694	7.9	0.860	0.844	0.838	0.878	0.881	8.2	36.5	103
PPYOLOE	0.781	0.859	0.771	0.723	7.9	0.868	0.873	0.851	0.889	0.857	8.1	7.6	17.1

4.4 The ablation experiments

To evaluate the improvement methods, we conduct extensive ablation experiments based on datasets SODA-C3 and GLD-HR. The validation results are shown in Table 3.

Firstly, different types of attention mechanisms are added at different positions in the original YOLOv8 network. Through comparison, the GAM attention mechanism module after the 8th layer of the original network structure yielded the best results, corresponding to the row of YOLOv8-AT8-GAM. Compared to the baseline model, it can be observed that there is a slight increase in the values of mAP and AP and little

decrease in the inference speed. We infer that the results benefit by the improvement of feature extraction capability of backbone network.

Secondly, the detection heads in the original network are redesigned, using combinations of "P2, P3, P4, P5", "P2, P3, P5", and "P2, P5", respectively. The results are indicated by the rows of YOLOv8-Head-P2345, YOLOv8-Head-P235, and YOLOv8-Head-P25. It is readily evident that incorporating the P2 detection head significantly enhances the detection performance of the network. Specifically, compared to the baseline model on Dataset 1, Model YOLOv8-Head-P2345 achieves mAP improvement of 6.1%, and AP improvement of 3.9% in P, 7.4% in R, and 6.9% in S. And compared to the baseline model on Dataset 2, the improvement Model achieves mAP improvement of 2.8%, and the AP improvement of 1.7% in NR, 2.0% in R, 2.3% in NS, 5.7% in S. It indicates that the proposed detection method particularly enhances the detection performance for small objects. Moreover, by removing the P3 and P4 detection heads while incorporating the P2 detection head, the model exhibits a slight decrease in accuracy but a significant increase in speed. This experimental result validates our theoretical analysis of using the P2 detection head to replace the P3 and P4 detection heads, which is mentioned in section 3.2.

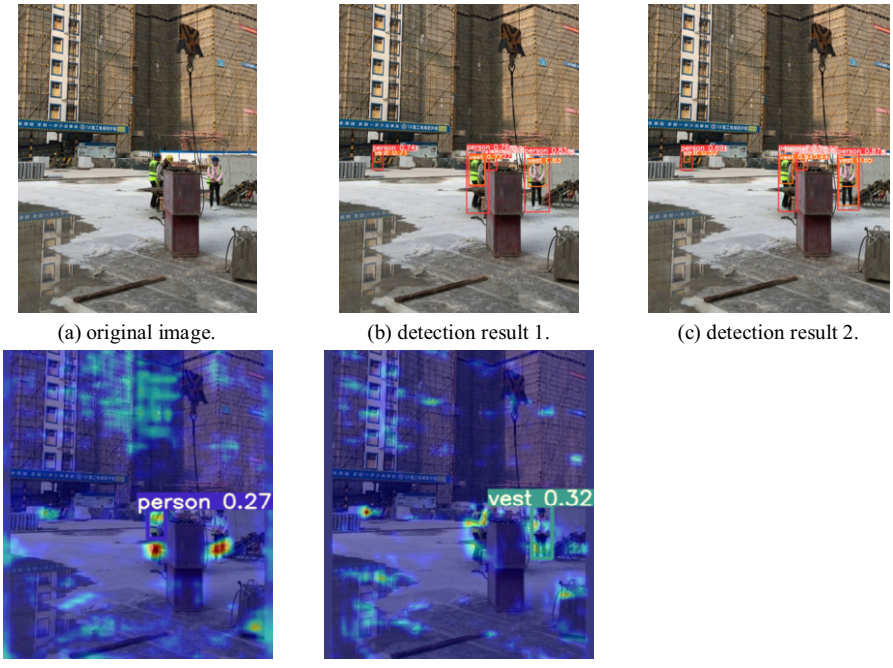
Thirdly, the CIOU loss function in the original algorithm is replaced, leading to the corresponding result in the row of YOLOv8-WIoU. Compared to the baseline model, it can be also observed that the method resulting in optimal accuracy without much degradation of inference speed.

Lastly, we conduct the experiments with all three improvement methods, and the results are in the row of YOLOv8-AT8-GAM-Head-P25-WIoU. Compared to the baseline model on Dataset SODA-C3, our improvement model achieves mAP improvement of 6.3%, and the AP improvement of 4.1% in P, 7.6% in R, 7.5% in S. And compared to the baseline model on Dataset GLD-HR, our improvement model achieves mAP improvement of 2.7%, and the AP improvement of 0.7% in NR, 1.9% in R, 2.3% in NS, 5.5% in S. Moreover, our improvement model has advantage of inference speed. After comprehensive comparison, our improved model outperforms all other models in terms of accuracy and speed.

Table 3. The results of ablation experiments for different improvement methods.

	SODA-C3					GLD-HR							
model	mAP	AP (P)	AP (R)	AP (S)	time	mAP	AP (NR)	AP (R)	AP (NS)	AP (S)	time	parameters	GFLOPs
YOLOv8s	0.768	0.858	0.759	0.688	**7.7**	0.858	0.869	0.843	0.882	0.84	**7.9**	11.1	**28.4**
YOLOv8s-AT8-GAM	0.773	0.86	0.763	0.695	8.3	0.861	0.872	0.845	0.884	0.841	8.5	17.6	33.7
YOLOv8s-Head-P2345	0.829	0.897	0.833	0.757	10.1	**0.886**	**0.886**	**0.863**	**0.905**	**0.897**	10.3	10.6	36.6
YOLOv8s-Head-P235	0.828	0.896	0.831	0.758	9.1	0.884	0.876	0.863	0.903	0.894	9.3	10.2	35.4
YOLOv8s-Head-P25	0.823	0.889	0.826	0.754	8.0	0.881	0.874	0.856	0.902	0.89	8.5	**10.0**	32.6
YOLOv8s-WIoU	0.772	0.859	0.763	0.694	8.0	0.860	0.870	0.845	0.882	0.840	8.1	11.1	**28.4**
YOLOv8s-im(ours)	**0.831**	**0.899**	**0.835**	**0.763**	9.0	0.885	0.876	0.862	**0.905**	0.895	9.1	16.5	37.8

To visually compare the detection performance of the models, Figure 3 presents the results of an image processed by different detection models. The original image contains four persons, four helmets, and four reflective vests. Therein, subfigure 3b shows the detection results of the baseline model, where the reflective vest of the second worker is falsely detected as a person, and the reflective vest of the third worker is missed. Subfigure 3c shows the detection results of our improved model, where the person of the fourth worker is falsely detected as a reflective vest. In addition, we illustrate heat maps of two false detection results. Subfigure 3d represents the heat map of the person generated by the baseline model. It assigns high attention to the person category at the location of the reflective vest of the second worker, leading to the false detection. Subfigure 3e represents the heat map of the reflective vest generated by our improved model. It assigns high attention to the reflective vest at the location of the person of the fourth worker, also resulting in the false detection. In summary, our improved model performs better in detecting objects in overlapping scenarios compared to the baseline model, but there is still room for further improvement.



(a) original image. (b) detection result 1. (c) detection result 2.

(d) heat map result 1. (e) heat map result 2.

Figure 3. The detection results of one test image and the corresponding heat maps.

5. Conclusion

To improve the object detection algorithm for worker-related instances in construction scenarios, this paper investigates various improvement methods and conducts multiple experiments based on the publicly available dataset SODA-C3 and our proprietary dataset GLD-HR. Considering both accuracy and speed of detection model, three improvement methods were selected: incorporating an attention mechanism module,

modifying the detection head, and refining the localization loss function. Experiments based on the SODA-C3 dataset show that compared to the baseline model, our improved model only increases the average inference time per frame by 1.3ms. And it achieves mAP improvement of 6.3%, the AP improvement of 4.1% in P, 7.6% in R, 7.5% in S. Similarly, experiments based on the GLD-HR dataset demonstrate that compared to the baseline model, our improved model only increases the average inference time per frame by 1.2ms. And it achieves mAP improvement of 2.7%, and the AP improvement of 0.7% in NR, 1.9% in R, 2.3% in NS, 5.5% in S. In summary, the proposed improvement methods have shown varying degrees of enhancement for the detection models trained on both datasets. Furthermore, there is a noticeable improvement in the accuracy of detecting small objects.

References

[1] CAO J L, LI Y L, SUN H Q, et al.A survey on deep learn-ing based visual object detection[J]. Journal of Image and Graphics,27(06):1697-1722.
[2] LU F, LIU H H, HUANG C Y, et al.Overview on Deep Learning-Based Object Detection[J]. Computer Systems and Applications, 2021, 30(3): 1-13.
[3] Joseph Redmon,Santosh Divvala,Ross Girshick,Ali Farhadi.You Only Look Once: Unified, Real-Time Object Detection[A].in:Institute of Electrical and Electronics Engineers.29th IEEE Conference on Computer Vision and Pattern Recognition: 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 26 June – 1 July 2016, Las Vegas, Nevada[C].2016.779-788.
[4] WANG X P, WANG X Q, LIN H, et al. Review on Im-provement of Typical Object Detection Algorithms in Deep Learning[J].Computer Engineering and Applications, 2022,058(006):42-57.
[5] WANG B, LI W J, TANG H.Improved YOLO v3 Algorithm and Its Application in Helmet Detection[J].Computer Engineering and Applications, 2020,056(009):33-40.
[6] XIAO T G, CAI L C, GAO X, HUANG H B, ZHANG C Y.Improved YOLOv3 Helmet Wearing Detection Meth-od[J].Computer Engineering and Applications, 2021,57(12):216-223.
[7] ZHENG X, WANG S Q, ZHANG W C, ZHENG J R, ZHOU Y.Safety Helmet Supervision System Based on Deep Learning[J].Computer Systems & Applica-tions,2021,030(011):118-126.
[8] GUO S H, JING R D, ZHANG X D, QIN X H.Research on detection of safety helmet wearing based on improved YOLOv4[J]. Joumal of Safety Science and Technology, 2021,17(12):135-141.
[9] QIU T H, WANG L, WANG P, BAI Y E. Research on Ob-ject Detection Algorithm Based on Improved YOLOv5[J].Computer Engineering and Applications, 2022,058(013):63-73.
[10] CHEN Y X, ALIFU K, LIN W L, YUAN X. CA-YOLOv5 for Crowded Pedestrian Detection[J]. Computer Engineering and Applications,2022,058(009):238-245.
[11] WANG L M, DUAN J, XIN L W. YOLOv5 Helmet Wear Detection Method with Introduction of Attention Mecha-nism[J]. Computer Engineering and Applications, 2022,058(009):303-312.
[12] YANG Y B, LI D. Lightweight Helmet Wearing Detection Algorithm of Improved YOLOv5[J].Computer Engineering and Applications, 2022,58(9):201-207.
[13] ZHANG J, QU P Q, SUN C, LUO M. Safety helmet wear-ing detection algorithm based on improved YOLOv5[J]. Computer Applications, 2022,42(4):1292-1300.
[14] Wei-Yen Hsu, Wen-Yen Lin.Ratio-and-Scale-Aware YOLO for Pedestrian Detection[J].IEEE Transactions on Image Processing,2021,30934-947.
[15] Cao J , Pang Y , Xie J ,et al.From Handcrafted to Deep Features for Pedestrian Detection: A Survey[J]. 2020.DOI:10.1109/TPAMI.2021.3076733.
[16] Redmon J , Farhadi A . YOLOv3: An Incremental Improvement[J]. arXiv e-prints, 2018.
[17] Li C, Li L, Jiang H, et al. YOLOv6: A single-stage object detection framework for industrial applications[J]. arXiv preprint arXiv:2209.02976, 2022.
[18] Wang C Y , Bochkovskiy A , Liao H . YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[J]. arXiv e-prints, 2022.
[19] Jocher, G., Chaurasia, A., & Qiu, J. (2023). YOLO by Ultralytics (Version 8.0.0) [Computer software]. https://github.com/ultralytics/ultralytics.
[20] MMYOLO Contributors. MMYOLO: OpenMMLab YOLO series toolbox and benchmark. https://github.com/open-mmlab/mmyolo, 2022.

[21] Liu, Yichao et al. “Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions.” ArXiv abs/2112.05561 (2021): n. pag.
[22] Duan R , Deng H , Tian M , et al. SODA: Site Object Detection dAtaset for Deep Learning in Construction[J]. 2022.
[23] Zhao Y. Fast Personal Protective Equipment Detection for Real Construction Sites Using Deep Learning Approaches[J]. Sensors, 2021, 21.
[24] Wu J , Cai N , Chen W , et al. Automatic detection of hardhats worn by construction personnel: A deep learning approach and benchmark dataset[J]. Automation in Construction, 2019, 106:102894-.
[25] An Xuehui, Zhou Li, Liu Zuguang, Wang Chengzhi, Li Pengfei, Li Zhiwei.Dataset and benchmark for detecting moving objects in construction sites[J].Automation in construction,2021,122(Feb.):103482.1-103482.18.